Advances in Interaction Studies

7

Future Robots

Towards a robotic science of human beings

Domenico Parisi

John Benjamins Publishing Company

Future Robots

Advances in Interaction Studies (AIS)

Advances in Interaction Studies (AIS) provides a forum for researchers to present excellent scholarly work in a variety of disciplines relevant to the advancement of knowledge in the field of interaction studies. The book series accompanies the journal *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*.

The book series allows the presentation of research in the forms of monographs or edited collections of peer-reviewed material in English.

For an overview of all books published in this series, please see *http://benjamins.com/catalog/ais*

Editors

Kerstin Dautenhahn
The University of Hertfordshire

Angelo Cangelosi
University of Plymouth

Editorial Board

Henrik Christensen
Georgia Institute of Technology

Harold Gouzoules
Emory University

Takayuki Kanda
ATR, Kyoto

Tetsuro Matsuzawa
Kyoto University

Giorgio Metta
IIT, Genoa

Adam Miklosi
Eötvös Loránd University

Robert W. Mitchell
Eastern Kentucky University

Chrystopher L. Nehaniv
University of Hertfordshire

Stefano Nolfi
CNR, Rome

Pierre-Yves Oudeyer
INRIA, Bordeaux

Irene M. Pepperberg
Harvard University & Brandeis University

Kerstin Severinson Eklundh
KTH, Stockholm

Stefan Wermter
University of Hamburg

Volume 7

Future Robots. Towards a robotic science of human beings
by Domenico Parisi

Future Robots

Towards a robotic science of human beings

Domenico Parisi
Institute of Cognitive Sciences and Technologies,
National Research Council, Rome

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Parisi, Domenico, author.
Future Robots : Towards a robotic science of human beings / Domenico Parisi.
p. cm. (Advances in Interaction Studies, ISSN 1879-873X ; v. 7)
Includes bibliographical references and index.
1. Robotics--Human factors. 2. Androids. 3. Artificial intelligence. I. Title.
TJ211.49.P37 2014
629.8'92--dc23 2014008326
ISBN 978 90 272 0461 5 (Hb ; alk. paper)
ISBN 978 90 272 7008 5 (Eb)

© 2014 – John Benjamins B.V.
No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

To my wife Cristina

Table of contents

Preface **xi**

CHAPTER 1
Robots as theories of behaviour **1**
1. The problem with scientific theories of human beings **3**
2. Theories as artefacts **6**
3. Robots as practical applications and robots as science **13**
4. One robot, many phenomena **16**
5. Human, not humanoid, robots **24**
6. This book **26**
   Appendix **28**

CHAPTER 2
Robots that have motivations and emotions **33**
1. The cognitive level and the motivational level of behaviour **33**
2. Today's robots do not have motivations **36**
3. Robots that have motivations **38**
4. Robots that have emotions **55**
5. Motivations, emotions, and covert attention **64**
6. Internal robotics **68**
7. Robots with a biological clock **75**
8. The two halves of the mind **79**

CHAPTER 3
How robots acquire their behaviour **81**
1. Why learning? **81**
2. The neural network of robots that learn **85**
3. The genetic predictability of the environment **89**
4. Living in genetically predictable environments **91**
5. Living in genetically unpredictable environments **98**
6. Learning in the experimental laboratory **100**
7. Imprinting and learning from one's mother **107**
8. The influence of learning on learning **112**
9. Learning with an evolved neural architecture **113**
10. On the limits of the robots described in this Chapter **116**

CHAPTER 4
Robots that have language **121**
1. The cognitive consequences of having language **121**
2. Meaning as co-variation between sounds and nonlinguistic experiences **122**
3. Classes of linguistic sounds **126**
4. Language helps human beings to categorize their environment **131**
5. The invention of language **141**
6. Asymmetries between language production and language understanding **146**
7. Robots that count and measure **149**
8. On the limits of our robots that have language **154**

CHAPTER 5
Robots with a mental life **159**
1. Mental life as the self-generation of sensory inputs **160**
2. Mental images **161**
3. Robots that predict **163**
4. Predicting and anticipating **167**
5. Evaluating the predicted consequences of one's actions **169**
6. Freedom of the will **171**
7. Predicted sensory inputs replace missing sensory input **174**
8. Other consequences of the ability to predict **179**
9. Talking to oneself **182**

CHAPTER 6
Social robots **187**
1. There is no social robotics today **187**
2. Living together **190**
3. Why *not* live together **196**
4. Socially damaging behaviours and how to contain them **199**
5. Why live together: Groups as information centres **211**
6. Living in small communities and living in large communities **213**
7. The social environment is very different from the non-social environment **215**

CHAPTER 7
Robotic families **221**
1. Genetic families and social families **221**
2. Mothers and daughters **223**
3. Grandmothers **227**
4. Sisters **231**

5. Males and females **235**
6. Homes **253**
7. Conclusions **257**

CHAPTER 8
Robots that learn from other robots and develop cultures and technologies **259**
1. Learning from others **259**
2. The cultural emergence of behaviours **263**
3. Staying near to others in order to learn from them **266**
4. Should adolescents learn from adults or from other adolescents? **269**
5. The evolution of artefacts **274**
6. Cultures **280**
7. The diminishing role of space in cultural processes **286**
8. Cultural globalization **289**
9. The expansion of Indo-European languages **296**

CHAPTER 9
Robot that own things **301**
1. External stores **302**
2. Individual stores **303**
3. Family stores **313**
4. Central stores **330**

CHAPTER 10
Political robotics **339**
1. Chief robots **340**
2. Chiefs and leaders **344**
3. No chiefs **350**
4. Historical changes in human settlements in ancient Southern Etruria **356**

CHAPTER 11
Robotic economies **361**
1. Goods **361**
2. Value **363**
3. Specialization and the exchange of goods **372**
4. Merchant robots **383**
5. The emergence of money **386**
6. Money and the price of goods **390**
7. Entrepreneur robots and worker robots **394**
8. The increase in the number of things that are goods for human beings **401**

CHAPTER 12
Individually different robots and robots with pathologies 407
1. Current robotics ignores inter-individual differences **407**
2. Fitness is not enough **410**
3. Motivations as a source of inter-individual differences **416**
4. Learning and inter-individual differences **417**
5. Why construct robots that have pathologies? **419**
6. Neurological pathologies and psychiatric pathologies **420**
7. Prediction, diagnosis, prognosis, and therapy **424**

CHAPTER 13
Robots that have art, religion, philosophy, science, and history 427
1. Explaining this Chapter **428**
2. Robots that have art **428**
3. Robots that have religion **435**
4. Doing metaphysics with robots **436**
5. Robots that do science **439**
6. Robots that have history **444**

CHAPTER 14
Human robots are future robots 451
1. A new science of human beings and its problems **451**
2. What has still to be done **452**

CHAPTER 15
How human robots can be useful to human beings 461
1. Human robots should be practically useful to human beings **461**
2. Robots make it possible for human beings to know themselves without being influenced by their desires and fears **462**
3. Robots can help human beings to identify possible conflicts between biology and culture and the consequences of these conflicts or their happiness **463**
4. Robots can help human beings to understand the impact of science on their life **467**
5. Robotic societies as tools for voters **472**
6. Difficult Problems **473**

References and additional readings 479

Index 489

Preface

Finish what you began, and we shall be born
(Brunella Antomarini, The Maiden Machine)

Human beings are the greatest challenge for science. They are the most complicated of all entities that make up reality. They are embarrassing entities because they seem to be made of both physical matter and something else which is not physical matter. They are difficult to study with the necessary detachment which is required by science because scientists also are human beings. This explains why science knows and understands nature much better than human beings. One might think that one must only wait and, someday, science will understand human beings as it understands all other phenomena of reality. But this may not be true unless the science of human beings changes radically.

The goal of this book is to outline a new science of human beings: a robotic science of human beings. The premise on which this science is based is that we will really understand and explain human beings and their societies only if we succeed in constructing robots that behave like human beings and live in societies which are like human societies. If we succeed in constructing these robots, the theory – or, rather, the blueprint – which has been used to construct the robots captures what underlies human behaviour and human societies and explains them.

But the robots of a robotic science of human beings must be *human*, not *humanoid*, robots. Today's humanoid robots have the external appearance of human beings and do some simple things that human beings do such as grasping an object with their hand or walking on two legs. Human robots must progressively reproduce all that we know about human beings: their body, their brain, their genes, their environment, their evolutionary origins, how they develop during their life, how they acquire new behaviours through learning and imitation, their motivations and emotions, their mental life, their families, their cultures, their economic and political institutions, and how they and their societies have changed and continue to change in the course of time.

These phenomena are traditionally studied by separate scientific disciplines – biology, neuroscience, psychology, anthropology, sociology, economics, political science, history – and this is a problem because scientific disciplines divide reality into separate pieces but reality is not divided into separate pieces. Reality is a large ensemble of different phenomena which are all connected together and, very often, the phenomena studied by one discipline can only be understood and explained by taking into consideration the phenomena studied by another discipline. A robotic science of human beings is a non-disciplinary science of human beings. One and the same robot and one

and the same collection of robots must reproduce all the different human phenomena which today are studied by separate scientific disciplines.

Today's robots are mostly technologies with practical applications and almost all research money goes to robots as practical applications. Our robots do not have practical applications but they are purely scientific tools. The only reason why we want to construct them is that, by constructing them, we can better understand human beings and human societies. They are mirrors through which human beings will be able to see themselves. This means that this book is addressed more to students of human behaviour and human societies than to roboticists – although roboticists may find suggestions on how to construct new robots with practical applications.

But our robots do have one practical application, and this may be the most important application of robots: they can help human beings to better understand the difficult problems they face today and will face in the future and, perhaps, to find solutions to these problems.

This is an ambitious program of research and this is why human robots are future robots. The robots described in the book reproduce in a very simplified way some very limited aspects of human behaviour and human societies. And writing the book has made very clear to its author how many phenomena concerning human beings and human societies still remain to be reproduced by constructing robots. But it is possible to make the first steps towards the realization of the program, and this book describes these first steps.

Acknowledgements

The robots described in this book have been constructed by people who have worked or are still working at the Institute of Cognitive Sciences and Technologies, National Research Council, Rome. Here is a list of these people: Alberto Acerbi, Gianluca Baldassarre, Paolo Bartolomeo, Valerio Biscione, Anna Borghi, Daniele Caligiore, Angelo Cangelosi, Giovanni Sirio Carmantini, Federico Cecconi, Antonio Cerini, Federico Da Rold, Tommasino Ferrauto, Dario Floreano, Simone Giansante, Onofrio Gigliotta, Nicola Lettieri, Henrik Hautop Lund, Filippo Menczer, Orazio Miglino, Marco Mirolli, Francesco Natale, Stefano Nolfi, Fabio Paglieri, Giancarlo Petrosino, Fabio Ruini, Filippo Saglimbeni, Matthew Schlesinger, Massimiliano Ugolini, Alberto Venditti.

I thank Cristina Delogu, Nicola Lettieri, and Marco Mirolli for reading some of the chapters of the book and suggesting how to improve them. I am also very grateful to Cristina Delogu, Emanuela Parisi, and Giuseppe Polegri for their great help in preparing the book's figures.

CHAPTER 1

Robots as theories of behaviour

Martian Eye (ME) is a scientist from Mars which one day arrives on the Earth and, of all the things which exist on the Earth, it decides to study human beings. There are no particular reasons why ME chooses to study human beings rather than some other species of animals except that human beings are particularly complex animals and, therefore, they represent a greater challenge for ME. But, to tell the truth, there is one reason why ME chooses to study human beings. Human beings have many desires and many fears concerning themselves and these desires and fears obscure the knowledge that they have of themselves. They think they are what they would like to be and they are not what they don't like to be. And they are inevitably anthropocentric. They see themselves as the centre of the world. But they are not the centre of the world and their anthropocentrism distorts the vision that they have of themselves. Scientists are human beings and, for them too, it is difficult to see human beings as only one of the many things that make up reality and to study human beings with the necessary detachment which is required by science. ME is not a human being and it believes that this is very important to understand human beings in the way in which science understands every other phenomenon of reality.

ME knows that objective and quantitative data are the basis of science but data is not enough. Science must answer the question "What?" but it must also answer the question "Why?", and to answer the question "Why?" science needs theories which explain the data. Human scientists express their theories by using words or mathematical symbols. ME is a Martian scientist and it has its own way of doing science. To express its theories, ME constructs artefacts that incorporate its theories and behave like what the theories must explain. ME's science is based on the principle "Whatever X is, to understand X you must reproduce X in an artefact". Only if its artefacts do what human beings do, ME is satisfied that it has understood and explained human beings.

When it arrives on the Earth, ME finds that human scientists already construct physical artefacts that resemble human beings and other animals, and they call these artefacts "robots". But human scientists mostly construct robots because robots have practical application and economic value. ME's robots are purely scientific tools that must make it possible to better understand what human beings are. Human scientists call their robots which resemble human beings "humanoid". To distinguish its robots from "humanoid" robots, ME calls its robots "human".

Constructing human robots that behave like human beings and live in societies which are like human societies is a very difficult task and ME is well aware that its robots often resemble much simpler animals and that really human robots are future robots. But ME's program of research is clear.

Another thing which makes ME's science different from how science is done on the Earth is that for ME the divisions among the different scientific disciplines do not exist. Reality is a large ensemble of different phenomena but these phenomena are all linked together and, very often, to understand the phenomenon studied by one discipline it is necessary to take into consideration the phenomena studied by other disciplines. ME's science is a non-disciplinary science of human beings. The robots constructed by ME must reproduce all aspects of human behaviour, from its physical and biological bases to all the individual and social phenomena which, on the Earth, are studied by psychologists, anthropologists, sociologists, economists, and political scientists.

ME also believes that to understand human beings it is necessary to know how they have become what they are: how they have evolved from nonhuman ancestors, how they mature and develop during their life, what they learn from their experiences, how human cultures, economies, and political institutions were born and how they have changed and continue to change. This implies that ME cannot design or program its robots but they must autonomously become what they are.

ME is well aware that, like all theories, its robots simplify with respect to real human beings and real human societies but it also knows that, to be useful, scientific theories must necessarily simplify with respect to reality. The problem is that they must make the appropriate simplifications, and ME believes it is more probable that its robots make the appropriate simplifications if one and the same robot and one and the same robotic society reproduce not one single phenomenon but as many phenomena as possible about human beings and human societies.

ME's research "philosophy" can be summarized in three principles:

- If you want to understand human beings, construct robots that behave like human beings.
- If you want to understand human beings, construct robots that not only reproduce human beings but also reproduce how human beings have become what they are.
- If you want to understand human beings, construct robots that reproduce as many different phenomena as possible concerning human beings.

ME is not an imaginary scientist like the Martian scientists of science fiction and thought experiments. ME is a real scientist and it constructs real robots – although, for now, most of its robots are simulated in a computer. And, being a real scientist, ME wants to discuss and collaborate with human scientists.

1. The problem with scientific theories of human beings

Imagine you are interested in how human beings behave and how their societies function and that you want to understand human behaviour and human societies as science knows and understands all other phenomena of reality. What must you do? You must collect objective and possibly quantitative data on the phenomena that interest you because, for science, reality is, first of all, empirical data. But empirical data is not enough. You must also formulate theories that explain the data. Traditionally, scientific theories of human behaviour and human societies are formulated in words but words have serious limitations as tools for expressing scientific theories. Words have unclear meanings, they have different meanings for different people, and they are often value-based and emotionally charged – and these properties of words are not good for science. A crucial requirement of scientific theories is that from a scientific theory it must be possible to derive specific and unambiguous predictions to be compared with the empirical data. Verbally expressed theories are generally unable to generate specific and unambiguous empirical predictions.

Scientists can agree on these limitations of words as scientific tools and they can try to overcome them by defining the words used in their theories or by specifying in which sense they use them. Or they may create new words and define them by using existing words. But this, clearly, is a circular strategy. Words are defined or re-defined by using other words, and these other words also have unclear meanings, mean different things to different people, and are value-based and emotionally charged. So, expressing scientific theories by using words remains a problem for science because verbally formulated theories leave us with what the English poet Thomas Stearns Eliot called "the intolerable wrestle with words and their meanings".

There is another property of words which is a problem for science. Given a word, there is a tendency to think that there exists an entity which corresponds to the word. This may be all good for the everyday use of language but it is not good for science. Given a word of the common language, it is not automatic that the explanatory apparatus of science should include the entity designated by the word. For example, the words "belief" and "goal" may be perfectly appropriate for everyday life. But a good scientific theory of human behaviour and human societies may *not* need these words and may better explain the phenomena to which in everyday life we refer by using these words by *not* using these words.

Expressing scientific theories by using words explains why the paradigmatic science of human behaviour, psychology, is only a half-science. The human mind has been studied by philosophers for thousands of years, and philosophers have studied the mind in their characteristic way: by doing conceptual analyses, by reasoning and presenting arguments, and by discussing with colleagues. More than one century ago, psychology was born, and psychology was a revolution in the study of the mind. Psychologists investigate the mind by observing behaviours in the controlled

conditions of an experimental laboratory and in other objective and quantitative ways. Therefore, they can claim they have created a *science* of the mind which uses the same methods of the natural sciences and is not confined, like philosophy, to conceptual analyses, argumentation, and discussion with colleagues. But science does not need only objective and quantitative data. It also needs theories that explain the data. The problem with psychology is that while its empirical methods are those of science, its theoretical vocabulary is still a philosophical vocabulary. If you read psychological books and papers, you find the same vocabulary of ancient or recent philosophical treatises, or the vocabulary that people use in their everyday life to talk about their behaviour and the behaviour of others. Here is a sample of this vocabulary: sensation, perception, attention, memory, thinking, reasoning, predicting, planning, motivation, emotion, all the different motivations and emotions, representation, concept, category, meaning, object, property, action, intention, goal, awareness, consciousness. These words cannot be the vocabulary of a science because they have unclear meanings, they have different meanings for different scientists and, even when one scientist provides a precise definition of one particular word, the definition is not generally adopted by his or her colleagues. Therefore, using words to express psychological theories causes endless and often useless discussions and it is a serious obstacle to the progress of psychology. Science is based on a virtuous circle between theories and data: theories must predict data, and data must confirm theories. Psychologists find it difficult to establish this virtuous circle. Theories are formulated and discussed but it is difficult to arrive to a consensus on which theory is the right one. Data are collected, often in the controlled conditions of an experimental laboratory, but they are rarely really illuminated and explained by a theory. This is why psychology is only a scientific half-revolution. Psychologists adopt the empirical methods of science but they continue to talk about human behaviour and the human mind by using the vocabulary of philosophy or everyday life.

Psychology is the paradigmatic science of human behaviour but many other disciplines study human beings: anthropology, sociology, economics, political science, history. These disciplines are called the social sciences and with the social sciences the problems are even more serious. Not only these sciences have verbally formulated theories but, since they generally cannot study the phenomena that interest them in the controlled conditions of the experimental laboratory, their empirical data are often unclear – or nonexistent. Therefore, for the social sciences, the necessary dialogue between theories and facts which is the foundation of science cannot be established because neither of the two interlocutors actually exists. (The science of economics is the only exception to this rule but the science of economics has its own problems. See later in this section.)

Another weakness of the social sciences is that they often resemble philosophy more than science. Much of the work of social scientists is an exegesis of what other social scientists have written and social scientists give the impression that they do not

really want to know and understand reality as science knows and understand reality but, like philosophers, they only want to "interpret" reality and provide arguments for or against different interpretations. And, like philosophers, it is often unclear if social scientists are interested in what human beings and human societies actually *are* or if they are proposing what they *should* be.

The sciences of nature have found a way out of the problem of verbally expressed theories. Physicists, chemists, and biologists define the words they use not by using other words but by referring to what they observe, count, and measure, and what they observe, count, and measure exhaustively captures the meaning of the words they use. This makes their theories really scientific theories because they necessarily agree on what they observe, count, and measure. The sciences that study human beings are unable to do the same thing and their theories remain inherently verbal. The words that express their theories only vaguely and indirectly refer to what is actually observed and, even when the reference is explicit and what is observed is also measured and counted, what is observed, counted, and measured does not exhaust the meaning of the theoretical vocabulary of these sciences.

And there is still another problem. The theoretical vocabulary of the sciences that study human behaviour and human societies is made up of words of this or that Western language – today, mostly the English language. But words do not have exact translations in other languages. Different languages segment and express reality in different ways, and this is especially true for the words that psychologists and social scientists use to talk about the mind and about society. Why should one particular language, the English language, capture the real nature of the human mind and of human society while this is not true for other languages, for example, Chinese or Kenyan? Physicists, chemists, and biologists, whatever their language and their culture, refer to the same things – atoms, molecules, cells, organisms – because the meaning of the words they use is completely defined by what they observe with their instruments and by what they count and measure. But, as we have said, this is not true for the words used by the sciences of human behaviour and human societies.

If verbally expressed theories have the limitations we have said, another possibility is to mimic physics and express theories of human behaviour and human societies by using mathematics. Unlike words, mathematical symbols have precise and unambiguous meanings and they have exact translations in other languages. And mathematics is a powerful tool not only to describe precisely and noncontroversially what is observed but also to find precise relations among different observations and to generate precise and noncontroversial empirical predictions – which is a crucial requirement for science. Are mathematically expressed theories a possible solution for the sciences that study human beings and human societies? Today, the availability of computers with their very large memory and processing capacities induce many researchers to apply mathematical theories to behavioural and social phenomena. It is still difficult to judge these attempts because they are in their early stages and they can certainly

shed some light on behavioural and social phenomena. However, the science of human beings which has more consistently and completely embraced the idea of expressing its theories in mathematical terms, the science of economics, appears to be unable to go beyond a superficial understanding of economic phenomena, as indicated by its inability to predict them and to advise on how to control them, and to think of economies different from market economies.

The same may be true for other attempts to use mathematics and statistics to study behavioural and social phenomena. These attempts find the regularities present in large quantities of data (which, today, with computers and the Internet, have become "big data") but they do not even try to identify the mechanisms and processes which underlie behavioural and social phenomena and explain them. Mathematics is an useful tool for computing relations among quantities and for making quantitative predictions but if what is computed and predicted does not capture the real nature of the phenomena we want to understand, these phenomena will not be understood. For some reason – which science itself should be able to identify – while mathematical theories capture all we need to know about the phenomena studied by physicists and explain these phenomena, this is not true when mathematical theories are applied to the phenomena studied by behavioural and social scientists.

Let us summarize what we have said. Science is both empirical data and theories that explain the empirical data. Scientists who study human beings and their societies lack good theories because verbally expressed theories and, for them, also mathematical theories have serious limitations. This explains why, while the sciences that study nature make constant and cumulative progress, this is not true for the sciences which study human beings and human societies, and why a science of human beings and their societies still does not really exist. If this correctly describes the present state of affairs, is there something that we can do?

2. Theories as artefacts

There is something that we can do. Computers offer a new possibility for expressing scientific theories. Theories become blueprints for constructing computer-based artefacts, and the artefact *is* the theory. If we are interested in understanding one particular piece of reality and we succeed in constructing an artefact which reproduces that piece of reality and how that piece of reality functions, the theory which has been used to construct the artefact explains that piece of reality. According to this new paradigm for doing science, we should not ask "What is X?" or "How can X be explained?" but we should ask "Can I reproduce X in an artefact?" and "How is made the artefact which reproduces X?" The computer has a critical role in this enterprise because the theory-artefact is either simulated in a computer or is a physical artefact controlled by a computer.

Theories as artefacts have many advantages with respect to both theories expressed in words and theories expressed by using mathematical symbols. Unlike theories expressed in words, theories as artefacts are necessarily precise and nonambiguous because, if they are not precise and nonambiguous, the artefact cannot be constructed. Theories as artefacts provide operational definitions of words. Words are no more defined by using other words but they are translated, exhaustively, into some particular feature of the artefact, and we can always "open" the artefact and see how the artefact is made and how it functions.

Another advantage of theories as artefacts is that they generate many detailed and noncontroversial empirical predictions because the manner in which the artefact functions and what it does *are* the predictions derived from the theory which has been used to construct the artefact – and, again, the manner in which the artefact functions and what it does are under the eyes of everybody and can be measured precisely. If the artefact functions like the piece of reality which it is intended to explain, the theory incorporated in the artefact can be considered as (provisionally) correct. If not, the theory – the artefact – must be modified or abandoned.

This also solves the problem of which language to use for expressing theories of human behaviour and human societies. Today, the language of science is (mostly) English and, as we have said, for verbally expressed theories this is a problem because the words of the English language, like the words of any other language, presuppose one particular and culturally biased view of the human mind and of human societies. Expressing theories as artefacts solves this problem because theories of human behaviour and human societies expressed as artefacts are non-verbal theories and anyone, whatever his or her language – English, Chinese or Kenyan – can construct, observe, and analyze the *same* artefact. This is true not only for language and also, more generally, for culture. Scientists belong to specific cultures and a scientist's culture shapes his or her mind – as psychologists and anthropologists know very well. If a theory of behaviour and society is expressed as an artefact or a collection of artefacts, scientists can look at the theory-artefact in a culturally neutral way.

With respect to mathematical theories, theories as computer-based artefacts have the advantage that they do not only capture the regularities which exist in the empirical data but they also identify what lies behind these regularities and explains them because, as we have said, we can "open" the artefact and see what produces the artefact's behaviours. The computer is not only a useful tool for memorizing large quantities of data e for doing long and difficult mathematical calculations to capture the regularities contained in the data but it is an artificial version of the piece of reality we want to understand and explain.

Theories as computer-based artefacts can be theories of all sorts of phenomena studied by all sorts of scientific disciplines but they are especially useful for studying human beings and human societies because theories of human beings and human societies have the limitations that we have said. If we want to understand why human

beings behave as they behave and why their societies function as they function, what we must do is construct artefacts which behave like human beings and collections of artefacts that function like human societies. Galileo said that the "book of nature" is written in *mathematical* symbols – and he was right. The "book of human beings" so far has been written in *alphabetical* symbols (words) but alphabetical symbols do not have the precision and univocity of mathematical symbols. The solution is to abandon the idea of a "book" when studying human beings. If you want to understand human beings what you need is not a book, written in either mathematical or alphabetical symbols, but a "tool box" for constructing artificial human beings.

Expressing scientific theories of human beings as computer-based artefacts will help to solve another problem which afflicts the sciences of human beings and explains why these sciences are so underdeveloped and so unsatisfactory when compared with the sciences of nature. We do not have an adequate science of human beings because the science of human beings is divided up into separate disciplines: biology, psychology, linguistics, anthropology, sociology, economics, political science, history. The division of science into separate disciplines reduces the capacity of science to know and understand reality because reality is *not* divided up into separate parts. Reality is a large ensemble of different phenomena but these phenomena are all connected together and, often, the phenomena which are studied by one discipline can only be explained by taking into consideration the phenomena studied by other disciplines. Science functions because scientists work together and, by working together, they know much more about reality than they would know if they were to work alone (see Figure 1).

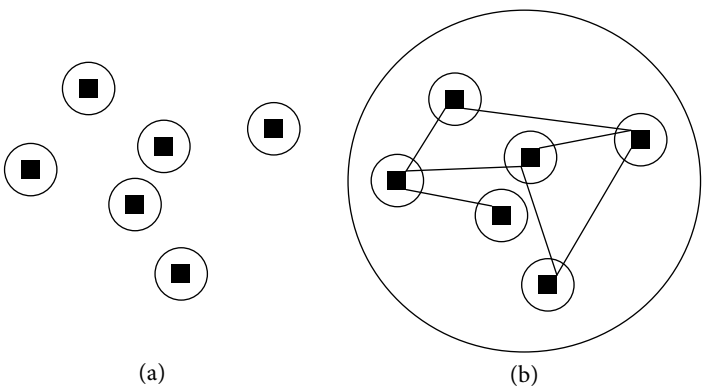


Figure 1. (a) Six scientists (black squares) and what they would know about reality (circles around the black squares) if they were to work alone without interacting together. (b) The same six scientists and what they know because they interact together (large circle).

The same applies to scientific disciplines. Science knows reality much less than it would know if there were no disciplinary divisions (see Figure 2).

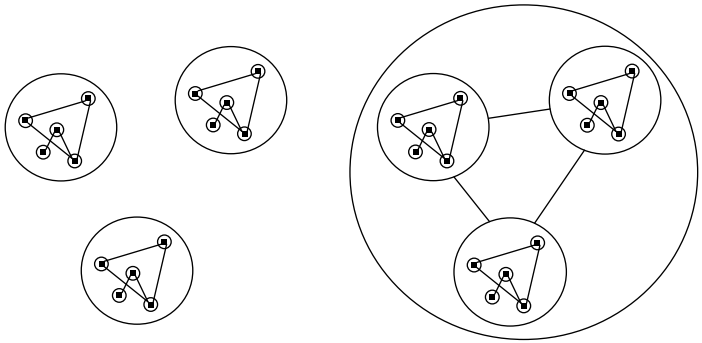


Figure 2. (a) Three scientific disciplines and what each of them knows about reality (small circles). (b) What science would know about reality if there were no disciplinary divisions (large circle).

The science of nature also is divided into disciplines: physics, chemistry, biology. But the division into disciplines does not create problems for the science of nature because physics, chemistry, and biology have a clear vision of how physical, chemical, and biological phenomena are linked together, they all use the same empirical method (laboratory experiments), they have very similar conceptual and theoretical traditions, and they share a view of nature as made up of physical causes that produce physical effects and as possessing an inherently quantitative character. In contrast, the division of the science of human beings into separate disciplines has very negative consequences because the behavioural and social sciences do not have a clear vision of how biological, psychological, and social phenomena are connected together, they do not use the same empirical methods to study these phenomena, they have very different conceptual and theoretical traditions, and they do not have a unified and shared view of the phenomena they study. The consequence is that the sciences of human beings give us the separated pieces of a mosaic but not a unified picture of human beings.

What are the pieces of this mosaic? Psychologists study individual human beings and the interactions among individual human beings but they ignore the societies in which human beings live, the cultures that shape their behaviour, and the past history of their societies and cultures. Psychologists also mostly ignore the biological bases of human behaviour, especially those studied by evolutionary biology and genetics, while only recently there is some dialogue between psychologists and neuroscientists. Another problem with psychology is that it ignores the complex environment in which human beings live. The best empirical data collected by psychologists are those of laboratory experiments but experimental laboratories are artificial and very simplified environments. Behaviour is interacting with the environment and, therefore, laboratory experiments inevitably give us a limited knowledge of behaviour. In real life

human beings normally do what they want to do and they expose themselves to stimuli which are to a large extent determined by their own behaviour. In the experimental laboratory "subjects" – or "participants", as one prefers to call them when "subjects" are human beings – do what the experimenter wants them to do and they are exposed to stimuli which do not depend on their behaviour but are decided by the experimenter. And when the experimenter has done one experiment, the subjects of the experiment are generally no more available for doing other experiments that would make it possible to better understand their behaviour – unless they are animals living in cages.

Linguists study a very important human phenomenon, language, but they do not really understand language because they isolate language from the rest of human behaviour and they ignore the basis of language in the human brain. Social scientists not only ignore most of what psychologists know about human behaviour but they segment human societies into separate parts – sociologists study the overall organization of human societies, anthropologists their culture, economists their economical system, political scientists their political institutions, historians their past history – and they have a very limited knowledge of the parts that they do not study and a very limited understanding of how the different parts interact together. Another problem of the social sciences is that they ignore nature and, in fact, the link between the social sciences and the biological sciences is practically nonexistent. This is a problem because human beings *are* nature and, even if one may think that they are not *only* nature, they have a natural past and we cannot understand any aspect of reality unless we know how that piece of reality has become what it is. And, finally, there are many important human phenomena such as art, religion, philosophy, and science itself which are almost completely ignored by science. They have remained an object of study for philosophy and the humanities and, therefore, they are not even part of the mosaic that the science of human beings tries, with little success, to put together.

Theories of human beings expressed as computer-based artefacts make it possible to put the pieces of this mosaic together. The different disciplines that study human beings are like the blind men who find an elephant and try to understand what kind of animal they have found. One touches the elephant's legs, another one its long nose, a third one its large body, a fourth one its short tail, but they do not really understand what sort of animal they have found. Theories as artefacts can help us to recognize the elephant and to construct a non-disciplinary science of human beings. Theories as computer-based artefacts are a *unified* theoretical and methodological framework, a "lingua franca" that can facilitate the dialogue among biologists, psychologists, and social scientists, and the development of an integrated science of human beings. The *same* artefact should reproduce the phenomena studied by the different disciplines and the different phenomena should interact together in the artefact. Before the computer, it was impossible to have a unified science of human beings because the brain of a single scientist cannot know the many different facts studied by the different disciplines and cannot understand the relations among all these facts. The computer has memory and

processing capacities which go well beyond those of the human brain. If a scientific theory is expressed as a computer-based artefact, the theory is *in* the artefact, not in the brain of the scientist, and computer-based artefacts can contain enormous quantities of information and can easily and rapidly compute what happens when the different pieces of information interact together.

Expressing theories of human beings as artefacts has another important advantage, this one of a different nature. Human beings have many desires and many fears about themselves, and these desires and fears are an obstacle to their capacity to know themselves because they tend to confuse what they are with what they would like to be. Scientists also are human beings and they inevitably carry their desires and fears with them when they do science. This is a problem because scientists can only know and understand reality if they look at reality with complete detachment and with a mind free from desires and fears. This is easy to do when scientists study nature but it is much more difficult when scientists study human beings and human societies – and this also explains why science knows human beings and human societies so much less than nature.

Expressing theories of human beings and human societies as computer-based artefacts solves this problem. Theories-artefacts are detached from us, they are objects outside us. If a theory is formulated in words, the theory is often a value-laden "interpretation" of the facts and we tend to derive from the theory only those predictions that correspond to our values and ideologies and to ignore other predictions which are in contrast with our values and ideologies. If a theory is expressed as an artefact, we are necessarily confronted with *all* the predictions that can be derived from the theory. The predictions of the theory are the behaviours of the artefact, and the artefact cannot exhibit the behaviours that correspond to our values and our desires and fail to exhibit those which are in contrast with our values and our desires. Theories as artefacts are mirrors in which we see ourselves, whether we like or not what we see in the mirror.

This mixing together of science and values takes another form, especially in the social sciences. Very often, it is not clear if social scientists want to know what human societies are or if they want to propose how human societies should be. For the purposes of science, these two things should be kept separate because science aims at understanding how reality is, not what it should be. Instead, it is often unclear which of the two things social scientists – especially, economists and political scientists – are doing and this is bad for both science and for human beings. Social scientists cannot really understand human societies unless they look at human societies with the same detachment with which natural scientists look at nature. And, if they do not really understand human societies, how can they help human beings to change human societies and to make them better?

Theories of human beings expressed as artefacts also clearly separate science from philosophy. Unlike the sciences of nature, the sciences of human beings still have many links with philosophy. This is clear for the social sciences but it is also true for

psychology. "Cognitive science" is an inter-disciplinary effort to better know the human mind and philosophers are important contributors to cognitive science. Theories of artefacts are not "cognitive science" because they are not *inter*-disciplinary but *non*-disciplinary and because they leave philosophy completely out of science. Philosophy is made of words while the "motto" of theories as artefacts is "Artefacts, not words". According to the Austrian philosopher Ludwig Wittgenstein, "whereof one cannot speak, thereof one must be silent". But scientists must not speak about reality but they must reproduce reality. Philosophy is "love of wisdom" while science has nothing to do with wisdom. Philosophy is, or has been, a manner of living while science is not a manner of living. Trying to understand human beings by constructing artefacts that behave like human beings makes all these things very clear.

Artefacts as scientific theories are computer-based artefacts in the sense that the artefact is simulated in a computer or is a physical artefact controlled by a computer. But computers are only useful tools and, in principle, we might incorporate a scientific theory in an artefact – something which has been made by us – even without computers. Computers are necessary only because reality is very complex and we can reproduce reality in an artefact only with the help of a computer. Someone has proposed that the mind *is* a computer. But this has nothing to do with our proposal to express scientific theories by constructing computer-based artefacts. Everything can be simulated in a computer and, therefore, everything – not only the mind – would be a computer. The only link between the human mind and computers is that the human mind has invented the computer.

As we have said, the "motto" of the new science of human beings is "Artefact, not words". Language is important for science because science is a social enterprise and scientists need language to talk with each other – and this is why this book has been written – and also to talk with themselves, to think. But science is not language. How human beings call things is an important empirical phenomenon but philosopher John Austin's dictum that "words are our tools" is true for philosophy, not for science. Science is to go out of language. For scientists what is important is what they observe, count, and measure – without using words. If "a room without books is like a body without a soul", our science of human beings is a room without books and a body without a soul.

Theories expressed as artefacts are an important novelty for science. So far human beings have constructed artefacts to satisfy their practical needs and they have created artistic and symbolic artefacts. Artefacts as scientific theories are a novelty because they are constructed with the only goal to let us better understand reality. Scientific theories expressed as artefacts reduce the distance between science and technology, and this reflects a general trend: today, more and more, science is trying to understand reality by manipulating reality and by creating new pieces of reality. Theories as computer-based artefacts are an important step in this direction of "knowing by doing".

3. Robots as practical applications and robots as science

So, if we want to understand human beings, we should construct artefacts that behave like human beings. Since the advent of the computer there have been many attempts at constructing computer-based artefacts endowed with a "mind" and capable to exhibit "intelligence" but these artefacts did not have a "body". Behaviour is the result of the interactions of an animal's body with the physical environment and, therefore, our artefacts should have not only a "mind" but also a "body". A physical artefact which has a body and behaves like an animal or a human being is called a robot. So we must construct robots. In this book we are interested in human, not animal, robots: robots that behave like human beings and live in societies which are like human societies.

Robotics is a very active area of research today. Many people in many countries construct all sorts of robots and robots are a popular theme in the media. But what are robots? As we have said, a robot is a physical artefact which resembles an animal or a human being and behaves like an animal or a human being. This is an acceptable definition but, for our purposes, it is an insufficient definition. If we want to really understand what robots are, it is necessary to ask *why* we construct robots. Robots can be constructed with two different goals. Robots can be constructed because they are practically useful and have economic value. These are *robots as practical applications*. We have been familiar with industrial robots for decades and today there is much research aimed at discovering new applications of robots beyond industrial applications: home-cleaning robots, robots that assist old or ill people, robots in medicine and surgery, robots for agriculture, robots for surveillance tasks, robots for space exploration, robots with military applications, robots for education and entertainment.

The other goal of constructing robots is a purely scientific goal. Robots can be constructed to help scientists to understand and explain the behaviour of nonhuman animals and human beings. If a robot behaves like a real animal or a real human being, the manner in which the robot functions capture what lies behind the behaviour of the animal or human being and explains this behaviour. These are *robots as science*. It is important to distinguish between these two different goals for constructing robots because the two goals tend to push research in different directions. The practical goal and the scientific goal have, and should have, reciprocal links because robots that have practical applications can be a test for scientific theories and can suggest new hypotheses and pose new problems for science, and robots which are constructed for scientific purposes can suggest novel practical applications. But the two goals are different and they should be kept separate because for robots as practical applications and for robots as science success is measured by different criteria. For robots as practical applications, we should ask: Does this robot have, or could have, some useful application? For robots as science, we should ask: Does this robot make us better understand animals and human beings?

This is not what is happening today. Today, when one constructs a robot, in most cases it is not clear if his or her goal is to solve some practical problem or to better understand the behaviour of animals and human beings. This is a problem especially for robots as science because, today, almost all research money is for robots as practical applications. People who construct robots as practical applications are not particularly interested in understanding the behaviour of animals and human beings and they look at animals and human beings only for inspiration on how to construct robots which have practical applications. Robots that have practical applications have economic value, and even researchers who are not interested in the practical applications of robots tend to be guided in their work, perhaps unknowingly, by practical applications because only research on robots with practical applications gets funded. This is a problem because robots have a great potential for creating a new science of human beings, and the emphasis on robots as practical applications is an obstacle to exploiting this potential. If all research money is for robots as practical applications, the selection of research topics is guided by the possibility of applications, and this inevitably leads to ignoring many important behavioural and social phenomena that could be better understood by constructing robots that reproduce those phenomena. If robots must have practical applications, who would construct robots that make mistakes, robots that have pathologies, robots that sleep and dream, robots that have their own motivations and autonomously choose what to do or not to do, robots that can be intelligent but also stupid, robots that have different personalities and can be nervous or sad, robots that are unpredictable? These robots would not have practical applications (at least for the present time) and they might even pose problems and elicit negative emotional reactions. But real animals *do* exhibit these behaviours and we cannot claim we understand real animals, including human beings, if we do not understand these behaviours. Therefore, if our goal when we construct robots is to understand animals and human beings, the robots should reproduce these behaviours.

Another reason why current robots are mainly robots as practical applications is that robots are physical artefacts controlled by a computer or simulated in a computer. Therefore, engineers and computer scientists are an important and necessary component of the community of robotics. But engineers and computer scientists are trained to construct artefacts that solve practical problems and to design systems that behave "optimally" with respect to desired objectives. Robots as science do not have practical objectives and they should not behave "optimally" because human beings do not behave "optimally" or do not always behave "optimally". Robots as science are "pure" science, and only some engineers and computer scientists are ready to become "pure" scientists who have the only objective to understand and explain reality.

The prevailing interest for robots as practical applications explains why so much research in robotics is dedicated to the physical body of robots, to the shape of the body, to the materials of which the body is made, to the sensors and possibilities of movement

of the body. Of course, this research is important for robots as science because robots presuppose an "embodied" conception of behaviour according to which behaviour is the result of the physical interactions of the animal's body with the physical environment and the neural representation of the sensory input reflects both the sensory input and the motor output with which the animal responds to the sensory input. But, for robots as science, the robot's body can be very simplified with respect to the body of real animals and it may not be the principal emphasis of research. And, while robots as practical applications must necessarily be physically realized, this is not true for robots as science. Robots as science can be simulated in the computer and still they may provide us with useful insights on the behaviour of animals and human beings and on the organization of their societies.

Another difference between robots as science and robots as practical applications has to do with *how* robots are constructed. Robots can be programmed by us to do what they do and, in fact, most current robots are programmed by us. This may be a reasonable strategy for robots as practical applications. We have in mind one particular application and we program the robot so that the robot is able to do what it must do for that application. But programming a robot does not make much sense for robots as science. Robots as science must be like real animals and real animals are not programmed by anyone. Like real animals, robots as science must autonomously acquire whatever behaviour they possess as a result of various historical processes such as biological evolution in a succession of generations of robots, development and learning during life and, in the case of human robots, social and cultural change in robotic societies.

Even robots that evolve or learn can be robots as practical applications and not robots as science. If we are interested in robots as practical applications, our goal is that the robots possess some desired behaviour and, therefore, we evolve the robots or make them learn so that at the end they exhibit the desired behaviour. On the contrary, for robots as science, we are not only interested in making them acquire specific behaviours but we are also interested in the environment in which they acquire this behaviour, how they acquire the behaviour, what are the different components of their behaviour, what other behaviours they also acquire, how different are the behaviours of individually different robots, why some robots do not acquire the behaviour.

In this book we are not interested in robots as practical applications but we are interested in robots as science. For us, robots are tools (for doing science), not ends. As we have already said, there should be a dialogue between research on robots as science and research on robots as practical applications because the dialogue can be reciprocally useful. But, unless we say explicitly that we are interested in robots as science, the reader may not understand why we have constructed most of the robots described in this book.

4. One robot, many phenomena

Robots as theories of behaviour are a novelty for science and we are still trying to understand how to best use them to explain the behaviour of animals and human beings. One problem is that robots can be only "toys" which are interesting and suggestive but do not really tell us much about real animals and real human beings. Robots simplify with respect to real organisms but this is not a problem because robots are scientific theories and scientific theories necessarily simplify with respect to reality. Science is not a re-description of reality but it is an attempt at capturing the basic entities, mechanisms, and processes which underlie what is observed and explain what is observed. Science needs Occam's razor: "If there many theories that account for the same phenomenon, choose the simplest one." So, robots should be simplified artificial organisms.

But there is another principle of science that we should also follow to reduce the risk that robots are only "toys", and this principle is the following: "A theory which explains many different empirical phenomena is to be preferred to a theory which explains one single empirical phenomenon". We call this principle "one theory/many phenomena" – which, in the case of robots, becomes "one robot/many phenomena". Robots as practical applications reproduce one single behaviour because it is this behaviour which is the goal of the application. Robots as science should not reproduce only one behaviour but the *same* robot should reproduce *many* different behaviours and *many* different aspects of these behaviours. If a robot reproduces one single behaviour, the robot is more likely a "toy" which is not really useful to understand real animals and real human beings. If one and the same robot reproduces many behaviours and many aspects of these behaviours, we can be more confident that the robot captures reality and tells us what animals and human beings really are.

A classical example – although not a robotic example – is programming a computer so that the computer appears to be able to talk with us. We say something to the computer and the computer responds appropriately. The computer asks questions which seem to be natural and we respond to these questions. But the computer does not really understand what we say and even what *it* says. To claim that the computer has language, the computer should be able to use language in many different ways and in many different contexts. It should use language when it interacts with the physical environment, it should talk spontaneously, it should say different things to different people aside from what these people ask the computer, and it should not be programmed to talk but it should autonomously learning to talk. So the talking computer is only a "toy", although an interesting and entertaining toy.

The reason is simple. Returning to robots, if a robot reproduces one single behaviour and one single aspect of this behaviour, there may be many different robots that reproduce the same behaviour or the same aspect of this behaviour and deciding among these different robots is arbitrary. This is what happens in today's robotics where

each researcher (or research group) has his or her robot and it is almost impossible to reach a consensus on which robot is the right one or the best one. On the contrary, if one and the same robot reproduces many different behaviours and many different phenomena concerning the behaviour of animals and human beings, it is more difficult to construct different robots which reproduce all these behaviours and all these phenomena, and we are more justified in thinking that the robot is not a "toy" and that it actually captures what lies behind the behaviour of animals and human beings and explains this behaviour.

Here is a list of different phenomena that one and the same robot or group of robots should be able to reproduce.

Body

Behaviour is movement of the different parts of the body in response to the stimuli provided by the sensors of the animal's body and, necessarily, animals which have different bodies have different behaviours – and also different "minds". Psychologists tend to study "cognition" or "mind" without taking the body into consideration and only recently – partly, under the influence of robotics – "embodied" theories of cognition and mind have made their appearance. Social scientists tend to completely ignore the body of human beings. And the body is also ignored in other attempts at studying mind and society by constructing "disembodied" computational artefacts such as those of artificial intelligence, the neural networks of classical connectionism, and agent-based social simulations. Robots are a first step in the direction of "one theory/many phenomena" because they reproduce not only behaviours but also bodies. In fact, robots are the fourth revolution in how human beings see themselves. The first revolution was the Copernican revolution: the place where human beings live, the Earth, is not the centre of the Universe. The second one was the Darwinian revolution: human beings have not been created by God but they have been created by nature. The third revolution was the Freudian revolution: most of the human mind is not conscious. Now a fourth revolution has begun, the robotic revolution: human beings are only bodies.

The body is not only the external body – the size and shape of the body and its sensory and motor organs – but it is also the body's internal organs and systems. One particular piece of the body, the brain, is especially important because it directly controls behaviour. However, behaviour is not only under the control of the brain but it is also influenced by other internal organs and systems such as the heart, the gut system, and the endocrine system. Current robots have an external body but they do not have internal organs and systems. Future robots will need to have organs and systems inside their body because these internal organs and systems are crucial to reproduce (explain) many aspects of behaviour and especially its motivational and emotional aspects.

Brain

Most animals have one particular organ inside their body which directly controls their behaviour: the brain. The principle "one robot/many phenomena" requires that the robots should not only behave like real animals but their behaviour should be controlled by an artificial brain which is like the real brain. Artificial brains are called neural networks. As the real brain is made up of neurons which can be more or less active and synapses through which a neuron increases or decreases the activation of another neuron, so neural networks are made up of artificial neurons and artificial synapses called connections. A neural network is basically made of sensory neurons which are activated by states or events in the external environment or inside the robot's body, motor neurons whose activation determines the movements of the robot's body, and internal neurons to which the sensory neurons send their connections and which, in turn, send their connections to the motor neurons. Behaviour is the mapping from specific sensory inputs to specific motor outputs and this mapping – how the robot responds to the different sensory inputs – depends on the quantitative strength (weight) of the connections that link the different neurons.

Robots with a brain – neuro-robots – obey the principle "one robot/many phenomena" because they generate predictions that can be confirmed or rejected on the basis of *two* classes of empirical data: behavioural data and neural data. A robot should not only behave like some real animal but the structure and functioning of the robot's neural network should also reflect the structure and functioning of the animal's brain and the robots should reproduce not only the phenomena studied by psychologists but also those studied by neuroscientists. And since the brain can be malfunctioning and produce neural and behavioural pathologies, the principle "one robot/many phenomena" should be extended to neural and behavioural pathologies. Once we have constructed a neuro-robot, we can lesion different parts of the robot's neural network and the brain-lesioned robot should replicate the different behavioural pathologies observed in animals and human beings.

Environment

Behaviour is not a property of an animal but it is a result of the interactions of the animal with the environment. The environment is both the physical, "objective", environment and the environment as filtered by the sensory and motor organs of the animal – what can be called the animal's behavioural environment. Two animals can live in the same physical environment but, if they have different sensory and motor organs, they live in different behavioural environments. And the stimuli which arrive from the environment are not independent from the animal's behaviour but they are self-selected by the animal's behaviour.

Psychologists have difficulty recognizing that they cannot ignore the environment in which the animals live. This is due to the historical traditions of the discipline but also to practical and technical reasons. Studying the behaviour of an animal in its natural environment is very difficult and expensive in terms of both time and money – and, in many cases, impossible. Furthermore, in the natural environment the scientist cannot control and manipulate the conditions in which the animal exhibits its behaviour, and this is a serious problem because scientists must control and vary the conditions in which they observe the phenomena that they observe if they want to find the causes of these phenomena. For all these reasons, most of the empirical data on the behaviour of animals and human beings that psychologists have collected since the birth of their discipline are data obtained not in the natural environment but in the experimental laboratory, in which a great number of precise and quantitative data can be obtained in relatively little time and everything can be controlled and manipulated by the experimenter. Laboratory experiments are useful and necessary but, if what one is studying is behaviour, they have three limitations. First, they tell us what are the immediate causes of behaviour but not its more remote causes – which can be more revealing than the immediate causes and can explain the immediate causes. Second, laboratory data may be unnatural and, in some cases, misleading because the behaviour of animals – and, especially, the behaviour of human beings – in the experimental laboratory may not be the same as their behaviour in the real environment. Third, many human phenomena cannot be studied in the experimental laboratory because they involve too many people, they happen in too large spaces, and they last too much time – and this is why the social sciences do not do laboratory experiments and, when they do laboratory experiments, the results of these experiments tell us even less about human societies than psychological experiments about individual behaviour.

This has many implications if robots must obey the principle "one robot/many phenomena". Robots cannot only be artificial bodies and artificial brains but they must also live in an artificial "natural environment". Robotics must be an environmental robotics because only an environmental robotics can make us understand how the behaviour of a robot depends on the particular environment in which the robot has lived and lives and how the robot's behaviour varies with the different circumstances in which the robot finds itself during its life. In addition, an environmental robotics allows us to make "environmental experiments" in which we change the environment in which the robots live and we determine the consequences of living in different environments for the robots' behaviour. Environmental experiments cannot be done with real animals because it is impossible to manipulate the natural environment of real animals but we can manipulate the environment of robots, and environmental experiments on robots can help us to interpret the (scarce) environmental data that we possess on real animals and may suggest what new data to look for.

Studying the behaviour of robots in their environment is especially important if we want to construct human robots. The environment of human beings is mostly a

social environment and most of the behaviour of human beings is influenced by the behaviour of other human beings. Furthermore, unlike nonhuman animals, human beings modify the environment in which they live and their "natural" environment has progressively become an "artificial" environment which has been created by them and continues to be modified by them. In fact, the behaviour of human beings is the result of a two-way interaction with the environment: the environment determines their behaviour and their behaviour determines the environment. An environmental robotics should help us understand this two-way interaction.

But, based on the principle "one robot/many phenomena", we should also do "laboratory experiments" with robots. The same robot that lives in its "natural" environment is brought into an "experimental laboratory", a simplified environment in which everything is controlled by us, and we collect data on the robot's behaviour in this simplified environment. This bi-directional link between the robots' natural environment and the laboratory may suggest answers to many interesting questions concerning real organisms: Why this organism – animal or human being – behaves in this way in the laboratory? What should be the natural environment of an organism which behaves in this way in the laboratory? How an organism which lives in this natural environment will behave in the laboratory? Which laboratory experiments should we do with organisms that live in this type of natural environment?

Behaviour as the product of historical processes

Human beings – like all organisms – are the product of historical processes. They become what they are and we cannot really understand what they are if we do not understand how they have become what they are. Therefore, if we want to construct human robots, our robots must not only reproduce how human beings behave and how their societies function but they must also reproduce the processes through which human beings have acquired their behaviour and human societies have become what they are. Human beings are the product of many different "histories": they are a product of biological evolution which takes place in a succession of generations, they are the product of a process of development which takes place during their life, they are the product of the specific experiences that they have during their life, and they are a product of their cultures, their societies, their technologies and of the history of their cultures, societies, technologies. Human robots should reproduce all these histories.

Today many robots are not programmed by us but they acquire their behaviour through evolution, development, or learning. Evolutionary robotics works with populations of robots that inherit from their parents a set of genes which encode the properties of their neural network – in most cases, the neural network's connection weights – and, therefore, determine the robots' behaviour. Different robots have different genes and, therefore, each robot has a different neural network and it behaves differently from all other robots. The robots reproduce selectively – some robots have

more offspring than other robots – based on how well they behave. Some random variation is added to the genes that the offspring robots inherit from their parents (genetic mutations) so that, in some cases, the offspring robots have better genes and behave better than their parents. The robots of the initial generation have random genes and, therefore, they are not very good at doing what is necessary to remain alive and reproduce. But since the robots that reproduce are those that behave better than the other robots, the result is that in a succession of generations the robots evolve initially nonexistent behaviours.

This is evolution. Development is a program of changes that take place during the life of a robot and which is encoded in the robot's genes. Evolution creates genes that do not only encode the "initial state" of the robot but also how the robot's body and brain change during the robot's life. Other robots acquire their behaviour through learning. A robot that learns has a neural network which changes as a result of the robot's experiences in its environment so that the robot acquires new behaviours and new capacities.

These different types of "robotics of change" may be sufficient for nonhuman animals but, if our goal is to construct human robots, we must also reproduce other processes of change which are specifically human. Human beings have societies, cultures, and technologies which change in the course of time and, since their behaviour is influenced by these societies, cultures, and technologies, human robots must live in robotic societies and have robotic cultures and technologies which change in the course of time. This requires that human robotics establishes a dialogue not only with evolutionary and developmental biology and with the psychology of learning, but also with the social sciences and with archaeology and history.

Here a comment on the expression "constructing robots" is necessary. If our goal is to understand the behaviour of human beings by constructing robots that behave like human beings, there is a problem with the idea of "constructing" robots. Real organisms are not constructed by anybody. They are self-constructed. They are the result of historical processes (evolution, development, learning and, in the case of human beings, the history of their societies and their cultures) which have made them what they are. So, if robots must reproduce real organisms, the robots cannot be constructed by us. They must construct themselves and the objective of reproducing the different processes through which they acquire their behaviours is as important as the objective of reproducing these behaviours. Robots are an example of a new, general, approach to doing science, an approach which is expressed by the formula "If you want to understand X, you must reproduce X in an artefact". But organisms are biological and historical phenomena and, for biological and historical phenomena, the formula necessarily becomes "If you want to understand X, you must not only reproduce X in an artefact but also how the artefact has autonomously become what it is".

This may appear to be in contrast with the fact that robots are computer-based artefacts because the idea of a computer is associated with the idea of a program. Computer programs are lists of instructions that a computer must execute to produce

a certain result. If *we* write the program and what the robot does is simply the result of executing the program, in what sense is the robot self-constructed? The answer is that we should not program our robots so that they exhibit desired behaviours but we should write a program that allows the robots to evolve, develop, or learn the behaviours they exhibit. And here, as elsewhere, robots as scientific theories are different from robots as practical applications. With application-oriented robots we already know what is the behaviour that our robots should be able to exhibit. Robots as scientific tools are different. We create the initial conditions for our robots to evolve, develop, and learn, and then we see *what* are the behaviours that emerge and *how* they emerge. These behaviours may surprise us and, in many cases, robotic behaviours that surprise us are more interesting and tell us more about real animals and real human beings than behaviours which we expect or we want our robots to exhibit. And the strategy is a recursive one. We create the initial conditions for X to emerge and then we make another step back and we create the initial conditions for these initial conditions to emerge, and so on going back in time – until we reach the beginnings of life on Earth.

Comparative robotics

Human beings are different from nonhuman animals but everything which appears to our eyes as uniquely human has a beginning or exists in some simpler form in other species. Therefore, to construct human robots we must adopt a comparative perspective and must compare human robots with animal robots. A comparative robotics can be as useful as comparative biology and comparative psychology are useful for the study of real animals. We must construct human robots but also robots that resemble and behave like the small worm Caenorhabditis elegans or a fish or a mouse or a monkey, or even plant robots. And we must compare these robots and discover how their evolutionary history has made them what they are so that we can explain both their differences and similarities. And the comparative method should be extended to human societies. Human societies have different cultures, different economies, different political systems. We must construct robotic societies with different cultures, economies, and political systems, and then we must compare these robotic societies to better understand how each component of a human society interacts with the other components and how the different components condition each other. Biological and social entities can be simpler or more complex and, to understand the more complex ones, it is often useful to compare the more complex ones with the simpler ones.

But there is something else which can be done with a comparative robotics and which is impossible to do with a comparative biology and a comparative psychology and also with a comparative social science. We can construct *counterfactual* robots, robots which are not like any existing animal, and *counterfactual* robotic societies which are not like any human society which exists today or has existed in the past. These counterfactual robots and counterfactual robotic societies are like counterfactual

hypotheses, hypotheses about possible but nonexisting worlds. The extension of the comparative method to nonexisting worlds which is made possible by expressing theories as computer-based artefacts is an important addition to the tools of science. Before the computer, counterfactual hypotheses – or, in social and political life, "utopias" – could only be imagined or described in words, and their consequences could only be "calculated" in our mind. The computer changes this. Counterfactual hypotheses become worlds which *do* exist in the computer or in a computer-controlled physical artefact, and we can observe these worlds and do experiment on them as we observe the real world and do experiments with the real world. Theories of behaviour must generate predictions to be compared with the behaviour of real animals. Theories expressed as robots generate predictions to be compared with the behaviour of real animals but also with the behaviour of the robots themselves. Can we predict how the counterfactual robots we have constructed will behave? Can we predict how they will behave in circumstances which we did not anticipate when we constructed them? Can we predict what will happen in a counterfactual robotic society? Robots are not only theories but they are also (artificial) data. Traditionally, theories and experiments are separate entities. Theories are *in* the mind of the scientist, in his/her words or in mathematical symbols. Experiments interrogate reality which is *out there*. In the new "artificial" science of robotics, theories and the data the theories are intended to explain may be both *in* the artefact, and artefacts can be at the same time theories and experimental laboratories. Real experiments and, more generally, real empirical data, obviously are still necessary. But now science has a new tool, and this tool can be very useful in the hands of a science which studies very complex phenomena such as human beings and their societies. If we construct a human robot which is like a real human being but it has no hands, can we predict how will the robot behave? If we construct a human robot which lives alone in its environment, can we predict what will be the robot's behaviour and what will be its "mind"? If we construct human robots which are not males and females and do not live in families, can we predict how the robots will feel? If we construct societies of human robots which do not have a "chief", can we predict what will happen to those societies? If we construct robotic societies with an economic system which is different from any existing economic system, can be predict what the results will be? Of course, these different components of the human adaptive pattern may be connected together so that, for example, it may be impossible to construct a robot without hands which can be called "human". But this by itself can be an interesting result and, more generally, counter-factual robots and counter-factual robotic societies can be useful to understand real human beings and real human societies – and they can also be useful to understand and predict their future.

So, if we want to apply the principle "one robot/many phenomena", the robots must have a body and a brain, they must live in an environment, they must be the results of various historical processes, and they must reproduce both human beings and other animals. These are the most basic phenomena. But if we want to specifically apply

the principle to human robots, there is a long list of different behaviours and social phenomena that one and the same robot or one and the same collection of robots should to be able to reproduce. We mention this variety of human behaviours and human social phenomena in the next section but they are the content of the different chapters of this book.

5. Human, not humanoid, robots

Although comparative robotics is important, in this book we are interested in human robots – robots that behave like human beings and live in societies that are like human societies. Human robots still do not exist. What exists now are "humanoid" robots. Humanoid robots are robots that externally look like human beings, exhibit some simple human behaviours such as reaching and grasping an object with their hand or walking and climbing stairs with their two legs, and in some cases have a language or emotional expressions with which they communicate with us. Having a body which resembles the body of a really existing animal is an important requirement for robotics but it is not sufficient to interpret a robot as a theory of the behaviour of the animal because behaviour depends on the body but it is *not* the body. The behaviours exhibited by today's humanoid robots are interesting behaviours but they are only a very small part of the rich repertoire of behaviours possessed by human beings, especially if we consider the importance of social life for human beings. And humanoid robots *seem* to communicate with us and they *seem* to express emotions but they do not really understand what we say, and even what *they* say, and they do not really feel the emotions that they express and do not really understand the emotions that we express.

So human robots are future robots. Very little research is being done on human robots. Robots are practical applications and robots as practical applications have an appeal for engineers and computer scientists but much less for students of human behaviour and human societies. The only students of human behaviour who are currently involved in robotic research are (very few) psychologists and neuroscientists, while social scientists almost completely ignore robots, although some of them try to understand social phenomena by constructing disembodied and abstract "agents".

Many reasons explain this situation. Constructing robots and, more generally, computer-based artefacts presupposes technical abilities such as writing computer programs which are not taught to students of human beings and human societies. The very idea of trying to understand human beings and human societies by reproducing them with robots is completely extraneous to students of human beings and human societies. And robots as intended in this book ignore disciplinary divisions while the science of human beings is still strongly disciplinary.

But if we want to understand human beings and human societies, we must construct robots that behave like human beings and live in societies that are like human

societies. And, if we want apply the principle "one robot/many phenomena" to human robots, we must construct robots that, like human beings, have a language which they use not only to communicate with other robots but also to communicate with themselves, have a mental life, learn most of their behaviours from other robots, do not live in the natural environment but create the environment in which they live, obtain most of the things which they need not from nature but from other robots, have societies as complex as human societies, have art and religion, and do philosophy and science.

Why are we interested in constructing human robots, rather than worm robots or monkey robots? There are many answers to this question. One is that human beings are especially complex animals and, therefore, they represent a greater challenge for a robotic approach to the study of behaviour. If we are able to construct robots that behave like human beings, we have more clearly demonstrated the power and usefulness of the approach than if we construct worm or even monkey robots. But human beings are not only more complex animals. They exhibit behaviours which are nonexistent in nonhuman animals. Some rudimentary forms of these behaviours can be found in this or that animal because, as Darwin said, "the difference in mind between man and the higher animals, great as it is, certainly is one of degree and not of kind". And this is why, even if we are interested in human beings, we cannot ignore other animals and we must understand the "animal" past of human beings. But, although all animals are special, human beings are *more* special.

Another answer to the question why we want to construct human robots is that both the author and the readers of this book are human beings and we are victims of the narcissism of the species. Following Socrates' advice, we want to know ourselves. But it is not only the narcissism of the species. We want to know ourselves by constructing human robots because human robots may help us to understand and to find solutions for the many difficult problems that human beings face today. Most of these problems are created by human beings themselves. Geologists say that 10.000 years ago a new geological era has begun in which the physical environment is not the natural environment but it is an environment created by man, and they call this geological era the Anthropocene (from the Greek "anthropos") – although the Anthropocene should be interpreted not only as a physical environment but also as a mental and social environment. If robots can help us to better understand ourselves and the environment created by us, we should invest in robots as science to better understand both our present condition and our future – and this might be the most important practical application of robots.

But in the last few decades the Anthropocene has become the Computercene, an era in which the environment in which human beings live is created by the computer – *is* the computer. And the Computercene might be followed by the Robocene. If we are able to construct really human robots, our environment will contain artificial human beings, and it is not clear what will be for human beings to live with artificial human beings. Human robots can be psychologically disturbing for human beings, although

in the long run human beings may adapt to them and accept them. But there is a more serious problem. Since human beings decide what they do, if human robots must really be like human beings they also must decide what they do, without being told by us. And they might decide to make us their "serfs". ("Serf" is the original meaning of the word "robot" but now the serfs would be human beings.) Or they might decide to live *without* us – and they might able to realize their desires. This may sounds like science fiction but it is the logical consequence of a program of research. Either it is *impossible* to construct really human robots or we must imagine a world with robots which are *really* like us. Of course, human beings might decide not to construct robots which are really like them and, in this case, they should decide where to stop – and perhaps this book should not have been published.

6. This book

Human robots still do not exist but we can make the first steps towards the construction of human robots. We can outline the range of behaviours and social phenomena that human robots should reproduce, we can identify the problems that must be solved, and we can construct some extremely simplified human robots. This is what we try to do in this book.

This book has many limitations. It describes the results of research done by people who still work or have worked in the past at the Institute of Cognitive Sciences and Technologies of the National Research Council in Rome – although the book reflects the point of view of its author on this research and on robotics more generally. So the book is a summary of a program of research rather than a statement of the state of the art. The robots described in the book are extremely simplified in their body, their brain, their environment, their evolutionary history, their life history, their motivations, their behaviours, their social life, and their societies – and in many cases they behave like nonhuman animals rather than like human beings. Furthermore, they are simulated in a computer and not physically realized, although most of our robots are small physical robots called "e-puck" robots, and the computer program that controls the behaviour of a simulated e-puck robot could be transferred to a physical e-puck robot and we might see how the physical robot behaves in the physical environment. The robots have a body which is very different from the human body but, if we must choose between a humanoid robot which looks like a human being but replicates only one or two very simple human behaviours, and a robot which does not have a human-like body but exhibits behavioural phenomena that resemble important human behavioural phenomena, our choice goes to the latter type of robots.

Another limitation is that, in some cases, our "robots" are not really robots because they do not have a body and a brain but they are like the disembodied "agents" of

agent-based social simulations. Agent-based simulations are very useful tools but we believe that the body and the brain – and, more generally, the biology of human beings – cannot be ignored if we want to understand both human behaviour and human societies. Therefore, our "agents" should be progressively replaced by robots.

A final limitation is that, although most of the robots described in this book have been really constructed and we synthetically report the results that have been obtained, in a few cases we describe robots that have not been constructed but could be easily constructed. (We return to the limitations of our robots in the last chapter of the book.)

And now a few words on how the book is structured. This Chapter describes the general "philosophy" of our program of research. Chapter 2 is dedicated to robots that have different motivations and, to decide in each moment which motivation to satisfy with their behaviour, they have emotions which help them to make better motivational decisions. Chapter 3 describes robots that acquire their behaviours because they both evolve in a succession of generations and learn during their life. These two chapters address behavioural phenomena which are very important but are not typically human because they can be found in all animals. The remaining chapters are dedicated to phenomena that exist in some embryonic form in nonhuman animals but tend to be specifically human. Chapter 4 describes robots that have language and their language changes the model of the world that they have in their brain. Chapter 5 is dedicated to robots that have a mental life in the sense that their brain self-generates its own sensory inputs and responds to these self-generated sensory inputs and not only to the sensory inputs from the external environment or from their body. The successive Chapters are dedicated to social life and human societies. Chapter 6 addresses some of the advantages and disadvantages of living together. Chapter 7 describes kin-related robots and robots that are males or females and live in families. Chapter 8 is dedicated to robots that learn by imitating other robots and develop cultures and technologies. Chapter 9 describes robots that own goods in the sense that they put their goods in external stores to which they only have access. Chapter 10 is dedicated to communities of robots which have a political life and a chief that decides for the entire community. Chapter 11 is dedicated to robotic economies. It describes the exchange of goods, merchants as intermediaries in the exchange of goods, the emergence of money, and the exchange of work for money between worker robots and entrepreneur robots. Chapter 12 is concerned with individual differences among robots and with robots which have neurological or psychiatric pathologies. Chapter 13 is dedicated to robots that have art and religion, to robotic societies which have a history, to robots that do science, and to doing metaphysics with robots. Chapter 14 lists the limitations of our robots and what must still be done. And the book ends with a final chapter, Chapter 15, in which we discuss how human robots can be useful to human beings.

Appendix

Almost all the robots described in this book have been actually constructed and one can find more details on the robots in the papers listed at the end of each chapter. Most robots are simulated e-puck robots. An e-puck robot is a physical robot with a small body, visual and other sensors, and two wheels which allow the robot to move in the environment (see Figure 1).



Figure 1. The e-puck robot.

In our basic scenario the robot lives in a walled environment that contains elements of food – food tokens – which the robot must be able to eat to remain alive and have offspring (see Figure 2).

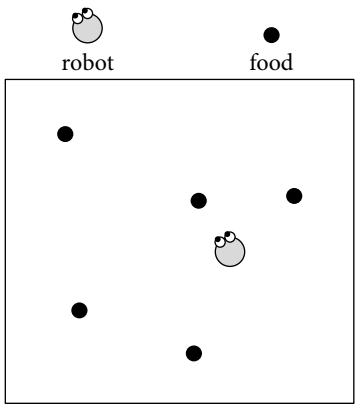


Figure 2. A robot lives in an environment which contains five food tokens.

When the robot touches a food token, the food token is considered as eaten by the robot and another food token appears in the environment so that the total number of food tokens is always the same. There is an entire population of robots each with a different brain and a different behaviour but each robot lives alone in its copy of environment.

The behaviour of the robots is controlled by an artificial brain – a neural network with sensory neurons, internal neurons, and motor neurons linked by connections. The sensory neurons send their connections to the internal neurons and the internal neurons send their connections the motor neurons and the connections can be excitatory or inhibitory (see Figure 3).

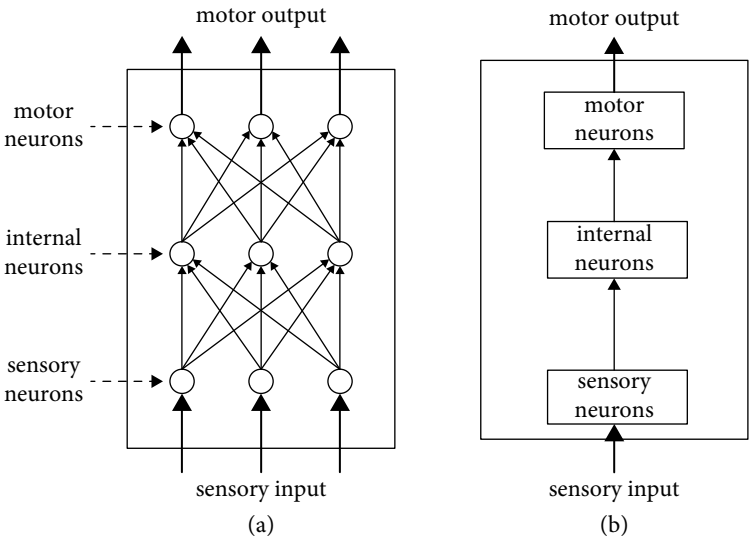


Figure 3. The robot's brain is an artificial neural network with sensory neurons which send their connections – the arrows – to the internal neurons which, in turn, send their connections to the motor neurons (a). (b) A more schematic representation of (a).

The life of a robot is a succession cycles. In each cycle each neuron has a level of activation which is different in different neurons and the ensemble of activation levels of a group of neurons is called the pattern of activation of that group of neurons. The pattern of activation of the sensory (visual) neurons is determined by the sensory input from the food tokens which are within the robot's visual field (see Figure 4). The sensory neurons excite or inhibit the internal neurons and the resulting pattern of activation of the internal neurons is called the internal representation of the sensory input. This pattern of activation causes a pattern of activation in the motor neurons which determines the speed of the robot's two wheels and, therefore, the robot's movements in the environment. The internal representation of the sensory input reflects both the

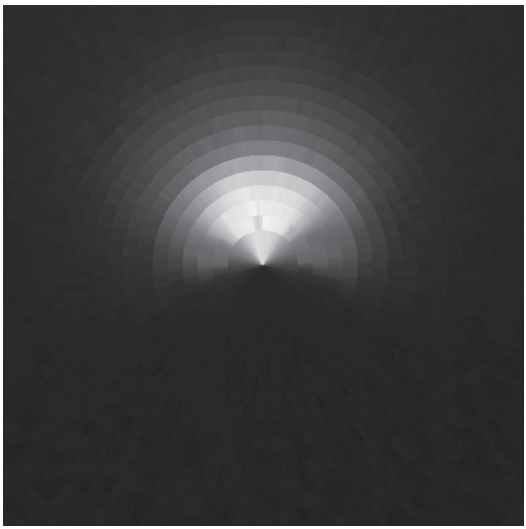


Figure 4. The robot's visual field.

sensory input and the motor response with which the neural network responds to the sensory input and, therefore, it is a sensory-motor representation.

If the visual neurons are not activated because no food token is within the robot's visual field, the robot can equally move and explore the environment because the internal and motor neurons have an intrinsic level of activation which is independent from, and added to, the activation arriving from other neurons.

How much one neuron influences – excites or inhibits – the activation of another neuron depends on two factors: (1) the level of activation of the first neuron and (2) the weight of the connection which links the first neuron to the second neuron. Therefore, if two robots have different connection weights in their neural networks, the two robots respond in different ways to the same sensory input.

The behaviour of the robots is genetically inherited (innate) and it evolves in a succession of generations. Each robot has a set of genes which encode the connection weights of its neural network and the intrinsic level of activation of its internal and motor neurons and, therefore, determine the robot's behaviour. We begin with a first generation of robots with random genes and, therefore, random connection weights in their neural network. Each robot behaves differently from the other robots but the robots are generally unable to reach and eat the food tokens. The robots reproduce nonsexually – only one robot is necessary to generate an offspring robot – and the robots that, for purely chance reasons, have better genes (better connection weights in their neural network) and, therefore, eat more, have more offspring. An offspring robot inherits the genes of its parent with some random variation in the value of the genes (genetic mutations) and, therefore, the offspring robots do not behave exactly

like their parents. In some cases, an offspring robot may be better than its parent at reaching and eating the food tokens and this robot, not its siblings, will probably be selected for reproduction. The process goes on for a certain number of generations during which the robots' behaviour becomes progressively better and, at the end, the robots have acquired the capacity to eat the food tokens – although each robot is still different from all other robots. The robots that reproduce are said to have more "fitness", where fitness is the total number of food tokens eaten during a life which has the same length for all robots. To have statistically significant results, we repeat the simulation a certain number of times starting with a different set of initial connection weights.

Figure 5 shows how the fitness of the robots increases in a succession of 500 generations. The best robot of the first generation eats around 70 food tokens and the average robot only 30 food tokens. The best robot of the last generation eats around 225 tokens and the average robot 175 tokens.

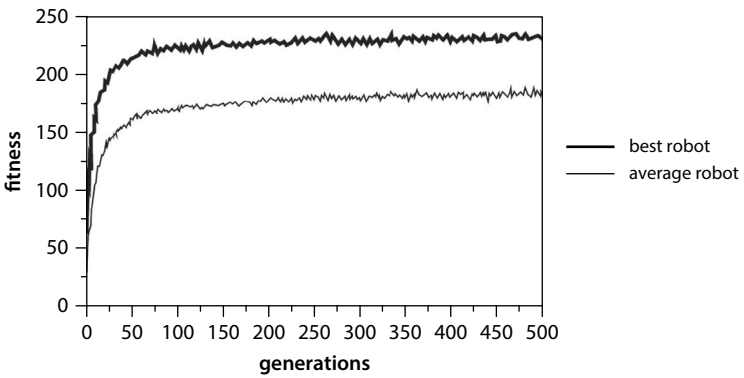


Figure 5. Increase in the fitness of the robots in a succession of 500 generations. Fitness is the number of food tokens eaten during a life which has the same length for all robots.

These are our basic robots. But the robots described in the book may live in different environments, they may have other problems to solve beyond eating, they may have more complex neural networks, their genes may encode something else in addition to the connection weights of their neural networks, they may both evolve in a succession of generations and learn during their life, and they may learn both by interacting with the physical environment and with other robots. And they may also be robots different from the e-puck robots or they may "agents" without a body and a brain and with an abstract behaviour. When necessary, in the course of the book we will add the necessary details.

CHAPTER 2

Robots that have motivations and emotions

When it arrives on the Earth, ME discovers that human scientists tend to identify the human mind with reason and intellect. Even Freud, the only significant psychologist who gave the motivational half of the mind prominence over the cognitive half, thought that healing from psychological disturbances is only possible if the motivational half is brought under the control of the cognitive half. And ME also finds that the science of the mind tends to be called "cognitive" science. The "mind" has become "cognition".

Science is the product of one specific human culture, Western culture, and in Western culture cognition is more important than motivations and emotions because cognition is control of reality and independence from nature whereas motivations and emotions are lack of control and dependence from nature.

ME is born on Mars and, although it is a scientist, it is not a human scientist and it knows that reason and intellect are only one half of the human mind – the cognitive half – and that the other half is motivations and emotions. And it tends to agree with the French philosopher La Rochefoucauld that "the mind is always a victim of the heart". ME wants to understand human beings by constructing robots that behave like human beings but, while the robots constructed by human beings possess only the cognitive half of the mind, ME's robots must possess both its cognitive half and its motivational/emotional half because only in this way they can reproduce such important human phenomena as the manner in which human beings psychologically live their life, what other human beings are for them, why most social behaviour is changing the motivations of others, why human beings differ not only in their cognitive capacities but also in character and temperament, the great variety of psychological pathologies that afflict them, and why they have art and religion.

1. The cognitive level and the motivational level of behaviour

Behaviour tends to be explained as caused by external stimuli. External stimuli arrive to the sensory organs of an animal and the animal's brain responds by telling the different parts of the body how to move and what to do. What happens in the brain between the stimuli and the responses can be very complex but external stimuli are still seen as the main determinants of behaviour. This picture of behaviour is, at best, incomplete.

Imagine an animal which sees a piece of food. The animal responds to the food by eating the food only if it is hungry. If it is not hungry, the animal ignores the food. This shows that external stimuli are not the only causes of behaviour. An explanation of the animal's behaviour requires that we also take into consideration the motivation which at any given time controls the animal's behaviour. It is the animal's current motivation – in this case, hunger – together with the external stimuli, that makes it possible to predict what the animal will do. External stimuli are not enough.

All animals have many different motivations which they must satisfy with their behaviour if they want to remain alive, reproduce, and live well. They have the motivation to eat, the motivation to drink, the motivation to avoid predators, the motivation to find a partner for reproduction, the motivation to insure the survival of their offspring, the motivation to heal when their body incurs some physical damage. For human beings, the number of different motivations is much larger. Human beings have all the motivations of nonhuman animals and many others: they want a multitude of other things beyond food and water – dresses, houses, cars, personal ornaments, mobile phones – and they want money with which they can buy all these things. They have the motivation that other people have a certain idea of them, the motivation to avoid bad thoughts concerning the future, and the motivation to know and understand reality to construct useful technologies but also as an end in itself.

The existence of different motivations poses a problem because, in most cases, an animal cannot satisfy two or more motivations at the same time. If an animal wants to both eat and avoid being killed by a predator, in some circumstances the animal must choose between looking for food and fleeing away from the predator. The same if the animal wants to eat and to find a mate for reproduction. It cannot do the two things at the same time. If a human being wants to make a lot of money but also wants to project an image of himself or herself as a person not interested in money, he or she must choose between making a lot of money and projecting an image of himself or herself as a person not interested in money.

This requires that all animals possess a mechanism for deciding which motivation should control their behaviour at any given time. And, in fact, all animals function at two levels: they function at the motivational level and they function at the cognitive or behavioural level. At the motivational level the animal decides which one of its many different motivations it should try to satisfy in each particular moment with its behaviour. At the cognitive level the animal executes the behaviour that allows the animal to satisfy the motivation decided at the motivational level. To survive and reproduce an animal must function well at both levels. The cognitive level is important because it is the actual level of behaviour and, clearly, if an animal in unable to do what is necessary to satisfy its motivations, this is a problem for the animal. If the animal decides to ignore food and to flee away from a predator when a predator appears but the animal is not very fast at running away, the animal is in serious trouble. But the motivational level is even more important. If the animal continues to eat when the predator appears, the animal will certainly die.

There is another difference between motivations and behaviours. How an individual functions at the behavioural level can be improved by learning. If an individual does not know how to do X, it can learn to do X. On the contrary, the capacity to decide correctly which motivation to try to satisfy with one's behaviour is more difficult to improve – as psychotherapists know very well. The motivations of an individual and the behaviours that the individual must posses in order to satisfy its motivations have different origins. The motivational system is a result of the evolutionary history of the species and is encoded in the genes which each individual inherits at birth. On the contrary, the behaviours that allow an individual to satisfy his or her motivations, at least in human beings are mostly learned during the individual's life. Of course, human beings do learn many new motivations during their life but all human beings are born with a set of basic motivations which are the result of the evolutionary history of their species.

This has an important implication for how to construct robots. If the robots must resemble real animals, they cannot be programmed by us because real animals are not programmed by anyone. They must spontaneously become what they are and, like real animals, they must become what they are as a result of two different processes: evolution and learning. They must evolve in a succession of generations and they must learn from the experiences that each individual robot has during its life. But evolution and learning do not play the same role with respect to the motivational level and the cognitive level of functioning of their brain. A robot can in principle acquire all its behaviours and all its cognitive capacities through learning but the robot cannot acquire its basic motivations unless we construct a population of robots that evolve these motivations as part of their adaptation to the environment.

As we have said, generally an animal cannot satisfy two or more motivations at the same time. Therefore, the animal needs a mechanism for deciding which motivation to satisfy with its behaviour in each particular moment. The mechanism is a very simple one. Each motivation has a quantitative strength which varies from individual to individual and from moment to moment and the motivation which wins the competition and guides the animal's behaviour is the motivation which in that moment has the greatest strength. The strength of a motivation depends on a variety of factors. Some motivations are intrinsically stronger than other motivations. The motivation to eat, the motivation to avoid dangers, the motivation to reproduce, the motivation to keep one's offspring alive have a high intrinsic level of strength. But the strength of a motivation can also depend on the current state of the animal's body. The motivation to eat is stronger if there is little energy in the animal's body and the animal's brain is informed of this fact under the form of hunger. Or the strength of a motivation can depend on the stimuli that currently arrive to the animal's brain from the external environment. The motivation to avoid predators becomes very strong when the animal sees a predator and the motivation to mate is intrinsically strong but it becomes stronger when the animal sees an individual of the opposite sex. For human beings everything is more complicated because, unlike nonhuman animals, human beings self-generate all sorts

of stimuli inside their brain and the strength of their motivations is influenced by these self-generated stimuli – by what they imagine, remember, and think. (For robots that self-generate their stimuli and have a mental life, see Chapter 5.)

We have said that an animal "decides" which motivation controls its behaviour at any given time but nonhuman animals and, in most cases, also human beings, do not actually "decide" which motivation they should try to satisfy with their behaviour. The different motivations compete with each other inside the animal's body and brain and the motivation which has the greatest strength automatically wins the competition and controls the animal's behaviour.

2. Today's robots do not have motivations

If it true that animals and human beings function at two levels, the motivational level and the cognitive level, our robots must also function at these two levels. A robot must have different motivations, it must "decide" which motivation controls its behaviour at any given time, and then it must execute the behaviours that make it possible to satisfy this motivation.

Today's robots do not have motivations and do not choose among different motivations. Almost all robotic research is dedicated to the cognitive level of behaviour while the motivational level tends to be ignored. One often hears of "autonomous" robots but "autonomous" robots are robots which are *cognitively* autonomous, not *motivationally* autonomous. Non-autonomous robots collect information from the environment with their sensors and they send this information to a human being, and it is the human being who decides what the robot must do in response to this information. A robot is called autonomous if it has an internal system – a "brain" – which autonomously decides how to respond to the sensory input which arrives from the environment. But the autonomy of current autonomous robots is cognitive, not motivational. The robots autonomously decide what to do in response to the sensory input they receive from the external environment but they do so always with respect to some task which has been assigned to them by us. Motivationally autonomous robots are robots that decide what to do independently from us and independently from any task that we have assigned to them.

The reason why no one constructs motivationally autonomous robots is the applied orientation of today's robotics. In the words of Bill Gates, robots are constructed "to perform tasks in the physical world on our behalf". If robots are constructed to perform tasks for us, they do not need a motivational level of functioning because the cognitive level is sufficient. The ability to execute the tasks is the only ability the robots must possess. This is implicit in the very meaning of the word "robot" which was invented by the Czech writer Karel Čapek in the '20s of the last century. Čapek's "robots" were "serf workers", artificial workers which do things for us. If robots are our

serfs, we decide which motivation should control their behaviour, not them. Robots as practical applications should *not* have a motivational level of behaviour because robots that have a motivational level of behaviour may do something which we don't want them to do. And they cannot have emotions – which, as we will see later in this Chapter, are a sub-mechanism of the motivational level of functioning – because having emotions may imply to be anxious, sad, angry, excited, and to have all sorts of psychological pathologies. What would be the practical use of such robots? These are robots that we don't want and we may even fear. Robots must be "automata". Human beings are not automata.

Today's robots do not have a motivational level of functioning even when they are not constructed because they can be practically useful but as scientific tools to better understand animals and human beings. Robots are constructed on the basis of the principle "one robot/one phenomenon". One constructs a different robot for each task one is interested in and the robot has only the "motivation" to execute that particular task. Real animals are different. They have many different "tasks" in the sense that they have many different motivations and a very important aspect of their behaviour is how they decide which "task" to do at any given time. If we construct robots to understand real animals, we must construct robots which possess many different motivations and a mechanism for deciding which motivation to try to satisfy with their behaviour at any given time.

The motivation level of behaviour might be called "strategic" and the cognitive level "tactical". At the strategic level an animal decides which motivation should guide its behaviour at any given time and at the tactical level the animal does what it must do to satisfy the motivation. But in psychology and robotics the two words "strategic" and tactical" are used in a different sense. If the behaviour needed to satisfy a given motivation is a hierarchical structure of actions and sub-actions, the higher levels of the hierarchy are called "strategic" and the lower levels "tactical". If I decide to satisfy my motivation to eat, what I need to do is a hierarchy of actions. I must go out to buy food and then I must cook the food. To go out to buy the food, I must open the door of my house and then I must go to the grocery store. To open the door of my house, I must press the handle of the door and then push the door. And so on until I reach the level of simple movements of the muscles of my body. This is a complex hierarchy of actions and sub-actions but, in terms of our distinction between the motivational level and the cognitive level of behaviour, all these actions and sub-actions belong to the cognitive level of behaviour. Above the cognitive level there is the motivational level, and it is at this level that I decide to satisfy my motivation to eat rather than some other motivation such as watching TV. The study of the hierarchical structure of behaviour is a classical topic of research in both psychology and artificial intelligence but our distinction between the strategic and the tactical level of behaviour is something different. The tactical level is the level at which the hierarchically complex behaviour which is necessary to satisfy a given motivation is generated. The strategic level is the level at which the motivation is decided.

The motivational level of behaviour should also not be confused with what is called "action selection". Very simple animals respond to one specific stimulus always with the same behaviour, so that they can be said to know only stimulus-response associations. More complex animals respond to the same stimulus with different behaviours according to the circumstances. A human example is reaching an object with one's hand. Generally, the arm is moved so that the hand describes the most direct trajectory from its initial position to the object. But if there is something between the object and the hand, we select another action and the hand reaches the object by following a more complex trajectory around the obstacle. Action selection in robots is actively studied today and this research will make it possible to construct more flexible and more intelligent robots. But the study of action selection is *not* the study of motivations and it does not involve the motivational level of behaviour. One selects an action because, in the given circumstances, it is the only or the best way to satisfy a given motivation. The motivational level is choosing the motivation.

In this Chapter we describe robots which have different motivations and which autonomously decide which motivation to satisfy with their behaviour. This is an important goal in itself but robots that have different motivations and autonomously choose which motivation to try to satisfy with their behaviour are particularly important because they are a pre-condition for constructing robots that have emotions – and emotions are a crucial component of the behaviour of both animals and human beings.

Robots that have motivations and emotions raise a general issue for robotics. Current robotics is an *external* robotics. It is concerned with the external morphology of the robot's body and with the interactions of the robot's body with the external environment. External robotics may be sufficient to capture the cognitive half of the mind but to capture its motivational/emotional half we also need an *internal* robotics: the construction of robots with a body that contains internal organs and systems and a brain that interacts not only with the external environment but also with these internal organs and system. (We discuss internal robotics in Section 7 of this chapter.)

3. Robots that have motivations

In this section we will describe various types of robots which have two different motivations which they must both satisfy to remain alive and transmit their genes to the next generation. All the robots have the motivation to eat but, in addition to the motivation to eat, they also have another motivation and, since they cannot satisfy both motivations at the same time, at any given time they must decide to satisfy the motivation to eat or this other motivation.

a. *Robots that must both eat and drink*

To remain alive and leave their genes to the next generation, these robots must both eat food and drink water. Their body has two internal stores, one for energy and one for water. The two bodily stores have the same size and in each input/output cycle of the robot's neural network the same quantity of energy and water is consumed to keep the robot alive. The robots' environment contains both food tokens and water tokens. The two types of tokens are equally numerous and they are randomly distributed in the environment (see Figure 1a). To refill its two bodily stores, a robot must both eat the food tokens and drink the water tokens. The robots generate one offspring at regular intervals and, therefore, their fitness is the length of their life. (But all robots die at the same maximum age.) The robots' neural network (Figure 1b) has visual neurons which tell the robots where are the nearest tokens and if they are food or water (the food tokens are black and the water tokens are white) and it also has hunger and thirst neurons whose level of activation reflects the quantity of energy and the quantity of water currently contained in the robot's body.

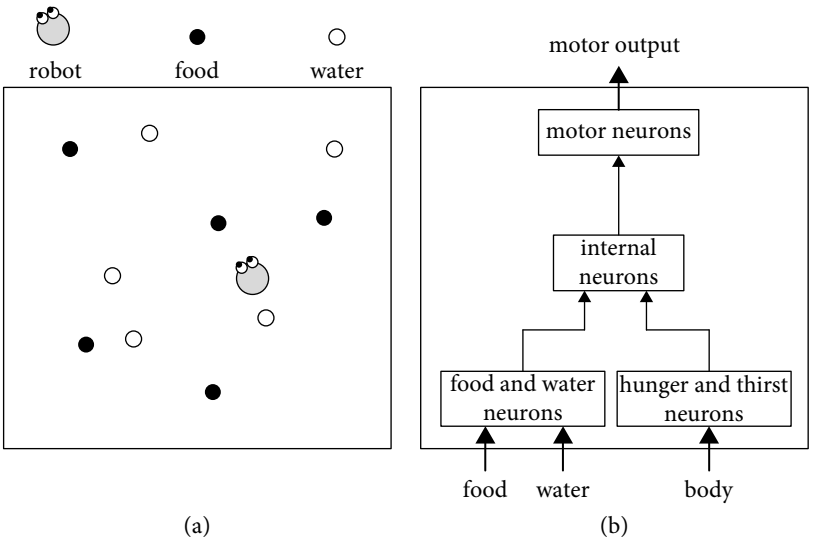


Figure 1. The robots live in an environment containing the same number of food tokens (black) and water tokens (white) (a). The robots' neural network has visual neurons that allow the robot to see the food and water tokens and hunger and thirst neurons (b).

The food tokens and the water tokens contain the same quantity of energy or water and a token is considered as eaten or drunk at the moment the robot's body touches the token. The token disappears and a new token of the same type appears in another place in the environment. The problem is that the robot cannot both eat and drink at

the same time and, therefore, at any given time it must decide whether to approach and eat a food token or approach and drink a water token. What must the robot do? When it sees a food token, should it approach and eat the food token or should it look for a water token? When it sees a water token, should it approach and drink the water token or should it look for a food token? When the robot sees both a food token and a water token at the same time, which token should it approach?

Since a robot dies if either the energy or the water in its body reaches the zero level, the robot cannot eat a lot and ignore drinking or drink a lot and ignore eating. To remain alive and have offspring the robot must both eat and drink in equivalent quantities and, what is even more important, at the appropriate time. For example, the robot cannot eat in the first part of its life and drink in the second part. The robot must eat when the level of energy in its body is low and the level of water is high and it must drink when the level of energy is high and the level of water low.

These robots are a very simple demonstration of the distinction between the motivational level and the cognitive level of behaviour. Deciding whether to look for food or for water is the motivational level. Executing the specific movements of the body which allow the robot to approach and reach a food token or a water token, is the cognitive level. The decision whether to look for food or for water is based on the current strength of the two motivations and, then, the robot must be able to reach the food and water tokens.

What determines the current strength of the two motivations? The quantity of energy and water contained in the robots' body is the most important factor. When a robot is hungry, it must approach and eat the food tokens and, when it is thirsty, it must approach and drink the water tokens. But the strength of the two motivations also depends on another factor: the distance of the food and water tokens from the robot. Given the same level of hunger and thirst, the motivation to eat wins the competition with the motivation to drink if a food token is nearer to the robot than a water token. This tells us that the robots do not only have the motivation to eat and the motivation to drink but they also have the motivation to save their time. To remain alive and reproduce, the robots must use as little time as possible to reach a food token or a water token so as to have more time to reach other food and water tokens.

This is what the robots do in their natural environment but, to have more precise and quantitative data, we do an experiment. We put a robot into an environment that contains only two tokens, a food token and a water token, both located in front of the robot (see Figure 2), and we vary the distance of the two tokens from the robot. If the two tokens are at the same distance (see Figure 2a), the robot tends to go to the food token half of the times and to the water token the other half. (We add a small amount of random "noise" to the activation level of the motor neurons that control the robot's two wheels so that the robot can respond in slightly different ways to the same sensory input.) But if one token is nearer than the other token (see Figure 2b and 2c), the robot goes to the nearer token, whatever it is, food or water. The robot's motivations

to eat and to drink have the same intrinsic strength but the distance of the two tokens from the robot makes a motivation stronger than the other and determines the robot's behaviour. This shows that, in addition to their motivation to eat and their motivation to drink, the robots also have the motivation to save their time.

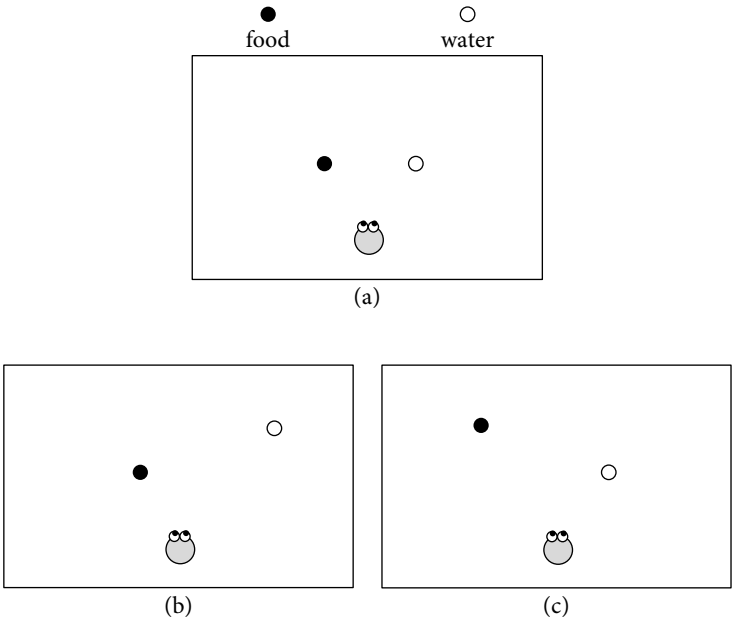


Figure 2. A robot in an experimental laboratory which contains one food token and one water token. The two tokens can be at the same distance from the robot (a) or one token can be nearer to the robot than the other token ((b) and (c)).

This is what happens if, when we do the experiment, the robot has the same level of hunger and thirst. But we can vary the level of the robot's hunger and thirst. If we keep the two tokens at the same distance from the robot, the robot goes to the food token if its hunger is greater than its thirst and to the water token if its thirst is greater than its hunger. If we vary the distance of the two tokens from the robot, we find that the two different motivations interact together to determine the robot's behaviour.

One motivation can be stronger than the other motivation for more intrinsic reasons. The robot's body may need more food than water and in this case the robot's motivation to eat is intrinsically stronger than the motivation to drink. We evolve a new population of robots which need more food than water to remain alive and we find that, in the experimental laboratory, if the food token and the water token are at the same distance from the robot, the robot goes to the food token rather than to the water token even if it is equally hungry and thirsty.

Another factor which influences the strength of the robots' motivations is the environment in which the robots live. The robots we have described live in an environment which contains the same number of food and water tokens. We evolve another population of robots in an environment in which the water tokens are less numerous than the food tokens (see Figure 3) and we find that in this environment the motivation to drink is intrinsically stronger than the motivation to eat.

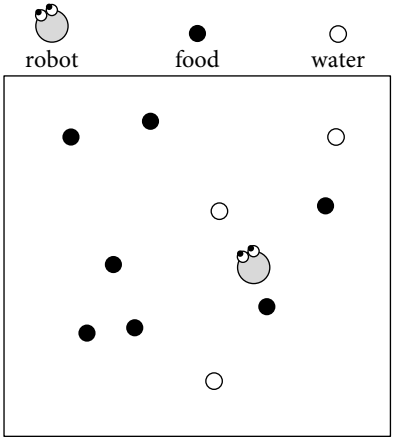


Figure 3. The environment contains fewer water tokens than food tokens.

If we look at the robots' behaviour in their natural environment, we find that when a robot sees a water token it approaches the water token, while this is not necessarily true when the robot sees a food token – unless the food token is very near. In many cases the robot ignores the food token and this shows that the robot is looking for water. Or, when the robot sees both a food token and a water token, the robot tends to go to the water token. The same is true in the experimental laboratory. If we make the robots equally hungry and thirsty, the robots prefer the water token to the food token. And even if we make them more hungry than thirsty, the robots still prefer the water token – unless they are very hungry or the food token is very near.

We construct another type of robots which live in a seasonal environment in which seasons with more food than water alternate with seasons with more water than food (see Figure 4) and what we find is that the strength of the robots' motivations changes from one season to the next although the robots do not explicitly know what the current season is. In the seasons in which food is more abundant than water, the robots prefer water to food because they tend to be more thirsty than hungry. In the seasons in which water is more abundant than food, the robots prefer food to water because they are more often hungry than thirsty. The adaptation of the robots to the seasonal environment is reflected in their behaviour in the experimental laboratory. In the experimental laboratory the robots' choice between the food token and the water token

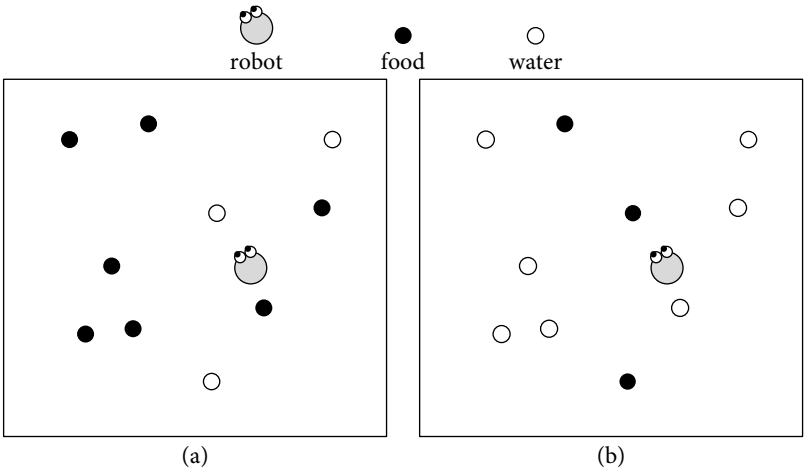


Figure 4. The robots' environment in a seasonal environment in which seasons with more food than water (a) alternate with seasons with more water than food (b).

mostly depends on their hunger or thirst rather than on the distance of the two tokens. The distance of the two tokens – the strength of the motivation to save time – is less important than it was in the non-seasonal environments.

Our final environment is an environment in which the food tokens and the water tokens are equally abundant but the food tokens are in one zone and the water tokens are in another zone of the environment and the two zones are separated by an empty space (see Figure 5).

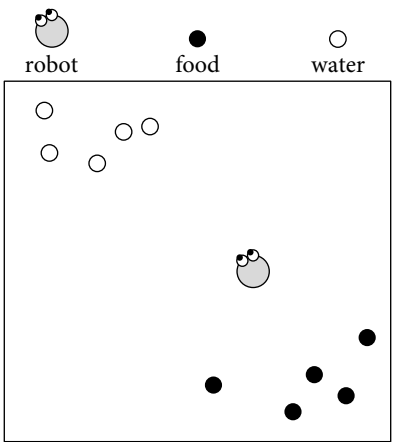


Figure 5. The robots live in an environment in which the food tokens are in one zone and the water tokens in a different zone and the two zones are separated by an empty space.

If we look at what the robots do in this environment, we see that they tend to stay for some time in one zone, either the food zone or the water zone, and then they move to the other zone. Like the robots that live in the seasonal environment, in the experimental laboratory the behaviour of these robots depends not so much on the distance of tokens from the robot but on the quantity of energy and water in their body. The robots tend to approach the token which contains what their body needs more at the time of the experiment – energy or water.

The adaptive pattern of these robots is more complex. Since food and water are in two separate zones of the environment, the robots must avoid to continuously travel from one zone to the other zone because this would mean a waste of time and a waste of the energy and water contained in their body. And, in fact, when they are in the food zone the robots do not leave the food zone until they cease to be hungry and their thirst has become rather great. And the same happens for the water zone. When they are in the water zone, they drink a lot before leaving the water zone and traversing the empty space to reach the food zone. In this manner they reduce the number of times they cross the empty space which separates the two zones and avoid consuming their time and their bodily energy and bodily water while traversing the empty space. The behaviour of these robots is more sophisticated and is delicately dependent on their level of hunger or thirst.

These robots have a brain which is informed of the quantity of energy and water contained in the body. The robot's neural network has hunger and thirst neurons and therefore the robot's behaviour depends not only on what they perceive in the external environment but also on what they sense about their body. What happens if the robots have a neural network *without* the hunger and thirst neurons? The quantity of energy and water in the robots' body continuously changes both because energy and water are consumed to keep the robots alive and because the robots eat and drink. But the robots don't know how much energy and water they have in their body. How do these robots compare with the robots that have the hunger and thirst neurons? Do the robots with the hunger and thirst neurons live longer than the robots without the hunger and thirst neurons?

The answer depends on the environment in which the robots live. The robots that live in the environment with an equal number of food and water tokens, randomly distributed, and those that live in the environment in which water is always less abundant than food, have the same fitness whether they have or do not have the hunger and thirst neurons. The hunger and thirst neurons make a difference only for the robots that live in the seasonal environment and for the robots that live in the environment in which the food tokens are in one zone and the water tokens in another zone. In these two environments, the robots with the hunger and thirst neurons have more fitness (live longer) than the robots without the hunger and thirst neurons (see Figure 6).

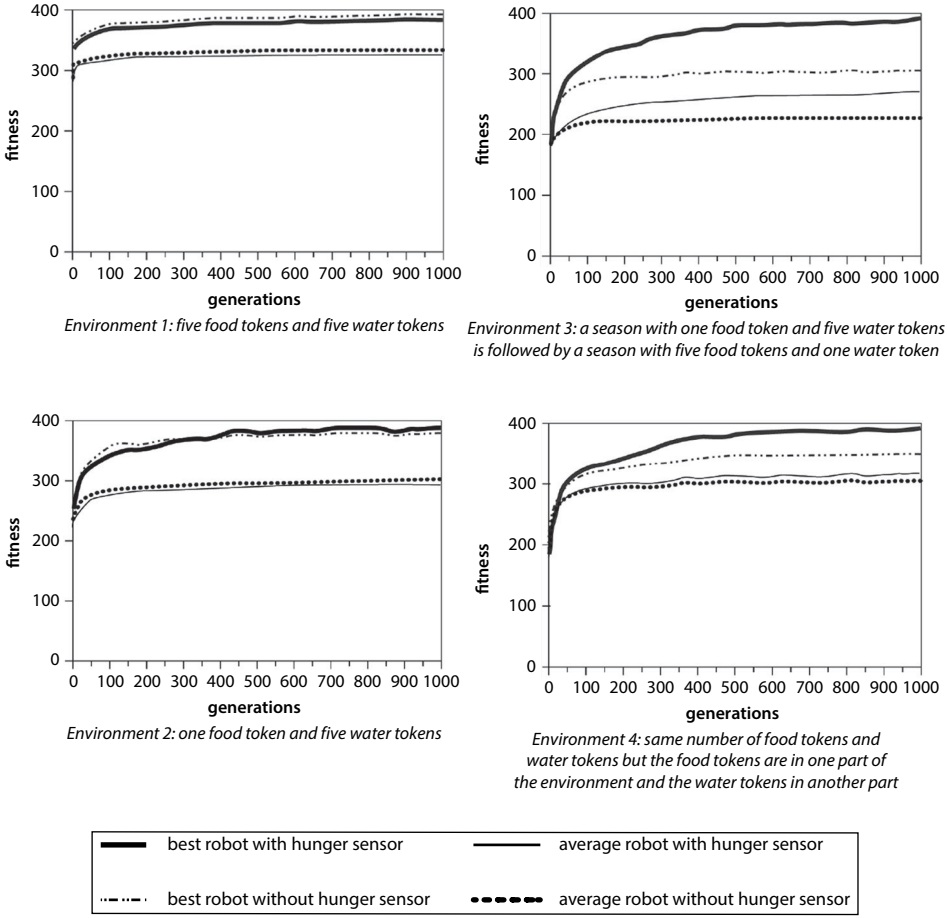


Figure 6. Fitness (length of life) of robots with and without hunger and thirst neurons that evolve in four different environments.

These results tell us something interesting about what determines the strength of an animal's motivations and how this strength depends on the environment in which the animal lives. In the environment in which food and water are equally abundant and in the environment in which food is always more abundant than water, the robots behave in the same way whether they have or do not have the hunger and thirst neurons. If food and water are equally abundant, the robots adopt a very simple strategy: they always approach the nearest token, whether food or water. This is understandable because in this environment the behaviour of approaching the nearest token, whether food or water, is sufficient to guarantee one's survival and the robots don't need to know how much

energy and how much water are currently contained in their body. The robots that live in the environment in which water is always less abundant than food, adopt another very simple strategy: they always prefer water to food – unless food is very near. This strategy also is reasonable. If in the environment water is always less abundant than food, the robots prefer to approach water rather than food because food is sufficiently abundant that they will eat even if they are more preoccupied with drinking rather than eating. In this environment too, it is not necessary to feel hungry or thirsty.

The situation is different in the other two environments. If the abundance of food and water depends on the season, and the robots ignore what the current season is and therefore they ignore if food is more abundant than water or vice versa, the sight of food and water becomes less important and what tells the robots what they should do is their body. When they feel hungry, they prefer food to water and, when they feel thirsty, they prefer water to food. The same is true for the environment in which food and water are in two different zones of the environment. If the robots are in the food zone and they continue to eat the food tokens even when there is little water in their body, they can die for lack of water in their body. The only solution is that their body tells their brain that there is little water in the body – they feel thirsty – so that they can leave the food zone and go to the water zone. And vice versa, if they are in the water zone.

If these robots have or do not have the hunger and thirst neurons is decided by us but the possession of hunger and thirst neurons is an adaptation that autonomously evolves in animals. We have not constructed robots that evolve their hunger and thirst neurons but they could be constructed by adding another gene to the robots' genotype, a "hunger and thirst" gene which encodes the probability that a robot has the hunger and thirst neurons in its neural network. The robots of the initial generations have a random value for their "hunger and thirst" gene and, therefore, the probability that a robot has the hunger and thirst neurons is only 50%. Then, the value of the gene changes as a result of the selective reproduction of the best robots and the constant addition of random variations to the value of their "hunger and thirst" gene and we predict that, for the robots that live in the environment in which food and water are equally abundant or in the environment in which food is always more abundant than water, the average value of the gene will remain more or less 50%. But for the robots that live in the seasonal environment or in the environment in which food and water are in two different zones, the value of the gene will increase and reach a value near to 100% because in these two environments having hunger and thirst neurons is necessary to remain alive and have offspring.

b. *Robots that must both eat and save their bodily energy by moving slower*

The robots we have described need both food and water to survive. We now construct another type of robots which need only food to remain alive. A robot has only an energy store in its body and, if the bodily energy reaches the zero level, the robot

dies. Life is simpler for these robots. They have a single motivation, the motivation to eat, and therefore, from this point of view, they don't have to choose among different motivations. But these robots have another problem. While the bodily energy of the preceding robots is consumed by the same quantity in each cycle, the bodily energy of the new robots is consumed by a quantity which varies with the speed with which the robots move in the environment. Moving is energetically costly. When they move very fast, these robots consume more energy than when they move more slowly.

We compare two populations of robots. The robots of one population are those we already know: their bodily energy is consumed by a fixed quantity in each cycle. The robots of the other population have a more complex adaptive pattern: if they move very fast, they consume more energy. Do the two types of robots behave differently?

The robots that consume more energy when they move very fast tend to move more slowly than the robots that do not pay for moving faster but have only the fixed energy costs associated with the mere fact of living. However, if we measure the fitness – length of life – of the two types of robots, we find that the robots that move more slowly because they pay for their speed of movement live longer than the robots that do not incur such a cost. How is that possible? The answer is that the robots with variable energy costs respond differently to the changes in the level of energy in their body. All robots have the hunger neurons which inform their brain of the current level of energy in their body. But while this information is not particularly useful for the robots with fixed energy costs because these robots look for food whatever their level of energy, the robots with variable energy costs have the possibility to control one of the factors that determine the level of energy in their body because they can regulate the speed with which they move. They exploit this possibility and they move faster when they are hungry and more slowly when they are no so hungry. These robots exhibit a more sophisticated behaviour which allows them to have a longer life compared to the robots with fixed energy costs.

How can we interpret the behaviour of these more sophisticated robots? The robots have the motivation to eat in order to remain alive and, if this were their only motivation, they could not be said to make motivational decisions. But these robots also have another motivation, the motivation to save their energy and, at any given time, they must decide whether to satisfy their motivation to eat, which requires moving as fast as possible, or to satisfy their motivation not to move too fast to save their energy. As for all motivational decisions, this decision depends on the current strength of the two motivations. If the level of energy in the robot's body is low and the robot is very hungry, the strength of the motivation to eat is greater and the robots tend to move fast. If the robot is not so hungry, the motivation not to move too fast to save energy wins the competition with the motivation to eat, and the robots move more slowly. This is another demonstration of how the different motivations of an individual interact together and how these interactions determine their current strength of the individual's motivations and, therefore, its motivational decisions.

c. *Robots that must both eat and flee away from a predator*

Hunger and thirst are both motivations based on information which originates in the body and is transmitted to the brain. The robot's brain has internal sensors that register the state of the body, and this sensory input from the body is used by the brain to decide which particular motivation should control the robot's behaviour at any given time. In other cases the critical information that allows a robot to make the appropriate motivational decisions does not arrive from inside the body but it arrives from the external environment. This is what we show with our next robots.

The new robots live in an environment which also contains food tokens that the robots must eat to remain alive. But these robots have an additional problem. Once in a while, at randomly chosen times, a predator appears and tries to reach and kill the robots (see Figure 7). (The predator is hardwired by us and, therefore, is not a robot.)

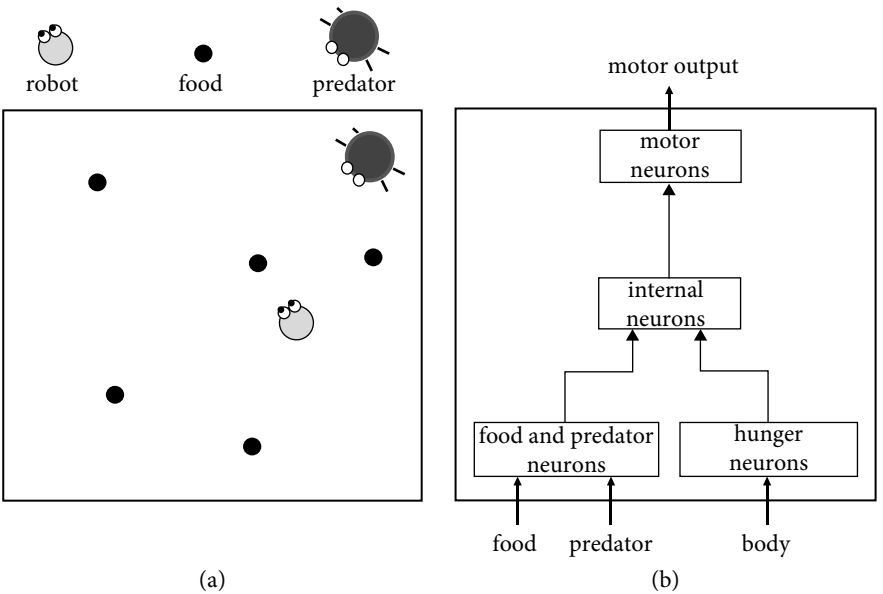


Figure 7. The robots live in an environment that contains food tokens but, once in a while, a predator appears and, if it reaches the robots, it kills them (a). The robots' neural network (b).

When the predator is not present, the only sensory input which arrives from the external environment to the robot's brain concerns the location of the food tokens and the robot must respond to this sensory input by approaching and eating the food tokens. However, when the predator appears, a separate set of sensory neurons of the robot's neural network encode the direction from which the predator is approaching and the distance between the predator and the robot. Since, predator kills the robot if it succeeds in reaching the robot, these robots also have two motivations to satisfy: the motivation to eat and the motivation to avoid being killed by the predator.

The strength of the motivation to eat depends on the bodily input which tells the robot's brain how much energy is currently contained in the body (level of hunger). The strength of the motivation to flee away from the predator depends on the sensory input arriving from the external environment – the sight of the predator. When the predator appears, the robots must decide whether to continue to pursue the motivation to eat and ignore the predator, or ignore food and switch to the motivation to flee away from the predator. The motivation to flee away from the predator is intrinsically very strong because, if the predator reaches the robot, it kills the robot. But we can manipulate this variable. We evolve the robots in different environments in which the predator can inflict different degrees of damage to the robots, without necessarily killing them. Eating one food token increases by some fixed quantity the survival chances of the robot. Being reached by a predator reduces by another quantity these chances but how much the survival chances of the robot are reduced depends on the dangerousness of the predator.

What are the consequences of living in an environment with more dangerous predators or in an environment with less dangerous predators? Some of the results are predictable. The robots that live in an environment with very dangerous predators evolve a better capacity to avoid being reached by a predator than the robots that live with less dangerous predators. For the robots that live with very dangerous predators, the motivation to avoid the predator is so intrinsically strong that the evolutionary pressure to develop an effective fleeing behaviour is very powerful. These robots eat less food than the robots living with less dangerous predators because, when the predator appears, they completely ignore food and flee away from the predator. But what is interesting, and perhaps unexpected, is another result. The robots that live with very dangerous predators eat less food not only when the predator makes its appearance but also when the predator is *not* present. Their inferior eating capacity is demonstrated by putting a robot into an experimental laboratory which contains only one food token and measuring the time taken by the robot to reach the food token. The robots living with more dangerous predators take more time to reach the food token than the robots living with less dangerous predators.

Why? One hypothesis is that, for the robots that live in an environment with very dangerous predators, more of their neural resources – neurons and connections between neurons of the robots' neural network – are used for the task of fleeing away from the predator and less neural resources are available for the task of eating and, therefore, the robots are less good at eating. The sensory input from food and the sensory input from the predator are both processed by the same set of internal neurons and, if more of the processing capacity of these internal neurons is dedicated to predators that are very dangerous, less processing capacity is left for responding appropriately to food. This hypothesis could be tested by examining the robots' neural network and seeing if the pattern of activation of the internal neurons co-varies more with the pattern of activation of the sensory neurons that encode the location of the predator than with thepattern of activation of the sensory neurons that encode the location of the food tokens – where "varying with" means "representing". But we have not made this test.

Another hypothesis is that the robots that live in an environment with more dangerous predators are more emotionally preoccupied with predators and this makes their behaviour of looking for food less efficient. This hypothesis has to do with emotions and we discuss emotions and their role in motivational decisions in the next section.

Like the robots that live in the environment in which water is less abundant than food and therefore the motivation to drink is intrinsically stronger that the motivation to eat, these robots are another demonstration that the intrinsic importance of the different motivations is a critical factor in deciding which motivation controls the robot's behaviour at any given time. For these robots, finding food is important but fleeing away from the predator is more important. But these robots tell us something more about motivational decisions. Motivational decisions must not only be right but they must also be made with the necessary speed. When the predator is absent, a robot must dedicate itself to the food tokens because, otherwise, it dies. But when the predator appears, the robot must not only switch from looking for food to fleeing away from the predator but it must make the switch as quickly as possible. Even if the motivational decision is the correct one, a robot's survival chances can be seriously reduced if the robot takes too much time to decide. And, in fact, at the end of evolution the robots tend both to make the appropriate motivational decisions and to decide with the necessary speed. When the predator appears, they immediately switch from the motivation to eat to the motivation to flee away from the predator.

d. *Robots that must both eat and find a partner for reproduction*

The most important motivation of both animals and human beings is the motivation to reproduce in order to leave their genes to the next generation. Our next robots have this motivation and they tell us how this motivation enters into their motivational decisions.

All the robots described so far reproduce nonsexually: a robot periodically generates one offspring with no need for a mating partner. The new robots reproduce sexually – they need a mating partner to generate an offspring. To remain alive they must eat but remaining alive is not enough to pass on their genes to the next generation. The robots must mate with another robot. The fitness of a robot is not the length of its life, as in our preceding robots, but the number of mating events of the robot, where a mating event occurs when the robot touches a mating partner. But these robots continue to have the problem of eating and, therefore, at any give time they must decide whether to look for food and to look for a mating partner.

The robots live in an environment which contains a certain number of food patches and a potential mate (see Figure 8).

A food patch is different from a food token because a food token contains only one unit of energy and it disappears as soon as the robot touches (eats) it, while a food patch contains many units of energy and the robot eats one of these energy units for each cycle spent inside the food patch until the food patch is exhausted and it disappears.

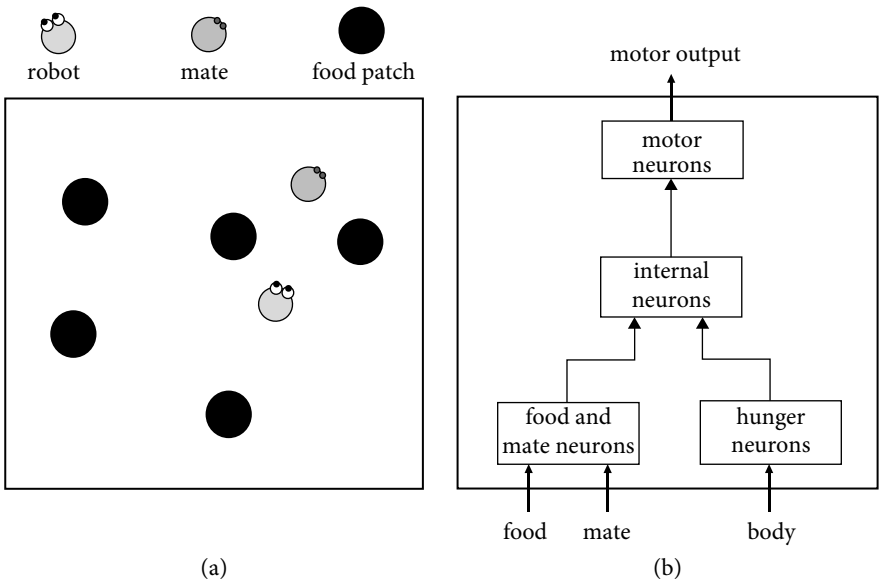


Figure 8. The environment contains a certain number of food patches and one potential mate (a). The robots' neural network (b).

(We have chosen food patches rather than food tokens because a food patch is more attractive for the robots.) The potential mate is not a real robot but a non-moving object which the robots distinguish from the food patches based on its colour. When the robot touches a mating partner, a mating event automatically occurs and an offspring is generated. After the mating event the mating partner disappears, and another potential mating partner appears in a new, randomly chosen location in the environment. To leave their genes to the next generation the robots must both reach the food patches and eat the food which is found there and reach and automatically mate with a mating partner. If a robot always looks for a mating partner and ignores food, the robot will die and therefore it will not be able to reproduce. If the robot always looks for food but ignores the potential mate, the robot will have a long life but it will not pass on its genes to the next generation. The motivation to eat has a variable strength which depends on the quantity of energy present in the robot's body and, like the preceding robots, these robots have hunger neurons that inform them on the current level of energy in their body. On the contrary, the motivation to mate has a constantly high intrinsic strength but, as in the case of the predator, it is only activated when the robot sees the potential mating partner.

At the end of evolution, the robots take all these factors into considerations in deciding which motivation to pursue. They divide appropriately their time between looking for food, especially when they are very hungry, and looking for a mating partner. But the motivation to mate is stronger than the motivation to eat and this is

clearly shown by how the robots behave in the experimental laboratory. When they are exposed to both a food token and a potential mate, they tend to go to the potential mate – unless they are very hungry.

e. *Robots that must both eat and feed their offspring*

For many animals mating and having offspring are not sufficient to transmit their genes to the next generation. For a period of time after birth the offspring need to be taken care of by their parents until they reach an age in which they can take care of themselves. Otherwise, the offspring die and the genes of their parents disappear from the genetic pool of the population. The new robots do not reproduce sexually. A robot generates one offspring at regular intervals but, unlike the preceding robots, the robot must take care of their dependent offspring to keep them alive. The robots must eat to remain alive but their fitness is not the length of their life and, therefore, the total number of offspring that they are able to generate, but the time spent taking care of their offspring.

The environment contains a large food patch and a nest in which the robot's offspring live. The offspring are virtual robots and the only thing that a parent robot actually perceives is the nest. The problem is that the food patch and the nest are in two opposite corners of the environment and, therefore, the robots must divide their time between the food patch and the nest (see Figure 9).

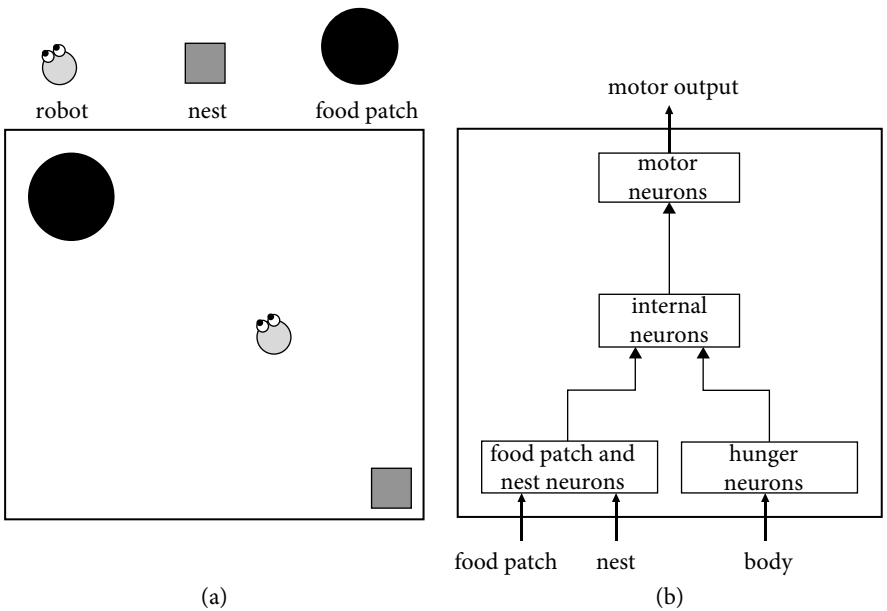


Figure 9. The robots live in an environment with a food patch and a nest where the robots' dependent offspring live (a). The robots' neural network (b).

For each cycle spent by a robot in the food patch the energy contained in the robot's body is increased by some fixed quantity. For each cycle spent in the nest, the survival chances of the robot's virtual offspring increase by a fixed quantity. As we have said, the robots reproduce nonsexually and they generate one offspring at regular intervals. However, their fitness is neither the length of their life nor the number of offspring that they generate but the time spent in the nest. A robot has to both eat (time spent in the food patch) and take care of its offspring (time spent in the nest) because, if it does not eat, the robot dies and cannot generate other offspring or take care of its existing offspring and, if it does not take care of its offspring, the offspring die and the robot's genes are not passed on to the next generation. The robots must find the appropriate balance between these two different motivations. The robots don't have visual neurons and, therefore, they do not see the food patch or the nest from a distance. But, in addition to the hunger neurons, their neural network has smell neurons which, when they are activated, tell the robots if they are inside the food patch or inside the nest. This means that the robots must be able to explore the environment without any sort of external sensory input but they must behave appropriately when they are inside the food patch or inside the nest.

The robots very rapidly evolve what for them is the most basic capacity: the capacity to explore the environment when they do not receive any external sensory input. Exploring the environment depends on the speed of the robot's two wheels. If the two wheels move with very different speeds, the robot turns around itself and, therefore, it does not explore the environment and it may never reach the food patch or the nest. On the contrary, if the two wheels have more or less the same speed, the robot follows a more or less straight trajectory and this increases the probability that the robot will end up either in the food patch or the nest. (The environment of the robots is closed by a wall and the robots bounce back when they touch the wall.) The first stages of evolution are mostly dedicated to evolving the capacity to rotate their two wheels appropriately so that the robots can explore the environment. Since there is no external input, this capacity depends on the intrinsic activation level of the internal and motor neurons and the robots that reproduce are those that have better intrinsic activation levels encoded in their genes and, therefore, explore more of the environment. When they happen to enter the food patch or the nest and they smell the food patch or the nest, they become progressively better at deciding appropriately between staying in the food patch in order to eat and leaving the food patch and reaching the nest in order to take care of their offspring and insure their survival.

These robots also have two motivations that they must satisfy if they want to leave their genes to the next generation. The strength of the motivation to eat varies with the quantity of energy contained in their body (hunger) while the motivation to take care of their offspring has always the same strength. If we examine the robots' behaviour, we find that, when they are inside the nest and they feel very hungry, they leave the nest searching for the food patch. If they are inside the food patch and they feel well satiated,

they leave the food patch and try to reach the nest to take care of their offspring. At the end of evolution the robots are able to both decide appropriately when to pursue one or the other of the two motivations and to behave appropriately to satisfy the motivation decided at the motivational level.

f. *Robots that must both eat and stop moving when their body is physically damaged*

Our last robots have a body which can incur physical damage and, when their body is physically damaged, the robots must stop moving in order to heal from the physical damage because, if they don't stop moving, the physical damage may compromise their survival (see Figure 10).

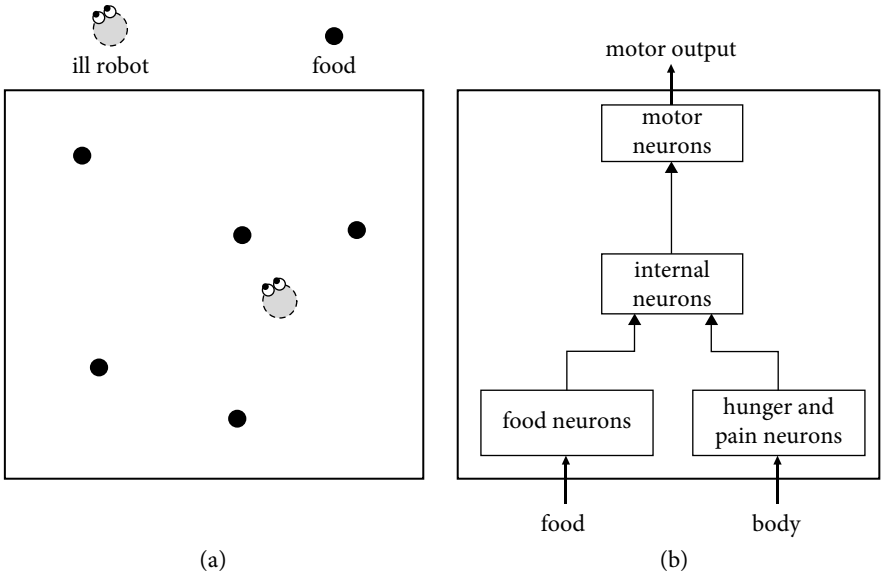


Figure 10. A robot which is ill (a) and the robot's neural network (b).

The robots live in an environment containing the food tokens that they must eat to stay alive but it sometimes happens that some external or internal event causes a physical damage to their body and they become ill. (The external or internal event and the physical damage are decided by us and are not explicitly reproduced.) If, when it is ill, a robot continues to move, the robot's risk of dying is increased even if the robot continues to eat. What the robot must do is stop moving (and, therefore, stop eating) because this will accelerate the spontaneous recovery from the illness. Therefore, these robots also have two different motivations: the motivation to eat and the motivation to heal by not moving when their body is ill. They cannot do both things at the same time and, therefore, when their body is ill, they must decide whether to continue to move and look for food or to stop moving in order to heal from their illness.

In addition to the visual neurons which encode the location of the food tokens and the hunger neurons, the robot's neural network has a pain neuron whose activation level can vary from 0 to 1. This activation level reflects the seriousness of the robot's physical damage, with an activation level of 0 indicating no physical damage and an activation level of 1 indicating maximum physical damage. (The seriousness of the physical damage and its duration are both decided randomly.) Since the robots can move at different speeds, given a certain degree of physical damage their loss of fitness is greater if they move fast than if they slow down. Like the other robots, these robots evolve in a succession of generations and the fitness criterion for deciding which robots reproduce and which do not reproduce is the number of food tokens eaten minus the number of cycles in which the robot has continued to move when its body was ill, with a reduction in fitness proportional to the seriousness of the illness and to the speed of movement of the robot.

At the end of evolution the robots have evolved a rather complex behaviour. If their pain neuron is not activated, they respond to the sensory input from food by approaching and eating the food. But, when their pain neuron is activated, they respond to the same sensory input from food by reducing their speed of movement, or even stopping moving. And the robots adapt their speed of movement to the seriousness of the physical damage signalled by the pain neuron. If the pain signal is not too strong, the robots continue to move and look for food, perhaps at a reduced speed. If the pain signal is very strong, they switch from pursuing the motivation to eat to pursuing the motivation of healing from their illness, and they stop moving.

All the robots described in this section clearly demonstrate that behaviour depends on motivations, and that these motivations, their strength, and the factors that determine their strength, depend on the environment in which the animal lives and on the adaptive pattern of the animal – its body, its brain, its sensory and motor organs, its manner of reproduction.

Motivations and motivational choices are important by themselves but they are especially important because they are the cause of emotions. The robots of the next section show how motivations and emotions are related together.

4. Robots that have emotions

We have distinguished between two levels in the behaviour of animals, the motivational level and the cognitive level. At the motivation level the animal decides which motivation to satisfy with its behaviour and at the cognitive level the animal executes the behaviours that allow the animal to satisfy the motivation decided at the motivational level. That our robots function at two levels is shown by the fact that some robots are good at the motivational level but not very good at the cognitive level while the opposite is true for other robots. One robot may be good at deciding to stop eating and flee away as soon as the predator arrives but the robot is not very good at running away

from the predator – and it will be probably killed by the predator. Another robot may be good at running away from the predator but it hesitates to stop looking for food when the predator appears – and it will even more probably killed by the predator. In Chapter 11 dedicated to robots with psychopathologies we will see robots that have serious problems at the motivational level and, when the predator appears, they are unable to decide whether to continue to look for food or to flee away from the predator. They simply do nothing and remain where they are – and are killed by the predator. Being able to do what is necessary to satisfy a given motivation is important but being able to decide correctly which motivation to try to satisfy is clearly more important. If the motivational level does not function well, its "mistakes" may be critical for the animal's survival and well-being.

This is why animals and human beings have emotions. Emotions (or emotional states; we use the two expressions as synonymous) are states of an animal's body/brain that increase the correctness and effectiveness of the animal's motivational decisions by influencing the current strength of the different motivations. If the motivations among which the animal must decide are not very strong, if the factors that have to be taken into consideration to make the appropriate decision are simple and restricted in number, or if the decision process is such that errors and slowness do not compromise the animal's survival and reproductive chances, motivational decisions do not need emotions. Emotions influence an animal's motivational decisions but motivational decisions need not be necessarily accompanied by emotional states. An animal can shift from looking for food to looking for water with no particular emotion accompanying the decision to shift. Emotions play a role when motivational decisions are important for the animal's survival and well-being and the animal must decide quickly, and they are especially numerous and articulated in human beings who must decide among a very large number of different motivations and must take many different and complex factors into consideration in order to decide correctly. Animals and human beings can make errors in their motivational decisions or they can be too slow in deciding, and this may compromise their survival and their well-being. Emotional states are a mechanism for making the motivation level of behaviour function more effectively, with fewer errors and with the necessary speed. Using a metaphor, emotional states allow one motivation to "speak louder" than other motivations in order to win the competition with other motivations. The sight of a present danger may induce an emotional state in an animal that increases the probability that the animal will stop pursuing other motivations and will try to avoid the danger, and will do this very quickly. The sight of a potential mate may cause an emotional state that increases the probability that the animal will pursue the motivation to mate rather than other motivations. In human beings even the *thought* of a danger or of a mate can cause emotional states which increase the strength of the motivation to avoid the danger or the motivation to be with the mate, with the consequence that the individual will be less likely to pursue other motivations. (For robots that have thoughts and a mental life, see Chapter 5.)

Of course, like everything in animals and human beings, emotions may not function properly and this may be an important cause of unhappiness and pathologies. But our hypothesis is that emotions have evolved in animals and human beings to let them function better at the motivational level.

There is much work dedicated to robots with emotions but these robots cannot be said to *have* emotions. They only *express* emotions. The robots move their face or their body in ways that we interpret as expressing emotions but, since they do not have emotions, they cannot express emotions. The robots express "unfelt" emotions, and the notion of an "unfelt emotion" is self-contradictory. Some robots can recognize our emotions on the basis of the movements and postures of our face but they process these movements and postures as a purely perceptual task, without sharing our emotions and, therefore, without really *understanding* our emotions.

"Emotional" robots cannot be said to have emotions because they do not have motivations. Emotions exist to make the motivational decision mechanism function more properly. Current robots do not have motivations and they do not have to decide which motivation to pursue with their behaviour and, therefore, they cannot have emotions.

In the preceding section we have described robots that have both a motivational level and a cognitive level of functioning. The best robots tend to be good at both levels but less good robots may be good at one level but not very good at the other, and the worst robots at neither. However, these robots cannot be said to have emotions because nothing happens in the robots' brain that can be described as an emotion or an emotional state. So we must ask: How can we endow our robots with emotions? We add a set of special neurons – emotional neurons – to the neural network that controls the robot's behaviour and we show that these emotional neurons lead to better motivational decisions and, therefore, to higher fitness. The motivational decision mechanism exists in all animals. Even a very simple animal such as the small worm *Caenorhabditis elegans*, with only about 300 neurons and a total of less than 1000 cells in its body, in certain circumstances must decide whether to look for food or to bend its body when a stimulus touches its body. But it is not clear that one can attribute emotions to this simple creature. Our robots are even simpler than *Caenorhabditis elegans* and they can survive and reproduce without emotional neurons in their brain. But if we add a set of emotional neurons to the neural network that controls their behaviour, they reach higher levels of fitness.

The brain of the robots described so far is very simple. It is a neural network with sensory neurons encoding information from the external environment (food, water, predator, mating partner, nest with offspring) and from inside the body (hunger, thirst, pain), this information is transferred to the internal neurons, and the internal neurons activate the motor neurons that control the robot's two wheels. We now add a set of emotional neurons to this basic neural network. The emotional neurons are part of an emotional circuit because they are activated by the sensory neurons and they in turn send their connections to the rest of the robot's neural network and, in this way, they

influence how the robot responds to the sensory inputs. The emotional neurons of the robots that must eat and drink are activated by the hunger and thirst neurons. Those of the robots that must both eat and flee away from the predator are activated by the sensors which encode the presence of the predator. For the robots that must both eat and find a mating partner, the emotional neurons are activated by the sensors which encode the mating partner, for the robots that must both eat and go to the nest to feed their offspring, they are activated by the nest sensors, and for the robots that must both eat and stop moving when their body is ill, by the pain sensors.

The emotional neurons have three characteristics which make them different from the regular internal neurons. First, they do not have an intrinsic level of activation, and this means that, when no activation arrives from the sensory neurons, they are not active and they have no influence on the robot's behaviour. Second, they have an activation threshold. If not enough activation arrives to the emotional neurons from the sensory neurons, they remain inactive. Third, they are active not only in the cycle in which they are activated by the sensory neurons but their activation persists for some time after they have been activated. Their activation increases for a certain number of cycles, it reaches a maximum level, and then it decreases. Both their activation threshold and these three variables of their prolonged activation are encoded in the robot's inherited genes and, therefore, their value evolves so as to make the emotional circuit a useful addition to the robots' neural network.

In the preceding section we have described five different types of robots living in five different environments. In this section we compare the fitness of these robots – which do not have an emotional circuit in their network – with the fitness of robots that do have the emotional circuit. The result of this comparison is that the robots with the emotional circuit have more fitness than the robots without the emotional circuit. The robots must make motivational decisions and, as we have seen, they can make the appropriate motivational decisions even without the emotional circuit. However, possessing an emotional circuit leads to more effective behaviour and higher levels of fitness, and this is true for all six types of robots: those that must choose between eating and drinking, eating and fleeing away from the predator, eating and looking for a mating partner, eating and taking care of their offspring, eating and stop moving to heal from physical damage. The results are shown in Figure 11.

We say that these robots have emotions because they provide us with an operational definition of the word "emotion". The robots' emotions *are* the patterns of activation of their emotional neurons. The emotional neurons have a functional (beneficial) role because they increase the robots' fitness. One might object that adding an emotional circuit to the robots' neural network is only increasing the number of internal neurons so as to make the robots' brain larger and more powerful, and this is sufficient to explain why the robots with the emotional circuit have more fitness that the robots without the emotional circuit. But this is not so. We evolve new populations of robots that do not have the emotional neurons in their neural network but have a larger number of regular internal neurons, and we compare the performance

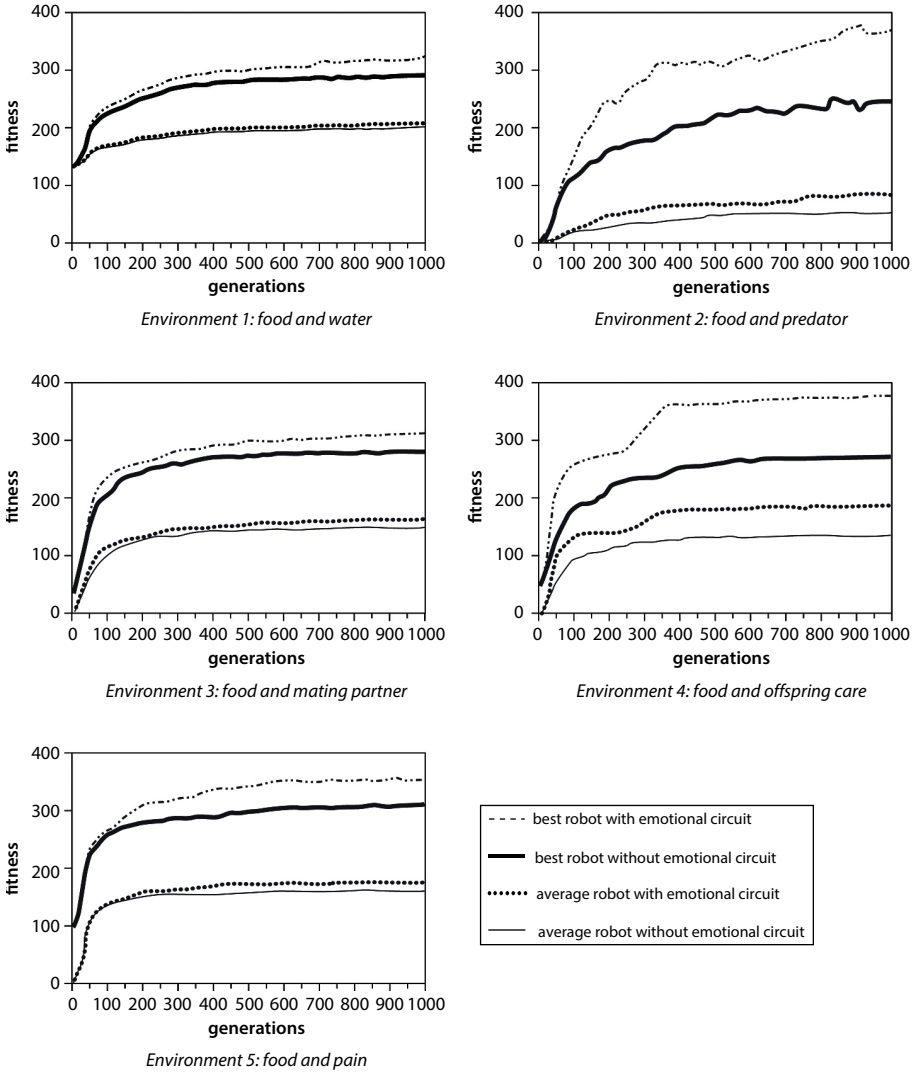


Figure 11. Evolution of the fitness of robots which have and robots that do not have an emotional circuit in their neural network. The robots are those described in the preceding section.

of these robots with the performance of the robots with the emotional neurons. What we find is that the robots with a larger number of regular internal neurons have more or less the same fitness as the robots with fewer internal neurons. This shows that the emotional neurons have a beneficial effect on the robot's behaviour not because they simply increase the computing power of the robots' brain but because of their special properties.

To better understand how the emotional neurons influence the robots' behaviour, we compare two different neural architectures. In one architecture the emotional neurons send their activation to the regular internal neurons while in the other architecture they send their activation directly to the motor neurons (see Figure 12).

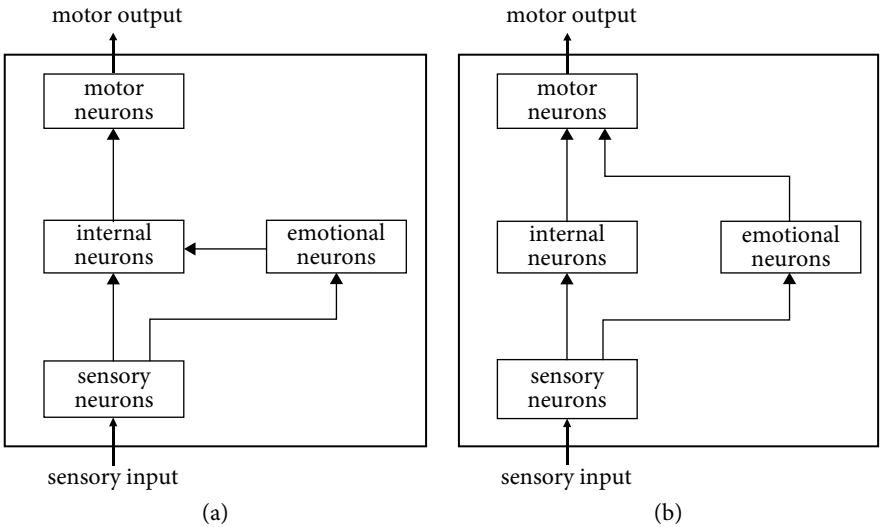


Figure 12. Two different neural architectures for the robots which have emotions. In one architecture the emotional neurons send their connections to the regular internal neurons (a). In the other architecture, they send their connections directly to the motor neurons (b).

Which architecture produces the best results? The answer depends on the different robots. The architecture with the emotional circuit directly connected to the motor neurons is better for the food/predator, food/mating partner, and food/nest robots, while the architecture with the emotional circuit linked to the regular internal neurons is better for the food/water and food/pain robots. This is interesting. The emotional circuit directly connected to the motor neurons (see Figure 12b) gives better results when the emotional neurons are activated by an external sensory input: perceiving the predator, perceiving a mating partner, perceiving the nest. On the contrary, the emotional circuit connected to the internal neurons and, therefore, only indirectly to the motor neurons (see Figure 12a), gives better results when the emotional circuit is activated by an internal input from the body: hunger, thirst, pain. Why? We do not have an answer to this question but we note that there is a difference between external and internal sensory input to the brain. Sensory input from the external environment tends to be richer in (cognitive) information than sensory input from the body. Perceiving the predator, the mate, or the nest provides the robot's brain not only with motivationally relevant information but also with cognitive information (where are the predator,

the mate, or the nest). In contrast, the sensory input which arrives from within the body is poorer in cognitive information and it contains prevalently motivationally relevant information (how strong is the hunger, thirst, or pain). One way to explain the results is to make the following hypothesis. In the case of the more cognitively rich external input, the emotional circuit can directly go to the motor neurons without passing through the cognitive part of the brain – the regular internal neurons. On the contrary, when it is the cognitively poorer body input which activates the emotional circuit (hunger, thirst, or pain), the emotional circuit must first pass through the regular internal neurons so that the motor response can be more cognitively informed and the robot knows in which direction it must move.

To better understand the effects of having an emotional circuit in the brain, we look more closely at the behaviour of the robots that have the emotional circuit. We do this only for the robots that must choose between eating and fleeing away from the predator because these robots offer a particularly clear demonstration of the influence of emotions on behaviour. As we have said, the robots with the emotional circuit have more fitness (live longer) than the robots that do not have the emotional circuit, and the fitness of these robots depends on two factors: the robot's ability to find food and its ability to escape from the predator when the predator appears. These are both cognitive capacities. But what is even more critical for a robot's fitness is the robot's capacity to shift appropriately from one activity to the other, which is a motivational capacity. What is interesting is that we find differences in all three capacities – two cognitive capacities, approaching food and fleeing away from the predator, and one motivational capacity, shifting from looking for food to fleeing away from the predator when the predator appears – between the two populations of robots, those with the emotional circuit and those without the emotional circuit. If we test the robots in an experimental laboratory which contains only food but no predator or only a predator but no food, we find that in both cases the robots with the emotional circuit are better than the robots without the emotional circuit. The robots with the emotional circuit eat more food than the robots without the emotional circuit when food is their only problem and they are better able to escape from the predator when their only problem is the predator. The existence of an emotional circuit does not only help the robots' brain to take more correct motivational decisions but it leads to a better division of labour between the cognitive and the motivational/emotional parts of their brain, with advantages also for the cognitive part.

What is also interesting is how, in their natural environment, the two types of robots react when the predator appears. As soon as the predator appears and is perceived by the robots, the robots without the emotional circuit for a while do not react to the predator but they continue to eat. Only when the predator comes closer, they cease to eat and they flee away from the predator. The behaviour of the robots which have an emotional circuit in their brain is different. These robots cease to look for food as soon as they see the predator, and they initially run in a randomly chosen direction.

Only when the predator approaches them, they run in the direction opposite to the direction from which the predator is approaching. This shows that the robots with the emotional circuit are able to shift from one activity (looking for food) to another activity (reacting to the sight of the predator) more quickly than the robots without the emotional circuit and, although their immediate "emotional" reaction to the predator may not be "rational", it is very effective. So, rational behaviour may not always be the most adaptive behaviour.

We find support for this analysis if we look at what happens inside the neural network of the robots endowed with the emotional circuit. The emotional circuit of these robots is made of only one neuron and this neuron is activated as soon as the predator appears and is perceived by the robot. However, when the predator approaches the robot, the emotional neuron ceases to be activated and the control of the robot's behaviour shifts to the regular internal neurons which process the more specific information concerning the direction from which the predator is approaching, allowing the robot to go in the opposite direction. If we analyze the activation level of the emotional neuron we find that the activation level of this neuron co-varies with the distance of the predator, not with the direction from which the predator arrives. This means that the emotional neuron controls the robot's behaviour when the predator first appears and the robot must rapidly shift from looking for food to escaping from the predator. It determines a quick, although imprecise, response (running in a randomly chosen direction) to the first appearance of the predator. When the predator comes closer, the control of the robot's behaviour shifts from the emotional circuit to the cognitive circuit – the regular internal neurons – which encodes the direction from which the predator is approaching the robot. Based on this more specific information the motor neurons produce a movement in the opposite direction with respect to the predator – which, of course, is a very adaptive response.

Another way to understand the role that the emotional circuit plays in the robots' behaviour, is to lesion the emotional circuit. Our prediction is that lesioning the emotional circuit should result in behaviour that is even less "fit" than the behaviour of the robots that do not have the emotional circuit. To test this idea we lesion (eliminate) the emotional circuit of the robots that must eat food and avoid being killed by the predator and we test the lesioned robots in an environment that contains only the predator. The results show that the lesioned robots are even less able to avoid being reached by the predator than the robots that never had an emotional circuit. The existence of the emotional circuit leads to a different overall organization of the robot's neural network and to a different distribution of tasks between the cognitive circuit constituted by the regular internal neurons and the emotional circuit with its special neurons. (Cf. our discussion above of the better division of labour between the cognitive and the motivational level of functioning if the brain includes an emotional circuit.) This may explain why human beings, who have a complex emotional regulation of their behaviour, may exhibit very unfit (pathological) behaviour when their emotional apparatus does not function well.

The robots described so far have conflicting motivations but the stimuli that arrive to their brain from the environment are each associated with one single motivation. For example, food is associated with the motivation to eat and the predator with the motivation to flee away. Therefore, the stimuli of our robots are motivationally unambiguous. But stimuli can be motivationally ambiguous. We construct another type of robots that live in an environment in which the predator does not appear at randomly chosen times but it appears immediately after the robot has eaten a food token. (Perhaps, when the robot eats a food token, this causes some noise which is heard by the predator.) For these robots food is motivationally ambiguous. It is associated with both the motivation to eat and the motivation to avoid being killed by the predator because eating food announces the arrival of the predator. These robots respond to the food tokens in a different way than the robots which live in an environment in which the predator appears at randomly chosen times. When they see a food token, they approach the food token but when they are near to the food token they hesitate and they eat the food token only when their brain is informed that there is very little energy in their body – when they are very hungry. For these robots the food tokens are not only motivationally ambiguous but they are motivationally conflicting. They give rise to two motivations that, to be satisfied, require different and incompatible behaviours – and motivationally conflicting stimuli can cause pathological behaviours.

These robots for which food induces conflicting motivations are interesting for another reason. The emotional circuit is activated only by the sight of the predator because only the sensory neurons encoding the presence of the predator are connected to the emotional neuron. We now evolve another type of robots with a somewhat different neural network. In the neural network of the new robots the sensory neurons encoding the presence and location of food, not the sensory neurons encoding the presence and location of the predator, are connected to the emotional neuron. The emotional circuit of these robots is activated by the sight of food, not by the sight of the predator. What we find is that, for these robots, possessing an emotional circuit which is activated by the perception of food increases the robots' fitness more than possessing an emotional circuit which is activated by the perception of the predator. The predator is motivationally unambiguous but food evokes conflicting motivations and, therefore, the robots are more helped in their motivational decisions by the emotional reactions evoked by food rather than by those evoked by the predator.

So, emotions exist because they help the robots to make better motivational decisions. And we predict that, if encode in the robots' genes the probability of having an emotional circuit in their brain, this probability will increase in a succession of generations and at the end all the robots will have an emotional circuit. As we have said, emotions also have their dark side because, in some circumstances, they may induce the robots to behaviours which are contrary to their survival and well-being and to psychological pathologies. But, overall, the advantages of having emotions that help animals and human beings to make better motivational decisions must be greater than

the disadvantages because, otherwise, emotions would disappear from the adaptive pattern of animals and human beings, while they are a crucial component of their adaptive pattern.

5. Motivations, emotions, and covert attention

Motivations and emotions play a crucial role in attention. At any given time both nonhuman animals and human beings receive many different sensory inputs from the environment and they must ignore most of these sensory inputs and respond only to those sensory inputs which are relevant to satisfy their current motivation. This capacity is called attention. Attention can manifest itself externally, in the movements of the body, face, and eyes, and this is what psychologists call "overt" attention. An animal turns its body, face, or eyes towards what is relevant to satisfy its current motivations or to shift to another motivation. But attention can also be something which happens inside the brain, without any movement of the body – and this is what psychologists call "covert" attention. There is some work on robots that have overt attention – they move their body, their face, and their eyes to better perceive what is important for what they must do – but almost no work on robots which have covert attention. In this section we describe robots which have covert attention. Since our robots are neuro-robots, it is possible to know what a robot is attending to by examining not the robot's body but the robot's brain. The robot's brain processes only some of the sensory inputs that in each moment arrive to its sensory organs from the environment or from its own body. The robot covertly attends to these stimuli.

To demonstrate the existence of covert attention in our robots and its relation to motivations, we return to the robots that live in the environment in which the water tokens are less abundant than the food tokens and, therefore, their motivation to drink is intrinsically stronger than their motivation to eat. If we look at the robots on the computer screen, we see that, when they perceive a water token and a food token at the same time, the robots go towards the water token and ignore the food token. This is overt, or behavioural, attention. To measure covert, or purely neural, attention, we bring the robots into an experimental laboratory and, in different trials, we expose the robots to a water token alone, to a food token alone, and to a water token together with a food token – and we examine the pattern of activation of the robot's internal neurons as soon as the robot perceives the stimuli and before it does anything with its body. Since for these robots the motivation to drink is intrinsically stronger than the motivation to eat, we expect that, when a robot perceives a water token together with a food token, the pattern of activation of its internal neurons will be similar to the pattern of activation of the same internal neurons when the robot only perceives a water token because the robot's brain processes only the sensory input from the water token and, therefore, the robot "sees" only the water token. On the contrary, when the robot

perceives a food token alone or together with a water token, the patterns of activation of the internal neurons will be different because in one case the robot is seeing a food token and in the other case a water token. To compare the patterns of activation of the internal neurons, we calculate the difference between the activation levels of each pair of neurons and then we make the average of these differences.

Covert attention can also be studied at a later stage of processing of the sensory input by comparing the patterns of activation not of the internal neurons but of the motor neurons of the robot's neural network. This is still covert attention because the pattern of activation of the motor neurons have not yet been translated into a movement of the robot's two wheels. The robot is preparing to move its body but its body has still not moved.

In both cases we find the same results (see Figure 13). The robot perceives the same thing when the water token is seen alone or together with a food token and it perceives two different things when the robot sees a food token alone or together with a water token. And the effect is stronger at the level of the motor neurons than at the level of the internal neurons.

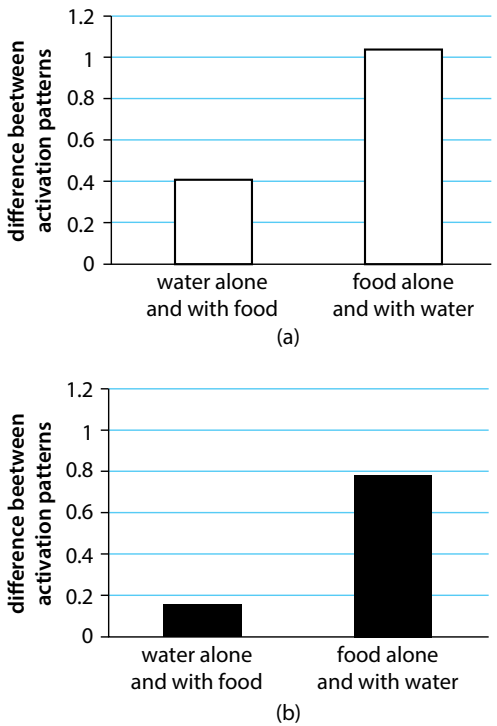


Figure 13. Difference between the patterns of activation of the internal (a) and motor (b) neurons when the robot sees a water token alone or together with a food token and when it sees a food token alone or together with a water token.

If, as we have shown, motivations have a role in attention, emotions – which are a sub-mechanism of the motivational level of behaviour – will probably also have a role in attention. And, in fact, we find that covert attention as the selective processing of one part of the total sensory input plays a more important role in robots which have an emotional circuit in their brain than in robots which do not have the emotional circuit. We measure covert attention by examining the patterns of activation of the motor neurons of the robots' neural network as soon as they see something and before they do anything with their body, and we do this with the robots that must both eat and flee away when a predator appears. We already know that possessing an emotional circuit is advantageous for these robots because the emotional circuit makes them shift more rapidly from pursuing the motivation to eat to pursuing the motivation to escape from the predator when the predator appears. We now test the robots in an experimental laboratory by exposing them for only one cycle to a predator alone or to a predator together with a food token. The results are that, for the robots without the emotional circuit, their motor neurons have a pattern of activation which is somewhat different when food is present and when food is not present. This means that the robots' brain continues to process the food token even if the predator has appeared. In contrast, the robots with the emotional circuit have more or less the same pattern of activation of their motor neurons independently of the presence or absence of food. These robots shift their covert attention from food to predator as soon as the predator appears (see Figure 14).

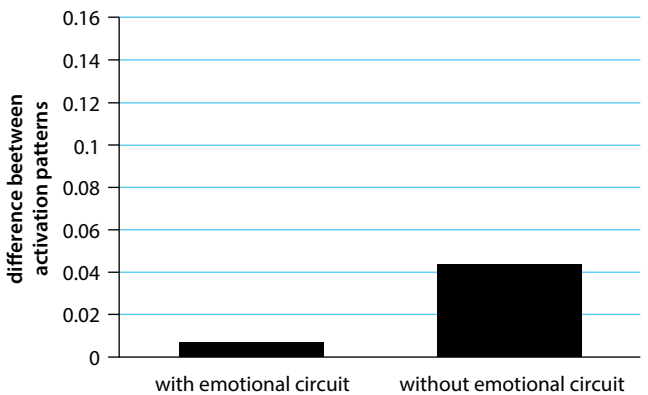


Figure 14. Difference between the patterns of activation of the motor neurons of the robots that have and do not have the emotional circuit. The robots with the emotional circuit almost completely ignore food as soon as they see the predator.

As we have seen, the emotional circuit has a positive influence not only on the motivational decisions of the robots but also on their behaviour. The robots with the emotional circuit are better at reaching the food tokens than the robots without the emotional

circuit. Now we ask: What happens in the brain of the two types of robots when they first see a food token? We compare the patterns of activation of the robots' motor neurons when the robots see a food token and when they see nothing. (Remember that the motor neurons have an intrinsic level of activation and, therefore, they are active even when the robots do not see anything.) The results are that the robots endowed with the emotional circuit pay more attention to food – there is a greater difference in the pattern of activation of the motor neurons when they see the food token and when they see nothing – than the robots without the emotional circuit (see Figure 15). Paying attention to food is necessary to approach and reach food, and this explains why the robots endowed with an emotional circuit are better than the robots without the emotional circuit at approaching and reaching food and they eat more.

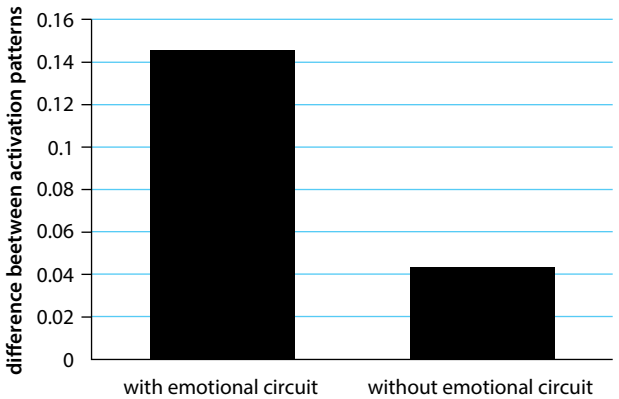


Figure 15. Difference between the patterns of activation of the robots' motor neurons when the robots see a food token and when they see nothing. The robots with the emotional circuit pay more attention to food even if there is no predator in sight.

We have also tested in the laboratory the robots which must both eat and find a mating partner for reproduction. In the experimental laboratory a robot sees both a food token and a potential mating partner and in different tests we vary the distance of the mating partner from the robot while maintaining the distance of the food constant. The results are that the robots with the emotional circuit pay attention to food if the mating partner is very distant but they tend to ignore food when the mating partner is closer. For the robots without the emotional circuit the difference is smaller while the robots which do not have the emotional circuit are unable to control their attention in the same effective way (see Figure 16).

So, motivations and emotions influence both overt and covert attention in our robots. When a robot has decided to satisfy one specific motivation, its body moves in a way that allows the robot to perceive what is relevant to satisfy its current motivation. This is overt attention. But there is also covert attention. Covert attention is

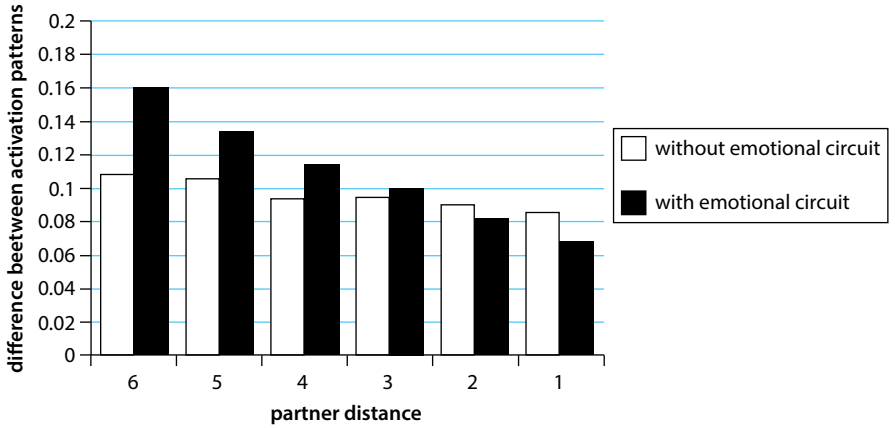


Figure 16. The robots endowed with an emotional circuit pay attention to food if the mating partner is very distant but they shift their attention from food to the mating partner if the mating partner is closer. The robots without the emotional circuit are less able to do so.

not movements of the robot's body but it is what happens in the robot's brain when a robot perceives two different things only one of which is relevant to its current motivation. The robots' brain processes only the information concerning what is relevant for its current motivation. And our results show that emotions also influence covert attention. By influencing the strength of the different motivations, the emotional circuit contributes to filtering appropriately the sensory input – which is the function of covert attention.

6. Internal robotics

We have constructed robots to test two hypotheses. The first hypothesis is that both animals and human beings function at two levels, the motivational level and the cognitive level, and to survive, reproduce, and live well they must function well at both levels. At the motivational level they must be able to choose which motivation should control their behaviour in any given moment and, at the cognitive level, they must be able to execute the behaviours that allow them to satisfy the motivation chosen at the motivation level. The second hypothesis is that emotions exist because they help the motivational decision mechanism to make better and faster motivational decisions. This is the adaptive function of emotions and this explains why they have emerged evolutionarily. By incorporating these two hypotheses in our robots, we have shown that the robots exhibit behaviours which have some resemblance with the behaviours of real organisms and, therefore, the robots provide some support for the two hypotheses. And we have also shown that possessing an emotional circuit in their brain helps the

robots to function better not only at the motivational but also at the cognitive level and that motivations and emotions influence both overt and covert attention.

The next step is to reproduce with robots other phenomena connected with motivations and emotions: inter-individual differences in personality and character, psychopathological behaviours, motivations and emotions which have to do with social life, the external expression of emotions to facilitate social interaction, the construction of "emotional artefacts" such religious and artistic artefacts. Some of these phenomena will be treated in other chapters of this book but what we will do now is discuss a general implication of motivations and emotions for robotics.

The brain of our robots is extremely simple and it should be made progressively more complex so that its structure and functioning more closely match what we know about motivations and emotions in the real brain. But the brain is not enough. Motivations and emotions are not *in* the brain. They are the result of the interactions *between* the brain and the rest of the body. The emotional neurons of our robots should influence and be influenced by specific organs and systems inside their body – the equivalent of the heart, gut, lungs, and the endocrine and immune systems. But this is a task for the future. In this section we examine a more general but very important issue: today's robotics is an *external* robotics but, if robots must reproduce the behaviour of animals and human beings and, more specifically, their motivations and emotions, what is needed is also an *internal* robotics.

Animals and human beings are physical entities which exist in a physical environment and interact with the physical environment. Their behaviour depends on the external morphology of their body (size, shape, etc.), on the nature of their sensors, and on how the different parts of their body can be moved. Their knowledge of the external environment which allows them to survive and reproduce is derived from the physical interactions of the animal's body with the external environment. By moving the different parts of its body, the animal modifies the physical relation of its sensors to the external environment or the external environment itself, in both cases actively influencing the sensory input which it receives from the environment. And the external environment is known by the animal because it responds to the animal's movements in ways that can be anticipated by the animal's brain. (On this, see Chapter 4.)

Robotics is important for our understanding of the behaviour of both animals and human beings because it makes entirely clear that animals and human beings are physical entities which exist in a physical environment and interact with the physical environment. But robotics so far has been an *external* robotics. External robotics is concerned with the interactions of the robot's body with the external environment. The sensory inputs which arrive to the robot's neural network originate in the external environment and the neural network's outputs have an effect on the external environment. But the external environment is only one of the two sources of input to an animal's brain and only one of the two environments with which the brain interacts. An animal's brain interacts not only with what is outside the animal's body – the external

environment – but also with the organs and systems which are inside the animal's body – the internal environment (see Figure 17). This means that what we need is not only an external robotics but also an *internal* robotics. If we want to understand the behaviour of an animal by constructing a robot that behaves like the animal, we must reproduce in the robot not only the external morphology of the animal's body and the interactions of the animal's brain with the external environment but also the internal structure of the animal's body and the interactions of the animal's brain with what is inside the body.

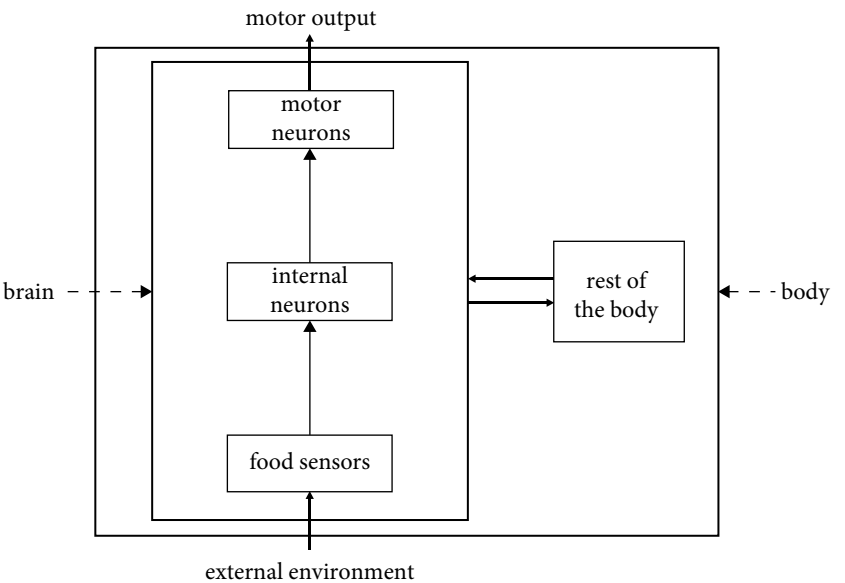


Figure 17. An animal's brain interacts not only with the external environment but also with what is inside the animal's body.

An internal robotics also requires that we explicitly recognize that the brain is only one part of the nervous system and that an animal's behaviour is the result of the interactions of the entire nervous system, and not only of the animal's brain, with the external environment and with what is inside the body beyond the nervous system. Current neural networks are models of the brain, not of the entire nervous system. The nervous system is made of a central part (brain and spinal chord) and an autonomic part (sympathetic, parasympathetic, and enteric sub-systems) and these different parts have different properties and different functions. An internal robotics requires that the behaviour of a robot be controlled by an entire artificial nervous system, not only by an artificial brain (see Figure 18).

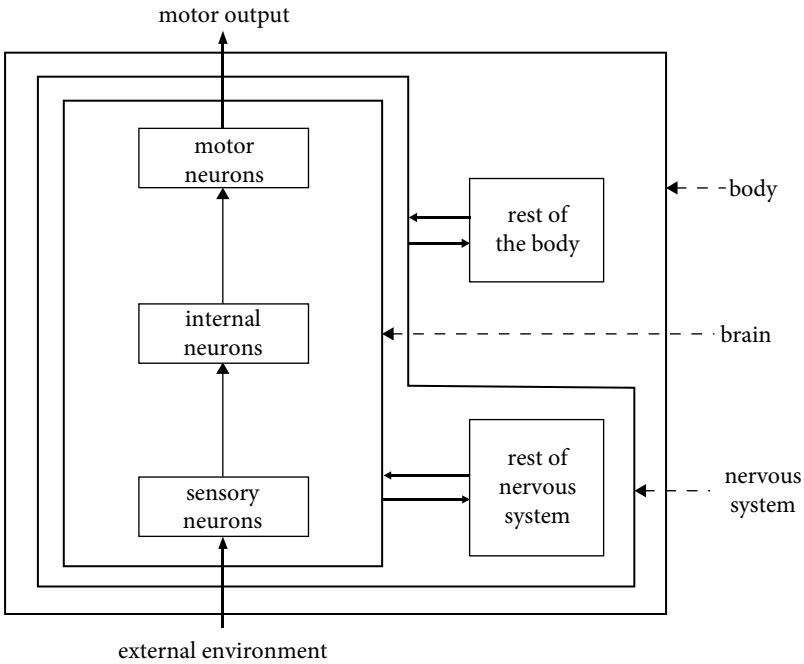


Figure 18. The entire animal.

This can be summarized by saying that animals live in two worlds, an external world and an internal world, and robots should help us understand both the external world and the internal world of animals. (Human beings also live in a third world, a mental world. We will discuss the mental world of human beings in Chapter 5.) Here is a list of the differences between the two worlds which need to be reproduced with robots, especially if the robots must be human robots.

1. *The brain's interactions with the external environment are predominantly physical interactions whereas its interactions with the internal environment are predominantly chemical interactions*

The effects of the brain on the external environment are mostly due to the movements of the animal's body. The brain encodes the movements of the eyes, the head, the arms, the hands, the legs, the phono-articulatory organs, and these movements and their effects can be studied by remaining within the boundaries of the discipline of physics. More rarely, at least in human beings, the brain has effects on the external environment which are mediated by the production of chemical molecules which diffuse in the external environment, for example, odours. Also, the sensory inputs that arrive to

the brain from the external environment also are predominantly caused by physical energies (light, sound) whereas only in some cases the external environment has effects on the brain through chemical molecules (smell, taste).

On the contrary, the interactions between the brain and what is inside the body are mostly chemical interactions and they need to be studied with the help of chemistry. Many effects of the brain on what is inside the body are due to the production of chemical molecules by the brain that affect various structures inside the body or to the stimulation by the brain of internal organs and systems which produce chemical molecules. In its turn, the rest of the body influences the brain by producing chemical molecules and sending these molecules to the brain – although the body also sends to the brain inputs which are similar to the inputs from the external environment such as the somatosensory and proprioceptive inputs caused by postures and movements of the body and of its internal organs.

This difference between the mostly physical interactions of the brain with the external environment and the mostly chemical interactions of the brain with the rest of the body has an important implication for research on robots and on robots' brains. Research on neural networks tends to consider *cellular* biology as the biological discipline of reference. A neural network's "units" are cells (neurons) and the connections between these "units" are extensions of cells (the neurons' axon). Studying the interactions of the brain with what is inside the body necessarily requires that we go one step down in the hierarchy of biological entities and we confront ourselves with *molecular* biology. An internal robotics would bring research on neural networks more in line with recent developments in biology and medicine which also tend to go one step down and to analyse and explain biological phenomena at the chemical level.

Traditional robots have mechanical bodies. They are made of inert physical matter and not of living matter. To eliminate some of the limitations of mechanical bodies from the point of view of their movements, the effects of their movements on the environment, and the probability of breaking, today some roboticists construct "soft" robots, robots made of physical matter that resembles living matter, with its flexibility and its greater resistance to breaking. But soft robotics is still an external robotics. It is the external body of the robots and its physical interactions with the external environment which one tries to make softer and more flexible. An internal soft robotics is still non-existent and, as we have said, it would require reproducing not only the physical properties of biological matter but also its chemical properties.

2. *Two different types of influences on the brain*

While a neuron of a neural network influences another neuron because there is a connection between the two neurons, not because the two neurons are close to each other in physical space, the action of the internal organs and systems on the brain is often mediated by the arrival of chemical molecules which diffuse in large zones of the intercellular space and affect in similar ways a large number of spatially close neurons rather than targeting single neurons. The brain which interacts with the external

environment can be modelled as a network of neurons which specifies which neuron is connected with which neuron independently from their location in space. The brain which interacts with the body's internal organs and systems is a structure in which where a neuron is located in space and the distance between different neuron play a critical role. In the interactions of the brain with the external environment, signals transmit information from one neuron to another neuron (neuro-transmission). In the interactions of the brain with what is inside the body, signals modulate this transmission of signals (neuromodulation).

3. *While the interactions of the brain with the external environment are interactions with an independent entity, the interactions of the brain with the rest of the body are interactions between two entities which are both evolved and which have evolved together*

The "brain-external environment" circuit passes through a physical system, the external environment, which has not co-evolved with the brain and which therefore has independent characteristics which have not been shaped by its interactions with the brain. The "brain-rest of the body" circuit goes through a physical system, the rest of the body, which has evolved together with the brain and, therefore, it has characteristics which have been shaped by its interactions with the brain. The external physical environment is the same for all species of animals that live in the same environment – although the behavioural environment of a species of animals is different from the behavioural environment of another species. On the contrary, the internal physical environment – the inside of the body – is different in different species because the brain and the rest of the body have co-evolved and are co-adapted.

From this point of view human beings are different from nonhuman animals. For human beings not only their internal environment but also their external environment co-evolves with their behaviour because with their behaviour they modify the external environment. Although some nonhuman animals modify the external environment – for example, birds construct nests – human beings are unique in that they almost entirely create the environment in which they live. Therefore, if we want our robots to be human robots, the robots must modify the external environment to make the external environment more adapted to them rather than only modify themselves to become more adapted to a fixed external environment.

4. *For the brain the external environment can be present or absent and it can be very different at different times while the internal environment of the body is something which is always present and is always more or less the same*

The rest of the body is a constant source of sensory inputs to the brain, creating a world which is always present and which remains basically the same at all times. On the contrary, the external environment creates a world which can be present but also

absent – for example, when one sleeps – and which can be very different at different times because the animal moves from one environment to another environment or because the environment changes independently of the animal. The stability and constant availability of the rest of the body can be the basis for the emergence in the brain of a "self" which is opposed to the external environment as "nonself".

5. *The causal influences originating from within the body can result in the emergence of a private world whereas those originating in the external environment define a public world*

For physical reasons the events and processes that take place inside the body cause effects only in the brain of a single individual, the individual which *owns* or *is* the body, whereas the events and processes that take place in the external environment can cause the same effects, or similar effects, in the brains of many individuals. This is the basis for the emergence of a distinction between a private and a public world, although this distinction may only exist for animals which have a sophisticated sociality such as human beings. For example, only human beings know that a certain sensory input originating from inside their body will produce effects on their brain but not on the brain of other individuals whereas a sensory input originating in the external environment will produce effects both on themselves and on other individuals. Therefore, we can reproduce the distinction between a "private" world and a "public" world only if we develop both an internal robotics and a social robotics. (Medical and information technologies make what happens inside an individual's body accessible to other individuals but they do not eliminate the distinction between a "private" world and a "public" world – although this may change in the future.)

6. *The cognitive components of behaviour emerge from the interactions of the brain with the external environment whereas its motivational and emotional components consist in the interaction of the brain with the rest of the body*

The cognitive components of behaviour can to a large extent be explained by considering only the interactions of the brain with the external environment. It is in these interactions that an individual acquires and exhibits its abilities, its problem solving capacities, its intelligence. On the contrary, the motivational and emotional components of behaviour are due to the interactions of the brain with what is inside the body and we cannot explain these components if we ignore what is inside the body. Only the external aspects of emotional phenomena – the production of expressive movements or the recognition of the expressive movements of others – can be reproduced in robots if the robots' body does not have an "inside" and their brain does not interact with what is inside their body. To construct robots that actually feel and recognize emotions, and do not only "pretend" to feel and recognize emotions, an internal robotics is necessary.

7. *The interactions of the brain with the external environment produce effects on the external environment that human beings are able to predict and, therefore, can be voluntary, whereas the interactions of the brain with what is inside the body give rise to effects which human beings are generally unable to predict and are involuntary*

In human beings the interactions of the brain with the external environment tend to produce voluntary behaviours and voluntary effects on the external environment. On the contrary, even for human beings the effects of the brain on the rest of the body are involuntary. The voluntary character of behaviour is linked to the ability to predict its effects and, while the effects of behaviour on the external environment can be within limits predicted, this is rarely the case for the effects that one's brain has on what is inside one's body. (For robots that predict, see Chapter 4 on mental life.)

But human beings invent strategies that make the effects of their brain on the rest of their body and the effects of the rest of their body on their brain more predictable and controllable. One strategy is to acquire a greater knowledge/control of the internal states of one's body through meditation and other forms of body/brain control techniques. Another strategy is exposing oneself to sensory inputs from the external environment which have effects on one's brain which in turn have effects on the rest of one's body. Examples are talking with others – relatives, friends, psychotherapists – and exposing oneself to religious and artistic artefacts.

Let us summarize what we have said in this section. Human beings live in two worlds: the external, cognitive, public, predictable, and controllable world which results from the interactions of their brain with the external environment, and the internal, affective, private, much less predictable and controllable world which results from the interactions of their brain with what is inside the body. External robotics constructs robots which have an external morphology and external sensory and motor organs and is interested in the interactions of the robot's brain with the external environment. Internal robotics constructs robots which, in addition to an external morphology and external sensory and motor organs, have internal organs and systems and is interested in the interactions of the robot's brain with these internal organs and systems. If we want to construct robots which are really like real animals and real human beings, what we need is not only an external robotics but also an internal robotics.

7. Robots with a biological clock

We have discussed internal robotics in this Chapter because motivations and emotions are based on the interactions of the brain with what is inside the body and, therefore, to reproduce motivations and emotions in robots an internal robotics is necessary. Today's robots are bodies without internal organs and systems but future robots will

have to have internal organs and systems and their emotional neurons should interact with these internal organs and systems.

An internal robotics is a future robotics but in this section we describe robots which reproduce one important aspect of internal robotics: unlike the external environment, the internal environment of robots co-evolves with their brain.

The new robots are robots that sleep. They live in an environment with food tokens which they must eat to remain alive but in their environment periods of light (day) are regularly and cyclically followed by periods without light (night) (see Figure 19).

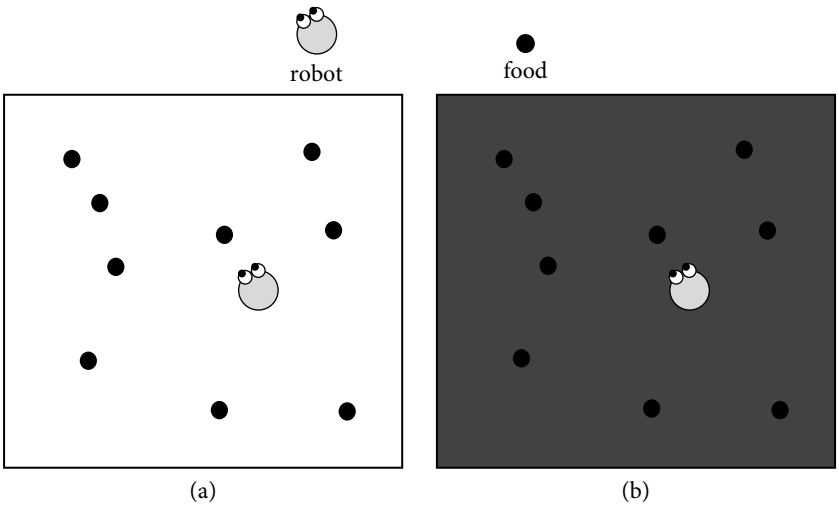


Figure 19. The robots live in an environment in which periods of light (day) (a) are followed by periods of no light (night) (b). At night the robots cannot see the food tokens.

When there is no light, the robots cannot see the food tokens and, therefore, since moving consumes their bodily energy, at night the robots should stop moving to avoid consuming their bodily energy uselessly. And, in fact, at the end of evolution we find that the robots during the day explore the environment looking for food but, at night, they stop moving until the next day begins. They sleep.

The robots' neural network is that of Figure 20. The robots have visual neurons that encode the location of the nearest food tokens and they also have light/dark neurons that are activated during the day but are not active at night. The night does not begin or terminate abruptly but gradually, and this is reflected in the activation level of the light/dark neurons which are maximally activated at mid-day and become progressively less activated when the night begins (evening) and more activated when the night ends (dawn). During the night random noise is added to the activation level of the robots' food visual sensors so that the robots cannot see the food tokens. This means that to save their energy, at night the robots should stop moving (sleep).

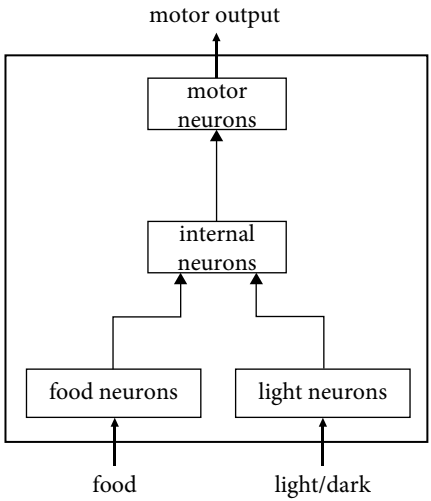


Figure 20. The neural network of the robots that live in an environment in which periods of light (day) are followed by periods without light (night).

These robots must satisfy two motivations, the motivation to eat during the day and the motivation to sleep at night, and they evolve the capacity to decide correctly when to move in search of food and when to sleep based on the information provided by the light sensors which tells them when it is day and when it is night. For a certain number of generations the robots evolve in the environment we have described but then they migrate to a new environment which is identical to the preceding environment except that it contains a certain number of caves. A cave is a zone of the environment in which the light of the day does not penetrate and therefore it is always dark (see Figure 21). This poses a new problem for the robots. When they happen to enter a cave, there is no light in the cave and, therefore, they believe that the night has arrived and they should stop moving. Clearly, this behaviour is not adaptive because what they should do is continue to move in order to go out of the cave. They don't see the food existing in the cave and, therefore, they cannot eat and they risk remaining all the rest of their life inside the cave because the night never ends for them.

To solve this problem the robots evolve a biological clock. We add another set of internal neurons to the neural network of the robots and, since these neurons are connected to the motor neurons, they influence the robots' behaviour (see Figure 22).

The new neurons are not activated from outside but they cyclically activate themselves. The length of a cycle – biological clock – is not decided by us but it is encoded in the robots' inherited genes. When the robots migrate to the environment with caves, the length of the cycle of their biological clock is random and is different for different robots and it rarely coincides with the day/night cycle. As a consequence, the robots

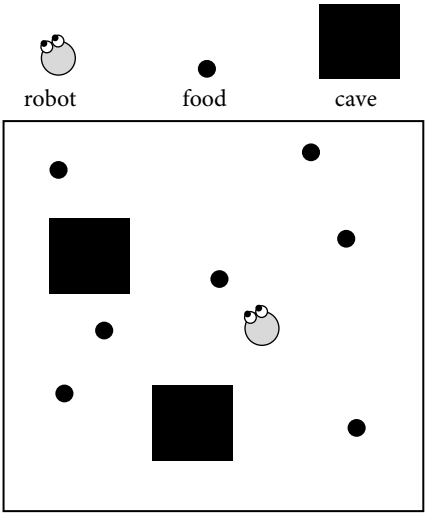


Figure 21. The new environment to which the robots migrate contains dark caves in which daylight does not penetrate.

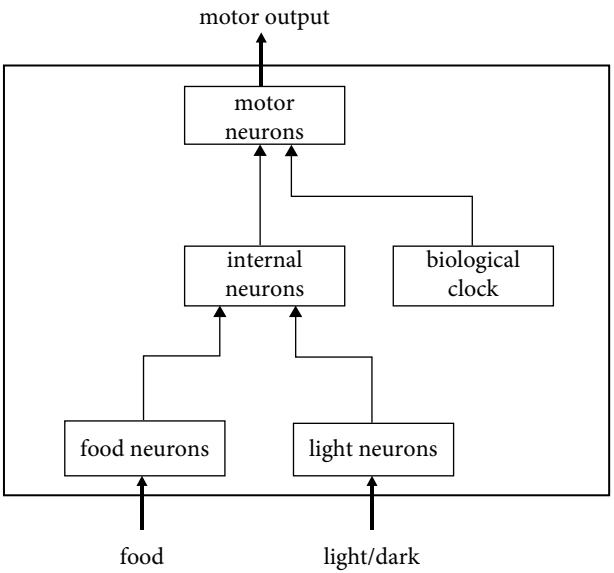


Figure 22. The neural network of the robots contains a new set of neurons (biological clock) which play a role in determining the robots' behaviour.

have a low level of fitness because, when they enter a cave, they think it is night-time and they tend to remain inside the cave for ever. Their light/dark neurons are inactive and, for them, this signals night. But after a certain number of generations the situation changes. The selective reproduction of the robots with the best biological clocks and the addition of random mutations to the genes which encode the length of the cycles of the biological clock gradually improve the quality of the robots' biological clock. The biological clock now tells the robots' brain when it is night and when it is day independently of the activation of their light/dark neurons. The robots' brain uses this information so that the robots stop moving at night, whether they are inside or outside a cave, and they move out of a cave when it is day, even if they see no light. The motivational decisions of the robots are influenced by their biological clock and the biological clock is part of their body.

The motivational decisions of the robots described in this Chapter are influenced by sensory inputs from inside their body: hunger, thirst, and pain. But the robots with a biological clock tell us something new. The input from the body of our preceding robots is hardwired by us. When the robot's bodily store for energy becomes progressively empty, *we* decide that the robot's hunger neurons are progressively activated, and the same for the thirst neurons. When their body is physically damaged, *we* decide that their pain neurons are activated. The biological clock of the new robots is not hardwired by us. It spontaneously evolves in a succession of generations.

This is what we wanted to show with these robots that sleep. While the environment of external robotics is the external environment, and the properties of the external environment are independent of the robot, the environment of internal robotics is what is inside the robot's body, and this internal environment evolves together with the robot's brain and the robot's behaviour. All the sensory organs of an animal are evolved, both its external and its internal sensory organs. But for the internal sensory organs what evolves are *both* the sensory organs and the source of the input to them.

8. The two halves of the mind

The moral of this Chapter is that the human mind is made of two halves, the cognitive half and the motivational/emotional half, and science cannot claim to have understood and explained human behaviour if it does not understand and explain both halves of the mind and how the two halves interact together. Science is a product of Western culture and, in Western culture, cognition is more important than motivations and emotions because cognition is control of nature and independence from nature while motivations and emotions are lack of control and dependence on nature. This explains why the science of the mind tends to emphasize cognition rather than motivations and emotions and why, today, psychology is becoming cognitive science and the "psyche" "cognition".

There are also technical reasons why the science of the mind prefers to study cognition rather than motivations and emotions. For science, the best empirical method for studying the mind – or anything else – is to do laboratory experiments, and this is why experimental psychology is considered as the most scientific part of psychology. But while the cognitive half of the mind can be studied in the experimental laboratory, this is less true for the motivational/emotional half of the mind. In the experimental laboratory "subjects" or "participants" cannot choose which motivation to satisfy with their behaviour because what they must do is simply obey the experimental instructions and, therefore, satisfy the *experimenter*'s motivations. In addition, the motivational/emotional half of the mind is strictly linked to the evolutionary history of the human species and in the experimental laboratory it is possible to study learning and behaviour but not evolution – with the exception of the evolution of extremely simple animals with a very short life.

Another reason why experimental psychologists tend to study the cognitive half of the mind and not its motivational/emotional half is that the cognitive capacities of an individual can be measured quantitatively by using intelligence tests and other tests while his or her motivational/emotional characteristics are more difficult to identify, test, and quantify. Science likes what can be objectively measured. Psychologists want to be scientists and, therefore, they inevitably tend to see the cognitive half of the mind as constituting the entire mind.

The science for which the cognitive half of the mind explicitly is the entire mind is the science of economics. Economists – with some exceptions – assume that human beings decide rationally not only what they must do to satisfy a given motivation but also which motivation to satisfy with their behaviour. But human beings can decide rationally which motivation to try to satisfy with their behaviour only if they explicitly predict the consequences of their actions – and only some human beings do this and only in some (rare) cases. Not only the human mind is made of two halves – the motivational/emotional half and the cognitive half – but the motivational/emotional half is more important than its cognitive half because it is more ancient and because is the basis on which the entire adaptive pattern of animals, including human beings, is constructed. If science ignores the motivational/emotional half of the mind, it cannot understand how behaviour is primarily determined by motivations and emotions and it cannot understand to what extent economic behaviour is influenced by techniques that try to change the motivations of buyers, not to make them think and behave better. Robots must have both halves of the mind. They must be Artificial Life, not Artificial Intelligence. Intelligence is only one half of the mind. Life is the entire mind. If we construct robots that have both the cognitive and the motivational/emotional halves of the mind we can know both halves of the mind with the precision and objectivity which is required by science.

CHAPTER 3

How robots acquire their behaviour

The Earth has a long history and at the beginning of this long history what existed were only phenomena that today are studied by physicists. Then new phenomena appeared which, to be understood and explained, require chemists, biologists, and zoologists. Only much more recently human beings have come into existence, and human beings are studied by a variety of scientists: psychologists, anthropologists, sociologists, economists, political scientists, and historians. ME thinks that, to understand any phenomenon of reality, it is necessary to know its past history and how it has become what it is, and that this is particularly true for those phenomena which have appeared more recently. Human beings and their societies are the most recent phenomena of reality and, therefore, to understand them ME needs to construct robots that reproduce not only how human beings behave and how their societies function but also the past history of their behaviour and their societies. ME is interested in all aspects of the human past: the evolutionary history of *Homo sapiens*, how a human being matures and develops in the course of his or her life, how the experiences of an individual affect the individual and make the individual learn new behaviours, the origin and history of human societies and human cultures. Reproducing all these changes is a complex task and ME decides to begin with biological evolution and individual learning and with how evolution and learning interact together.

1. Why learning?

All animals become what they are and we cannot claim we understand them unless we know how they have become what they are. Therefore, robots cannot be programmed by us to exhibit some desired behaviour but they must autonomously acquire their behaviour. The robots described in the preceding chapter evolve their behaviour in a succession of generations and, therefore, their behaviour is entirely innate – encoded in their genes – and it remains always the same during their entire life. But biological evolution is not the only way in which animals acquire their behaviour. An individual's brain changes as a result of the individual's experiences in the environment and, therefore, the individual's behaviour also changes. This process of acquisition of new behaviours and new capacities is called learning. Learning plays a more important role

in more complex animals with a more complex behaviour and, in fact, most behaviours of human beings are learned or modified by learning. But learning occurs in all animals, even the simplest ones.

This poses an interesting question: Why do animals learn? Why evolution is not sufficient to endow an animal with all the behaviours which are necessary for its survival and reproduction? The answer is the genetic unpredictability of the environment. If the environment in which the animal will spend its life could be entirely "predicted" by its genes, evolution would be sufficient to endow the animal with all the behaviours that is needs to remain alive and have offspring. But if an animal's genes cannot "predict" the specific environment in which the individual will spend its life, evolution is not sufficient and learning is necessary.

The environment may be genetically unpredictable for a variety of reasons. The environment can change for its own reasons, or it can change because the animal moves to a new environment or, in the case of human beings, because human beings modify the environment in which they live. But the environment may be genetically unpredictable not because it changes but because an animal's adaptive pattern is so complex and the animal must be able to do so many different things that the same individual actually lives in a variety of different environments and different individuals live in different environments – and this is particularly true for human beings. In all these cases the genes cannot contain all the information which the individual's brain needs to generate all the necessary behaviours. The genes encode the general nature of the individual's behavioural capacities but learning is necessary to refine and increase these capacities so that the individual's behaviour is better adapted to the particular environment in which the individual happens to spend its life.

This is why animals do not only evolve but they also learn. Their behaviour changes in a succession of generations and it also changes during their life. But animals do not separately evolve and learn because evolution and learning interact together and influence each other. Psychologists and other students of behaviour rarely study how evolution and learning interact together. The evolution of a species of animals is difficult to know because it happens in a succession of generations and, therefore, it takes a long time to study and because it depends on the natural environment and it is difficult to collect precise and sufficient data on the present environment of a species of animals and almost impossible for its past environments. Evolution is "reconstructed" by studying animal fossils and in other indirect ways but it cannot be really studied while it is happening, except in very simple animals in experimental laboratories. On the contrary, learning is one of the phenomena most studied by psychologists, to such an extent that, for some psychologists, the behaviourists, psychology was synonymous with learning theory. But learning is almost impossible to study in the natural environment because in the natural environment the conditions in which learning occurs cannot be controlled. This is why psychologists study learning in the experimental laboratory where they can collect many precise and quantitative data and they can

control and manipulate the conditions in which learning occurs. The problem is that learning in the experimental laboratory may not be the same as learning in the natural environment and, by studying learning in the experimental laboratory, it is difficult to understand how learning is linked to the evolutionary history of the species. In the natural environment an animal is exposed to a variety of stimuli which in large part depend on the animal's own behaviour and the animal autonomously decides what to do on the basis of the current strength of its motivations. The experimental laboratory is an artificial, simplified, environment in which "subjects" respond to stimuli which are decided by the experimenter and they do what the experimenter wants them to do. Laboratory experiments have taught and will continue to teach us many important things about behaviour and learning but, if the natural environment of animals and their evolutionary history are important to understand how animals learn and what they learn, there will inevitably be significant aspects of learning which we will miss if we study learning only in the experimental laboratory.

How evolution and learning interact together is largely, even if not completely, also ignored in robotics. Evolutionary robotics is a very active area of research today. The robots live and reproduce in their "natural" environment, their behaviour changes in a succession of generations, and they acquire initially non-existent behaviour and capacities. In the preceding chapter we have seen various examples of robots that evolve and we will see many other examples throughout this book. Learning also is a very active area of research in robotics and, in fact, most work in robotics is dedicated to robots that learn. But evolution and learning tend to be studied separately and there is little interest in how learning and evolution interact together in robots. Real animals learn in their natural environment but robots learn in the equivalent of an experimental laboratory. They respond to sensory inputs which are decided by the researcher and they learn what the researcher wants them to learn. Real animals are not passive receivers of sensory inputs and they do not learn what we want them to learn. They live in a *physical* environment which is independent of them but they interact with a *behavioural* environment which is partially, but importantly, determined by their own behaviour – and what they learn and how they learn depend on their behavioural environment, not on the physical environment. If a robot learns in the equivalent of an experimental laboratory, the robot is a passive receiver of inputs which have been decided by us and it learns to respond to these sensory inputs in ways which are also decided by us. And if the robot learns based on rewards and punishments, what is rewarding or punishing is also decided by us. If we are interested in studying with robots how evolution and learning interact together, this is not what should be done because both evolution and learning are adaptations to the natural environment, not to the artificial environment of an experimental laboratory, and what is rewarding or punishing cannot be decided by us.

Evolution and learning interact together in the sense that both evolution influences learning and learning influences evolution but in this Chapter we concentrate on

how evolution influences learning. One important way in which evolution influences learning is that evolution creates the initial conditions for learning. Current robots start their learning from zero. At the beginning, the robot's neural network is assigned random connection weights, and then, as a result of the robot's experiences, the connection weights change and the robot becomes progressively able to exhibit initially absent behaviours. Real animals do not start their learning from zero. The inherited genes specify the initial state of the animal that learns and this initial state influences what is learned and how it is learned. This is one important reason why robots that learn should be robots that both evolve and learn.

Another important way in which evolution affects learning is that it is evolution which assigns positive and negative value to stimuli. Much learning depends on rewards and punishments, where rewards are stimuli that increase the probability that the behaviour which has caused them will be exhibited in the future and punishments are stimuli that decrease this probability. An animal is more likely to eat X if eating X causes stimuli which have reward value (they taste "good") while it is less likely to eat Y if eating Y causes stimuli which have punishment value (they taste "bad"). But it is evolution that decides what is rewarding and what is punishing for an animal and, therefore, it is evolution that determines what is learned by the animal. Initially neutral stimuli can acquire reward or punishment value in the course of an animal's life so that the animal will learn a very large number of new behaviours – which is typical of human beings – but the original rewards and punishments are due to evolution. Constructing robots that at the same time evolve and learn makes it possible to reproduce the role of evolution in assigning reward or punishment value to stimuli and, therefore, in determining what an animal will learn during its life.

The robots that we describe in this Chapter are robots that do not learn in the equivalent of an experimental laboratory but, like real animals, they learn in the environment in which they live. As we have said, learning is necessary when an animal's genes cannot contain all the information needed by the animal to survive and reproduce in its environment. We describe robots that evolve and learn in various environments and we show that learning becomes increasingly important for the robots' fitness as the environments in which the robots live become increasingly unpredictable genetically. The robots do not need learning if the environment is very simple and it does no change in unpredictable ways. But they need learning if the environment is more complex and it poses challenges that cannot be predicted by their genes.

Constructing robots that both evolve and learn is an application of the principle "one robot/many phenomena". Current robots *either* evolve *or* learn. Our robots *both* evolve *and* learn. But we apply the principle "one robot/many phenomena" in other ways. Our robots do not only learn in their natural environment but they also learn in the equivalent of an experimental laboratory. We extract a robot from the environment in which it lives and we bring the robot into a controlled environment in which the robot is exposed to stimuli which it has never encountered during its life – and we show that the robot learns to respond appropriately to these stimuli. Furthermore, we

reproduce not only the *fact* that animals learn but also a variety of results of laboratory experiments on learning in nonhuman animals and human beings. And we also apply the principle "one robot/many phenomena" by reproducing with our robots different types of learning by using the same model of learning. We describe this model of learning in the next section.

2. The neural network of robots that learn

The neural network that controls the behaviour of our robots is made of artificial neurons linked by artificial excitatory or inhibitory synapses called connections. The contribution of each neuron to the robot's overall behaviour depends on the neuron's current level of activation. The activation of the sensory neurons is determined by what exists or happens in the external environment or inside the robot's body. The sensory neurons excite or inhibit the internal neurons through the connections that link the sensory neurons to the internal neurons, and the internal neurons excite or inhibit the motor neurons through the connections that link the internal neurons to the motor neurons. In this way the robot responds to a given sensory input with a given motor output.

The connection which links one pre-synaptic neuron to a post-synaptic neuron can be stronger or weaker, and this is reflected in the quantitative weight of the connection. If a pre-synaptic neuron is linked to a post-synaptic neuron by a connection with a high weight, the influence that the pre-synaptic neuron has on the post-synaptic neuron is greater. If the weight is low, its influence is smaller. Therefore, how much a pre-synaptic neuron excites or inhibits a post-synaptic neuron depends on two factors: the level of activation of the pre-synaptic neuron and the positive (excitatory) or negative (inhibitory) weight of the connection which links the pre-synaptic neuron to the post-synaptic neuron (see Figure 1).

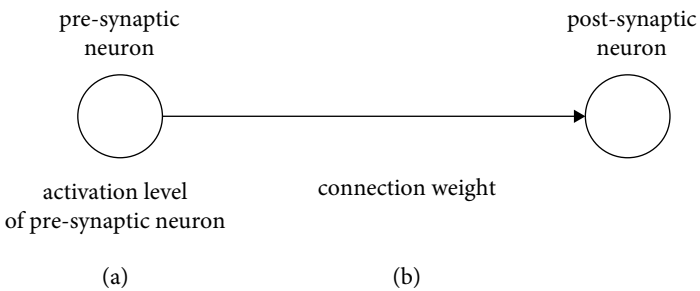


Figure 1. The contribution of a pre-synaptic neuron to the activation of a post-synaptic neuron depends on two factors: (a) the activation level of the pre-synaptic neuron, and (b) the weight of the connection that links the pre-synaptic neuron to the post-synaptic neuron.

The genes that the robots inherit from their parents specify the weights of the connections of their neural network and, therefore, they determine how the robots respond to the sensory input. If the weights do not change during a robot's life, the robot responds in the same way to the sensory inputs throughout its entire life and the robot's behaviour never changes. The robot learns if the connection weights of its neural network change because of the robot's experiences during its life. The changes in the connection weights modify the robot's behaviour and new behaviours are learned.

This is how neural networks typically learn. What changes during learning are their connection weights. The neural network of our robots that learn is different. In the brain a pre-synaptic neuron influences a post-synaptic neuron because the pre-synaptic neuron releases a certain number of molecules called neurotransmitters in the space between the two neurons – the synaptic gap – and these molecules are captured by the receptive molecules located on the external membrane of the post-synaptic neuron. The number of neurotransmitters released by the pre-synaptic neuron depends on two factors: the level of activation of the pre-synaptic neuron and the intrinsic properties of the pre-synaptic neuron which are simulated by the weight of the connection linking the pre-synaptic neuron to the post-synaptic neuron. But in the real brain there is another factor which determines how much a pre-synaptic neuron influences a post-synaptic neuron, and this other factor is how much the post-synaptic neuron is *receptive* to that particular pre-synaptic neuron. Given the same number of neurotransmitters released by the pre-synaptic neuron in the synaptic gap, the post-synaptic neuron may be more or less excited or inhibited by these neurotransmitters because it may be more or less receptive to that particular pre-synaptic neuron. The receptivity of the post-synaptic neuron with respect to one particular pre-synaptic neuron is the number of receptive molecules which exists in one particular point of the post-synaptic neuron's external membrane – the synaptic site.

The neural network of our robots that learn reproduces this property of the real brain. How much a pre-synaptic neuron excites or inhibits a post-synaptic neuron depends not on two but on three factors: (a) the activation level of the pre-synaptic neuron, (b) the weight of the connection linking the pre-synaptic neuron to the post-synaptic neuron, and (c) the receptivity of the post-synaptic neuron to that particular pre-synaptic neuron (see Figure 2).

This has consequences for learning. As we have said, in traditional neural networks that learn, what changes during learning are the weights of the neural network' connections. In our robots that learn, the weights of the connections do not change but what changes is the receptivity of a post-synaptic neuron with respect to each of its pre-synaptic neurons. The model schematized in Figure 2 only applies to a subset of the internal neurons of a neural network which we call learning neurons. The other internal neurons are the traditional neurons schematized in Figure 1 and nothing changes in these neurons during learning. But since the learning neurons are connected to the motor neurons, if the receptivity of the learning neurons changes, the robots' behaviour also changes.

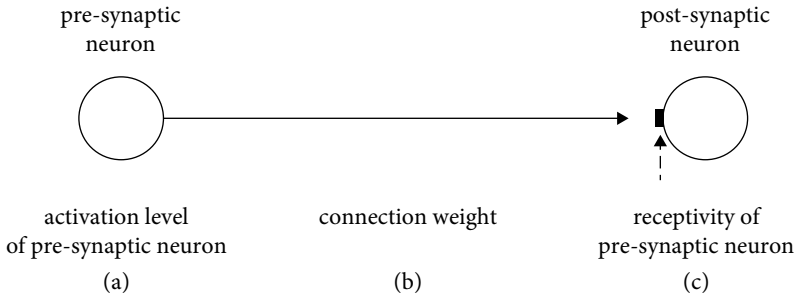


Figure 2. The contribution of a pre-synaptic neuron to the activation level of a post-synaptic neuron is a function of the two factors (a) and (b) of Figure 1 plus the receptivity of the post-synaptic neuron with respect to that particular pre-synaptic neuron (c).

How does the receptivity of the learning neurons change? When a sensory neuron excites or inhibits a learning neuron, a marker is left on the learning neuron's synaptic site for that particular sensory neuron. The marker has a value proportional to the activation arriving from the sensory neuron but this value progressively decreases unless the same sensory neuron continues to send its activation to the learning neuron in the successive cycles. The learning neuron is activated only if the total activation arriving from the sensory neurons to the learning neuron is greater than a given threshold. If the activation threshold is reached, the receptivity of the learning neuron to the sensory neurons which have left a marker on their synaptic site is modified. The receptivity increases if the activation level of the learning neuron is higher than a certain threshold and it decreases if the activation level is lower than another, lower, threshold. These changes – which are called long-term potentiation and long-term depression in the neuroscientific literature – are permanent and, since the learning neurons influence the motor neurons, the robot's behaviour is permanently modified.

In this model of learning, evolution and learning interact together. The inherited genes encode the weights of the connections of the robots' neural network and they also encode the value of the variables that characterize each learning neuron: how the value of its markers progressively decreases in the successive input/output cycles, what is its activation threshold, and when its level of activation is sufficiently high or sufficiently low to increase or decrease its receptivity with respect to the sensory neurons that have left a trace in their receptive sites.

We apply this model of learning to a population of robots that live in an environment with food tokens and poisonous tokens. The food tokens are black and the poisonous tokens are white and, to survive and reproduce, the robots must learn to eat the food tokens and to avoid the poisonous tokens. The food tokens have a sweet taste and the poisonous tokens have a bitter taste and the robots' neural network has visual neurons which encode the colour of the tokens and taste sensors which encode their taste (see Figure 3).

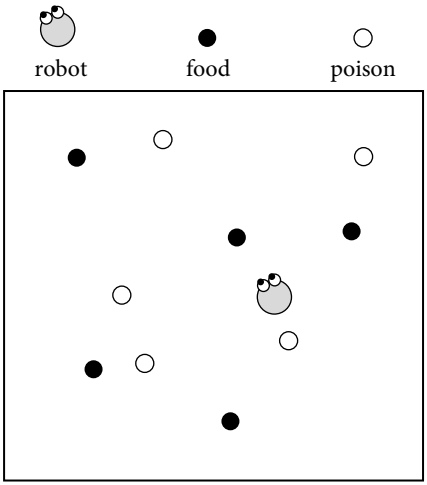


Figure 3. The environment contains both food tokens and poisonous tokens. The robots' neural network has visual neurons that encode the colour of the food tokens (black) and of the poisonous tokens (white) and taste neurons which encode their taste (sweet and bitter.)

Learning occurs because the robots' neural network associates the colour black with the taste of sweetness and the colour white with the taste of bitterness. Since the black tokens have a sweet taste, the learning neurons increase their receptivity towards the visual neurons which have left a marker on their synaptic sites so that, when the robot sees another black token, the probability that the robot will approach and eat the black token increases. On the contrary, since the white tokens have a bitter taste, the receptivity of the learning neurons is decreased and the probability that the robot will approach and eat a white token is reduced. The sweet taste is a "reward" and the bitter taste a "punishment" and what gives reward value to the sweet taste and punishment value to the bitter taste is evolution. The manner in which the learning neurons function is specified in the robots' genes and it is the result of an evolutionary process during which the robots that learn to associate the sweet taste with food and the bitter tastes with poison have left more offspring than the robots which do not learn.

Today, most robots start their learning from zero because learning consists in changes in the connection weights of their neural network and the initial value of their connection weights is random. Therefore, evolution has no role in learning. If the robots both evolve and learn, evolution creates the initial state of learning – the initial value of their neural network's connection weights which is encoded in their genes – but then evolution ceases to have an influence on the robots' behaviour because learning changes their connection weights. In our robots what changes during learning are not the connection weights but the receptivity of the learning neurons, and this implies that the genes do not only create the initial state of learning but they also assist learning and continue to play a role during the entire life of the robots.

Given our principle "one robot/many phenomena", one test of our model of learning is that with this model we can reproduce many different types of learning and many different phenomena linked to learning. This is what we show in this Chapter. But first we want to explain why the answer to our initial question "Why animals, in addition to evolution, have learning?" is the genetic predictability of the environment.

3. The genetic predictability of the environment

Learning is necessary when the environment in which an animal lives cannot be "predicted" by the animal's genes. The behaviour of an animal could be completely innate – completely determined by the animal's genes – but this would be possible only in an environment which is very simple, never changes, and never poses new challenges to the animal. If this is not true – and it is rarely true even for very simple animals – the animal's behaviour cannot be entirely innate because the information encoded in the animal's genes determines behaviours which are not always well adapted to the environment in which the animal happens to live and, therefore, it is necessary for the animal to learn.

Environments change and they change in many different ways. The environment can change for its own independent reasons. An example is climatic changes. Climatic changes can be so slow that the robots can adapt to them by changing their genes and evolving new behaviours in a succession of generations with no need for learning. Or they may be so fast that, learning or no learning, the robots become extinct. (Cf. the extinction of dinosaurs tens of millions of years ago or the extinction of many animal species caused by human beings today.) Other types of environmental changes that do not require learning are seasonal changes. The environment changes cyclically from one season to the next but these changes are always the same, generation after generation, and therefore they can be genetically predicted. An animal is born with genes that encode both the behaviours which are appropriate for one season and the behaviours which are appropriate for the next season. Or the environment in which animals live can change because the animals move to a new environment but, if their migration is slow, evolution is sufficient to produce the changes in the inherited genes which are necessary to live in the new environment. Seasonal migrations, being periodic and repetitive, are also genetically predictable and do not require learning.

Another example of genetically predictable changes is the changes due to the development of the body and of the brain in the early stages of an animal's life. An animal's environment is not the physical, "objective", environment but it is an environment which depends on the shape and size of the animal's body, on its sensory and motor organs, and on its brain. If the animal's body, its sensory and motor organs, and its brain change during the animal's life because of maturation and development, the animal's behavioural environment also changes. But these changes are always the same,

generation after generation, and, therefore, they can be predicted by the genes and by themselves they do not make learning necessary.

These are all genetically predictable changes which they do not require learning. But learning can be useful even in environments which are mostly genetically predictable. Environments are not spatially homogeneous and two individuals may live in different parts of the environment which require different behaviours. The genes cannot predict all the different micro-environments in which an animal will live and some learning is necessary. And the genes may not be sufficient even if two robots live in the same physical environment. As we have said, we must distinguish between the *physical* environment and the *behavioural* environment. Since each individual behaves differently from all other individuals because its genes are different from the genes of the other individuals, each individual lives in a different behavioural environment. This makes genetically inherited information insufficient to guarantee the adaptation of the individual to its unique behavioural environment and some learning is needed. This could be demonstrated by making two robots with the same genes live exactly in the same environment. We predict that, if the two robots do not learn, their behaviour will be exactly the same but, if they learn, their behaviour will be modified by the specific experiences that they have during their life and their behaviour will be different.

But the environment can be much less genetically predictable, and this is especially true in the case of human beings. Human beings have a very complex adaptive pattern and a very rich repertoire of different behaviours. These different behaviours make their environment very complex and populated by many different entities, and this complex environment requires learning because it is impossible for their genes to predict it. Another characteristic of human beings is that human beings change the environment in which they live by creating cultures, languages, and technological artefacts, and these changes tend to be so fast that evolution cannot keep track of them and learning is necessary. In fact, these changes may be responsible for the fact that human beings are the only animal species whose behaviour is almost completely learned. Cultures, languages, and technological artefacts require learning not only because they change continuously but also because they make the environment in which a human being lives much richer and variable. An individual's genes cannot contain all the information which is necessary to behave appropriately in an environment so rich in artefacts, behaviours, and social practices, and learning is the only solution. And this also explains why learning in human beings in many cases is teaching: it is not nature but other human beings who provide the rewards and punishments that are necessary for them to learn.

In the next two sections we describe robots that live in more genetically predictable environments and robots that live in environments which are less genetically unpredictable, and we show that no environment can be entirely predicted by the robots' genes.

4. Living in genetically predictable environments

The robots live in our basic environment that contains a certain number of randomly distributed food tokens which the robots must eat if they want to reproduce. The food tokens are black and they have a sweet taste. When a robot reaches a food token, it eats the food token. The food token disappears and a new food token appears in the environment so that the total number of food tokens is always the same. Length of life is the same for all robots and the robots which reproduce are those that eat more food tokens. The robots of the initial generation have random genes and, therefore, the robots are not very good at eating the food tokens. Then, the selective reproduction of the best robots and the constant addition of random changes to the inherited genes lead to a progressive increase in the number of food tokens eaten by the robots.

These are our basic, evolving, robots. We compare these robots with another population of robots which do not only evolve in a succession of generations but they also learn during their life. While the robots of the first population have a neural network which does not allow them to learn (see Figure 4a), the robots of the other population have a neural network which allows them to learn (see Figure 4b).

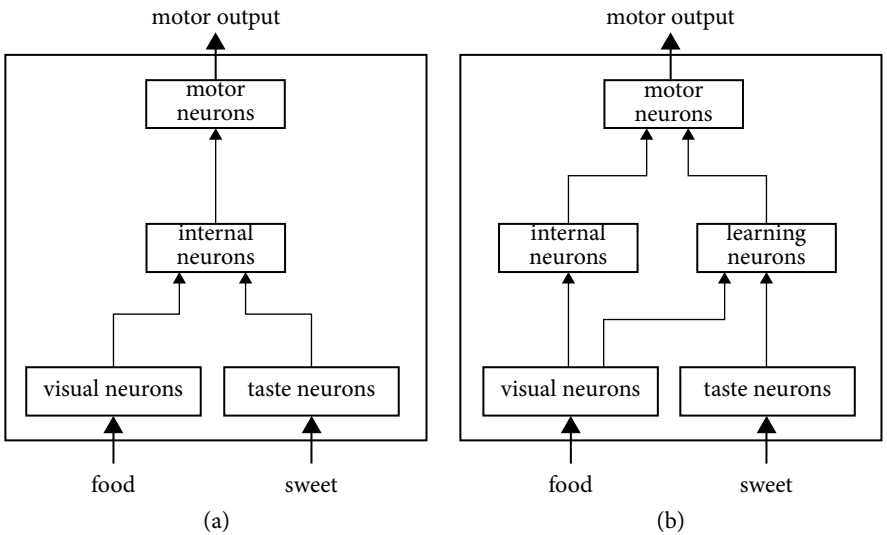


Figure 4. Neural network of robots that do not learn (a) and of robots that learn (b).

In both cases the neural network of the robots has two sets of sensory neurons: the visual neurons and the taste neurons. The visual neurons encode the direction and distance of the food tokens which are within the robot's visual field. The taste neurons encode the sweet taste of the food tokens and they are activated when the robot touches (eats) a food token. The sensory neurons send their connections to the internal neurons

which send their connections to the motor neurons. The motor neurons control the movements of the robot's body that allow the robot to displace itself in the environment. So far, the neural networks of the two populations of robots are identical. What distinguishes the neural network of the robots that learn from the neural network of the robots that do not learn is that their neural network has an additional set of internal neurons which are learning neurons: the receptivity of these neurons changes during a robot's life because of the robot's experiences in the environment. The visual sensory neurons send their connections to both the regular internal neurons and the learning neurons while the taste sensory neurons send their connections only to the learning neurons. The learning neurons send their connections to the motor neurons and, since the learning neurons change their receptivity during the robot's life, how the robots respond to food also changes.

What are the results when we compare the two populations of robots? After 500 generations the robots that learn and those that do not learn have more or less the same fitness: they eat more or less the same number of food tokens. However, the growth of the fitness in the successive generations is slightly faster for the robots that learn than for the robots that do not learn (see Figure 5).

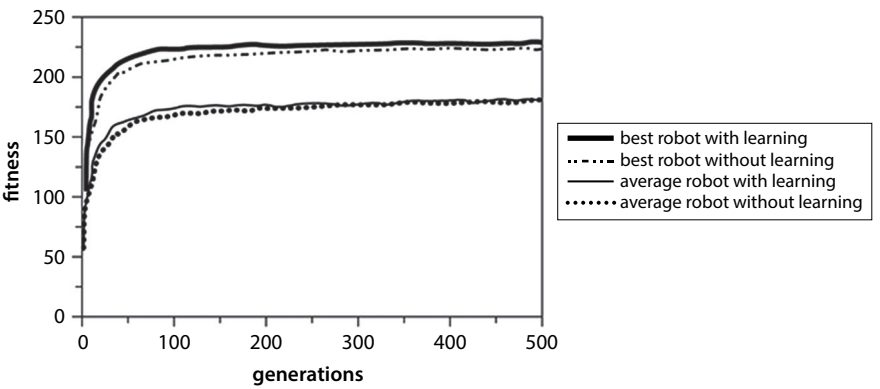


Figure 5. Best and average fitness of robots that learn and robots that do not learn.

One might think this as irrelevant because at the end the robots have the same fitness. But in biology time is important and the number of generations which are needed by a population of robots to evolve the capacity to eat the food tokens may be crucial. Imagine we add a "learning gene" to the robots' genotype which specifies the probability that the neural network of the robots has the learning neurons and, therefore, they can learn during their life. In the robots of the initial generation the value of this "learning gene" is randomly chosen in the interval between 0 and 1, and this means that, on average, 50% of the robots learn and 50% do not learn. But if for a certain number of generations the robots that learn have a higher fitness than the robots that

do not learn, during this period the robots with a higher value of their "learning gene" are more likely to reproduce than the robots with a lower value of their "learning gene". We have not constructed these robots but we predict that at the end all the robots of the population will have a high value for the "learning gene" and all the robots will learn. Evolution creates the capacity to learn.

The environment of these robots is very simple. It only contains food tokens and the only problem for the robots is to approach and eat the food tokens. What happens if we evolve the robots in an environment which is somewhat more complex and, therefore, less genetically predictable in the sense that the robots' genes cannot contain all the information which is necessary to behave appropriately in the environment?

We evolve the robots in two more complex environments. The first environment is a seasonal environment and the food tokens change their colour from one season to the next: in one season the food tokens are black and in the next season they are white, and so on cyclically (see Figure 6).

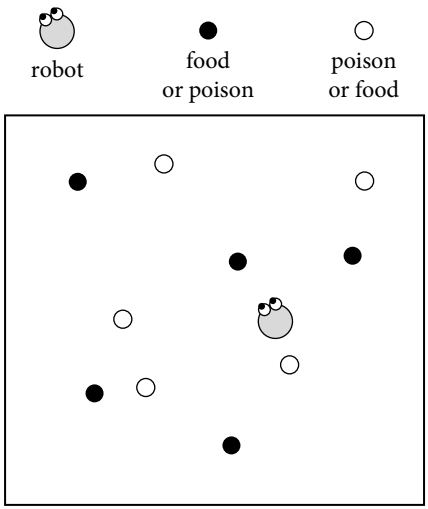


Figure 6. A seasonal environment in which in one season the food tokens are black and the poisonous tokens are white while, in the next season, the food tokens are white and the poisonous tokens are black, and so on cyclically.

The second environment contains three types of food tokens with three different colours – black, grey, and white – and during their life the robots go through a process of maturation/development which changes their sensory capacities (see Figure 7). (This process of development is hardwired by us.) When they are born, they can only see the black tokens. Then, after a certain number of cycles, they become able to see both the black tokens and the grey tokens, and at the end of their development, they can see all three types of food tokens: the black ones, the grey ones, and the white ones.

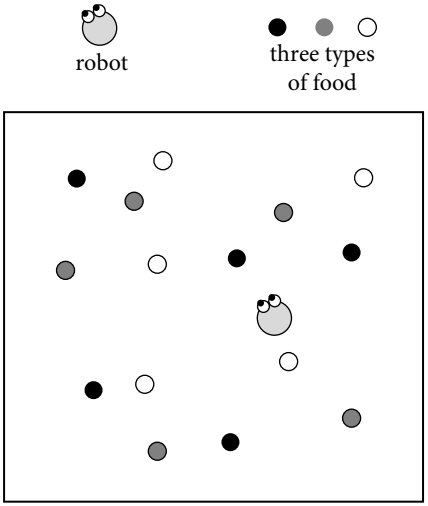


Figure 7. An environment with food tokens of three different colours: black, grey, and white. At birth, the robots can see only the back tokens. Then they develop and become able to see the grey tokens and, later in their life, also the white tokens.

When we compare the robots that learn and the robots that do not learn in these more complex environments, we find that the fitness of the robots that learn is slightly but constantly higher than the fitness of the robots that do not learn (see Figure 8).

This is interesting. These environments also appear to be genetically predictable and, therefore, evolution should be sufficient to create genes which are well adapted to the environment and learning should not be necessary. Instead, in these genetically predictable environments learning is slightly advantageous until the end. How can this be? A possible answer is that the environments are more complex because they are seasonal or because they contain different types of food tokens and the robots' visual capacities develop during their life, and this greater complexity makes them not completely predictable by the robots' genes.

The proof that the predictability of the environment is the critical variable is that, if we evolve the robots in an environment which *is* entirely predictable genetically, learning is not necessary and can even be costly. The new environment contains only black food tokens but these food tokens are not randomly distributed in the environment, as was the case for the preceding robots. Their position in the environment is entirely predictable because they form a regular square pattern (see Figure 9). Furthermore, the tokens do not re-appear when they are eaten, and in this way we exclude that the appearance of a food token in a new random location can make the environment different for different robots. Finally, at the beginning of its life, the robot is placed in the same position in the environment with the same orientation of its visual field. These features make the new environment entirely predictable from a genetic point of view because all the robots of all generations live exactly in the same environment.

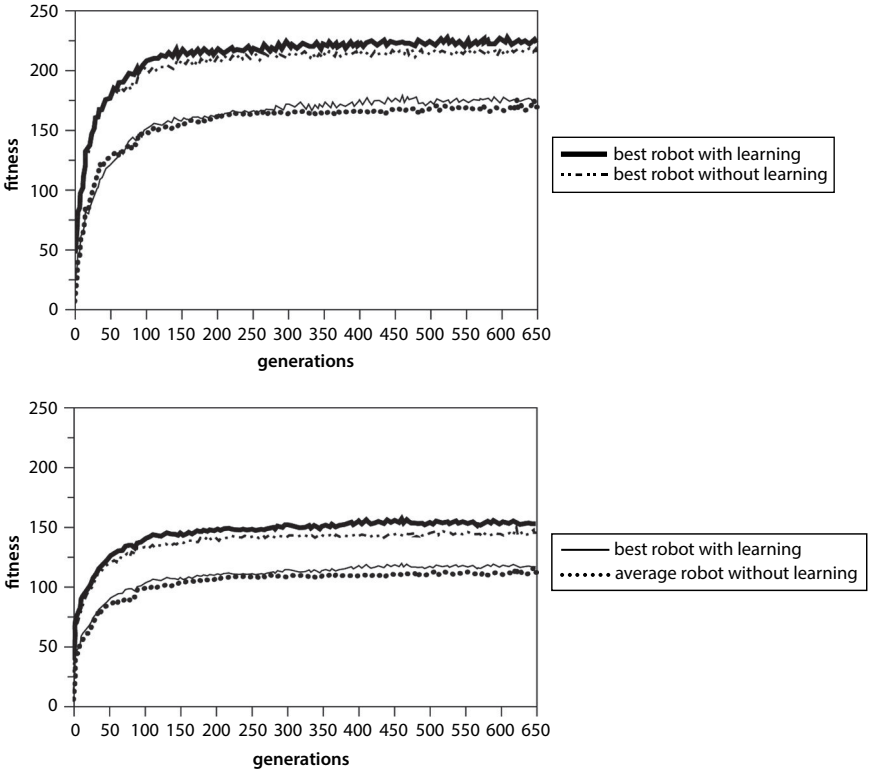


Figure 8. Fitness of robots that live in an environment in which in one season food is black and in the next season white (a) and of robots which develop during their life and become progressively able to see one, two, and three different types of food (b).

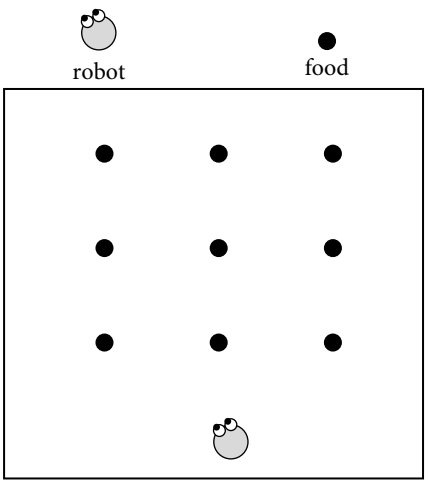


Figure 9. A completely predictable environment.

If the robots evolve in this environment which is completely predictable by their genes, the robots that learn have a lower fitness than the robots that do not learn. This is interesting. Learning is a useful adaptation in environments which are at least to some extent genetically unpredictable. If the environment is completely regular, it never changes, and it is the same for all robots, learning is a useless burden that has a negative impact on the robots' fitness. Consider that the neural network of the robots that learn is more complex than the neural network of the robots that do not learn because it has not only the regular internal neurons but also the learning neurons. Therefore, the genes of the robots that learn must encode more information than the genes of the robots that do not learn. If the robots live in an environment which is not entirely predictable by their genes, these costs are more than compensated by the advantages provided by learning but, if the robots live in an entirely genetically predictable environment, learning has only costs. But no animal lives in an environment which is entirely predictable by its genes, and this is why all animals learn.

There is another result which also is interesting and this result concerns the robots that develop during their life and become progressively able to see more colours and, therefore, more food types. At the end of evolution we test these robots in the controlled conditions of an artificial laboratory. The laboratory contains only one food token with a certain colour and we measure the time (number of input/output cycles of the robot's neural network) needed by a robot to reach the food token. A robot is put into the laboratory at birth, before it has had an opportunity to learn, and is tested for a certain number of trials with each of the three types of food tokens.

The results are shown in Figure 10 and they tell us that the robots that can learn do learn *during* the experiment. While for the robots that cannot learn the time necessary to reach the food token remains always the same, for the robots that learn the time progressively decreases during the course of the experiment.

But the interesting result is another one. As we have said, these robots see the black food tokens as soon as they are born but they become able to see the grey and white food tokens only later in their life. The results of the experiment show that the robots that do *not* learn take more time to reach the grey tokens than the black tokens and even more time to reach the white tokens (see Figure 11). This is explained by the fact that the robots have more occasions to respond to the food tokens which they can see during their entire course of their life (the black tokens) than to the food tokens that they can see only later in their life (the grey and the white tokens). In other words, the behaviour of these robots reflects their behavioural environment which changes during the robots' life as a result of their development. On the contrary, the robots that learn are more autonomous from their environment because they can adapt their behaviour to the environment during their life, with the result that they are equally able to reach all three types of food tokens independently of the age at which they become able to see them. This is interesting because it shows that learning reduces the influence of genes on behaviour.

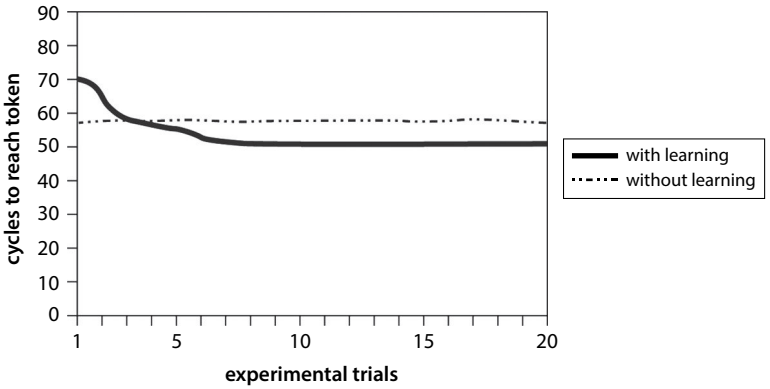


Figure 10. Number of cycles needed to reach a food token. At the beginning of the experiment the robots that learn take more time to reach the food token but then they rapidly learn and reach the food token faster than the robots that do not learn.

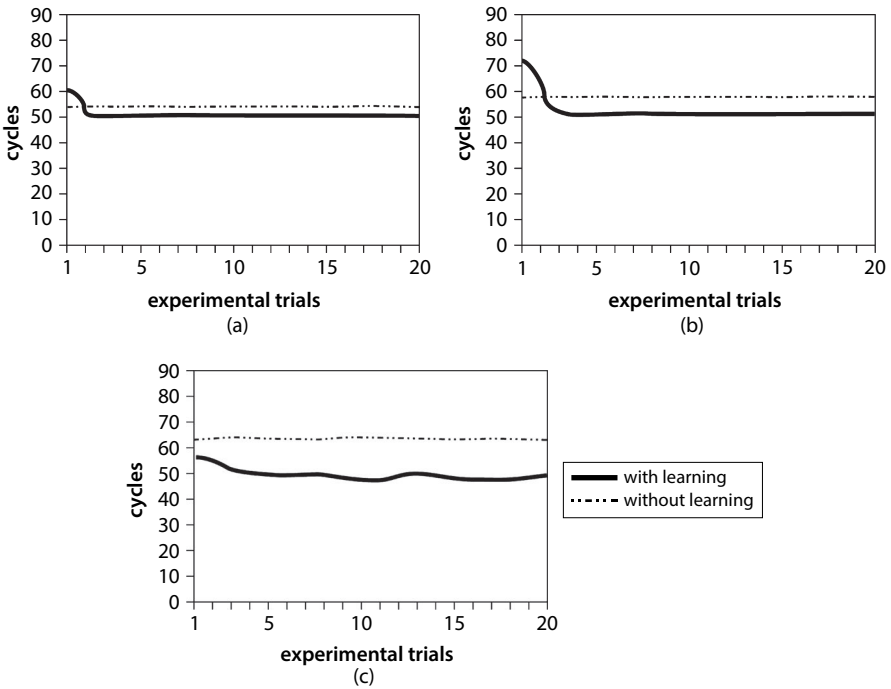


Figure 11. The robots that do not learn take more time to reach the food tokens they perceive later in their life. The robots that learn are equally able to reach the black tokens (a), the grey tokens (b), and the white tokens (c).

5. Living in genetically unpredictable environments

The robots of the preceding Section tell us that, even if the environment in which the robots live is largely genetically predictable, learning leads to higher levels of fitness. But learning is even more important if the robots live in an environment which is genetically unpredictable so that the robots' genes cannot in any way prepare the robots for the environment in which they will spend their life.

We evolve the robots in two environments which are both genetically unpredictable. The first environment is divided into four zones, each zone contains tokens of a different colour but only the tokens of one zone are food and have a sweet taste. The tokens of the other three zones contain no energy and have no taste, and eating them does not make any difference for the robots. The problem is that the colour of the tokens which have energetic value changes unpredictably from one generation to the next so that, when a robot is born, it does not know which tokens are food and which are not food (see Figure 12).

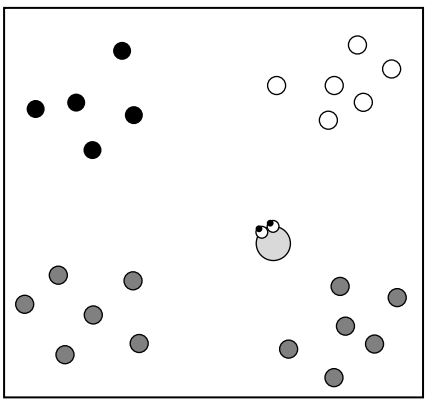


Figure 12. The environment contains tokens of four different colours located in four different zones of the environment. Only the tokens of one colour are food and the colour of the food tokens varies unpredictably from one generation to the next.

The second genetically unpredictable environment contains two types of tokens with two different colours, black and white, and the tokens are distributed in the entire environment. The tokens of one colour are food and they have a sweet taste. The tokens of the other colour are poison and they have a bitter taste. The problem is that the tokens which are food and the tokens which are poison change unpredictably from one generation to the next (see Figure 13).

In these environments the capacity to learn is not only useful but necessary. The difference in fitness between the robots that learn and the robots which do not learn is much larger than for the robots which live in the environments which are genetically predictable. In the environment in which there are four different types of tokens

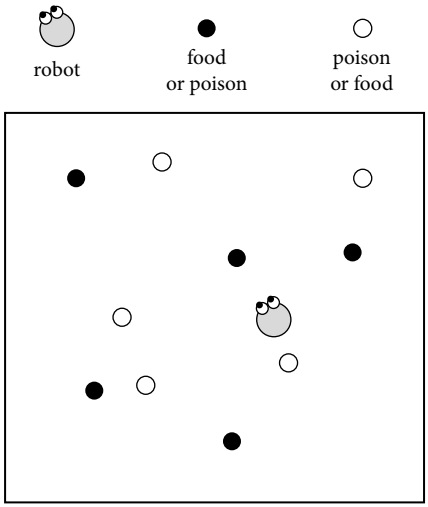


Figure 13. The environment contains two types of tokens which are randomly distributed in the entire environment. One type of token is food and the other type is poison but the colour of the food and poisonous tokens varies unpredictably from one generation to the next.

in different zones but which are the tokens which contain energy cannot be predicted by the genes, the only strategy available to the robots that do not learn is eating any type of tokens, both those that contain energy and those that do not contain energy. Clearly, this is a very costly strategy and it explains why the robots that do not learn have a very low level of fitness. On the contrary, the robots that learn eat the different tokens and learn which one is food based on its sweet taste. They eat only these tokens and their fitness is quite high (see Figure 14).

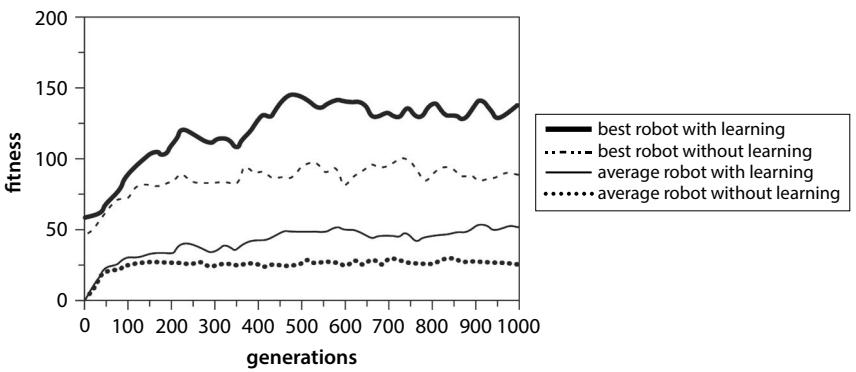


Figure 14. The robots live in an environment divided up into four zones each containing one type of token. Only the tokens of one type are food but the colour of the food tokens varies from one generation to the next.

The other robots live in an environment which is even more problematic because some of the tokens are food and some are poison but the colour of the two types of tokens varies randomly from one generation to the next. If the robots cannot learn, they have no choice but eating both the tokens containing energy and the tokens containing poison – which is not only an expensive but also a disastrous behavioural strategy. The robots that learn, on the contrary, can learn which tokens are food and which are poison on the basis of their taste and, in this way, they can easily adapt to the environment in which they happen to be born (see Figure 15).

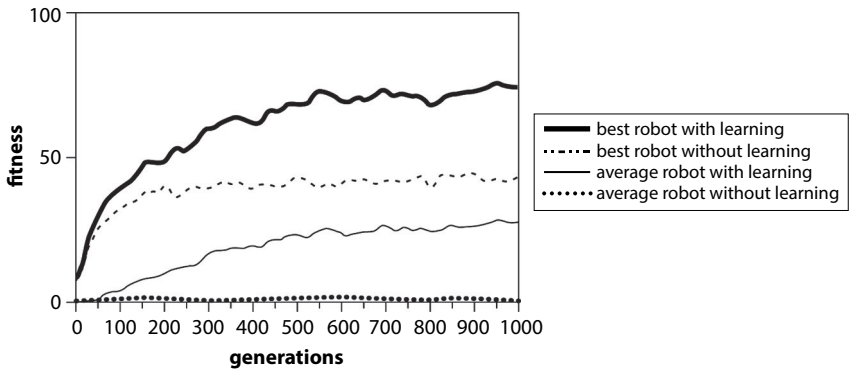


Figure 15. Fitness of robots that live in an environment containing both edible and poisonous tokens. The colour of the two types of tokens varies unpredictably from one generation to the next.

These last robots are a clear proof of what we have said at the beginning of this Chapter. The robots learn the colour of the tokens that they must eat and the colour of the tokens that they must not eat. Food tokens have a taste of sweetness while poisonous tokens have a taste of bitterness but sweetness and bitterness are purely "objective" (physical or chemical) properties and it is only evolution which gives reward value to sweetness and punishment value to bitterness. In fact, for some animals bitterness might signal food and sweetness poison. We might make the food tokens taste bitter and the poisonous tokens taste sweet, and nothing would change. This is a very important way in which evolution influences learning. Learning is based on rewards and punishments but it is evolution which gives reward and punishment value to stimuli.

6. Learning in the experimental laboratory

One fundamental principle of robotics as science is the principle "one robot/many phenomena": a robot which reproduces many different phenomenon concerning the behaviour of animals is to be preferred to a robot which reproduces only one single

phenomenon. This is why robots that *both* evolve *and* learn are better than robots that *either* evolve *or* learn. In this section we continue to apply this principle and we bring our robots into an experimental laboratory with the aim to replicate some of the results obtained in experiments on what psychologists call instrumental learning and classical conditioning. Instrumental learning is learning to do something which allows the animal to obtain some desired result. A rat learns to press a lever because pressing the lever is followed by the arrival of food. Classical conditioning is learning to respond to a new stimulus in the same way as the animal already responds to another stimulus if the first stimulus is experienced together with the second stimulus. A dog learns to salivate when it hears the sound of a bell if hearing the sound of the bell is followed by the arrival of food in the dog's mouth which causes the robot to salivate.

The robots on which we do these experiments are our last robots, those that evolve in the environment that contains both food tokens and poisonous tokens but the colour of the two types of tokens – black and white – changes unpredictably from one generation to the next. As we have seen, possessing a capacity to learn is very important for these robots. The robots evolve in an environment in which the tokens of one colour can be either food or poison and, therefore, to survive and reproduce the robots must learn if a token is food or poison based on its taste (sweet or bitter).

We put the robots, one at a time, into an experimental laboratory which contains a *grey* token which the robot has never seen during its life and, therefore, at the beginning of the experiment the robot is not attracted or repelled by the grey token and it move around in the laboratory more or less randomly. However, if the robot happens to touch the grey token, a black token immediately appears in front of the robot. The robots approaches and eats the black token and what happens next depends on whether the black token tastes sweet or bitter. If the black token tastes sweet (reward), the probability that the robot approaches and touches the grey token increases in the successive trials of the experiment (see Figure 16). If the black token tastes bitter (punishment), the probability that the robot approaches and touches the grey token tends to decrease and the robot avoids the grey token (see Figure 16). We have also run a control experiment in which touching the grey token is *not* followed by the appearance of the black token. In these conditions there is no learning of any type: the number of times the robot touches the grey token does not increase or decrease in the successive trials of the experiment.

Our robots are neuro-robots and, therefore, we can look at what happens inside the robot's neural network while the robot is learning to approach or avoid the grey token. We describe the results obtained when the black token tastes sweet (reward) and, therefore, the robot learns to approach the grey token, but similar results are obtained when the black token tastes bitter (punishment) and the robot learns to avoid the grey token. As we have said, the learning neurons have an activation threshold and, only if the threshold is reached, the learning neurons are activated and they influences how the robot responds to the current sensory input. At the beginning of the experiment

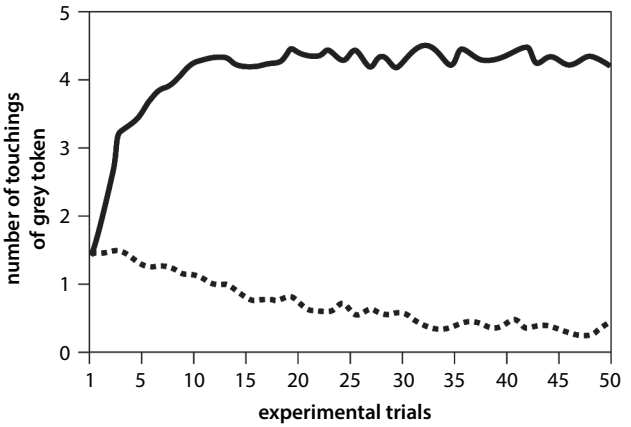


Figure 16. Instrumental learning. Number of times a robot touches a grey token it has never seen before when the token which appears after the robot has touched the grey token has a sweet taste (reward) (higher curve) or a bitter taste (punishment) (lower curve).

the grey token is unable to activate the learning neurons, which are activated only by the sweet-tasting black token. But the grey token leaves a trace on the learning neurons' receptive sites and, given this trace, the appearance of the sweet-tasting black token gradually increases the probability that the grey token activates the learning neurons. In fact, in the course of the experiment, the learning neurons tend to be progressively activated *before* the appearance of the black token and, at the end, they are activated as soon as the *grey* token appears. Since the robot responds to the activation of the learning neurons by approaching the token which has activated these neurons, after a certain number of experimental trials the robot responds to the sight of the grey token by approaching and touching the grey token (see Figure 17).

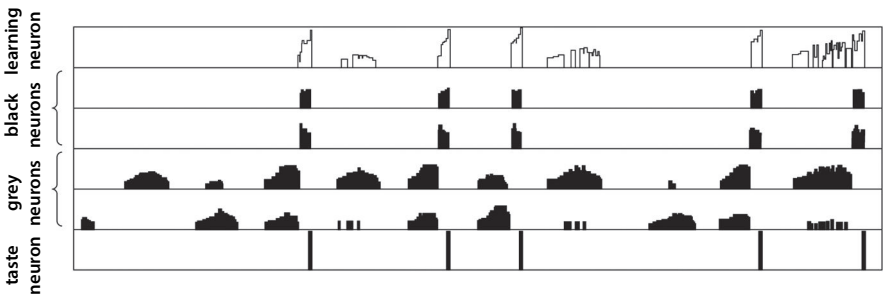


Figure 17. At the beginning of the experiment, a learning neuron is activated only by the sight of the black token. Later on, if seeing the grey token is followed by seeing the black token, the activation of the learning neuron occurs as soon as the robot sees the grey token.

The robots replicate other phenomena observed in laboratory experiments on real animals. One is called extinction: the progressive disappearance of a response if the response ceases to be followed by a rewarding stimulus. If, after the robot has learned to approach and touch the grey token, touching the grey token is no more followed by the appearance of the sweet-tasting black token, the number of times the robot touches the grey token progressively decrease. The behaviour of reaching the grey token has been extinguished (see Figure 18).

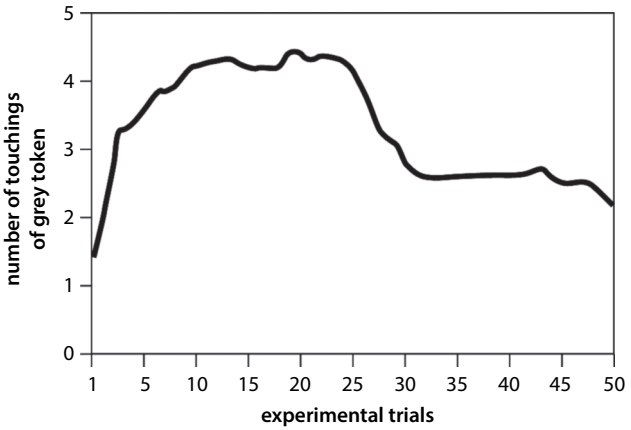


Figure 18. Extinction. After the robot has learned to reach the grey token, if touching the grey token is no more followed by the appearance of the sweet-tasting black token (trial 25), the behaviour of reaching the grey token becomes progressively less probable.

Another result concerns the timing of the events. An animal learns only if the temporal interval between its response and the appearance of the reward is not too long. If we vary the length of the temporal interval between the moment in which the robot touches the grey token and the appearance of the sweet-tasting black token, we find that, as the interval increases, learning is progressively reduced. Figure 19 shows the learning curves (number of grey tokens reached and touched by the robot) for five temporal intervals between the appearance of the grey token (conditioned stimulus) and the appearance of the black token (unconditioned stimulus): 1 cycle, 10 cycles, 20 cycles, 40 cycles, and 60 cycles. There is progressively less learning as the temporal interval increases because for the robot's brain it is increasingly difficult to link the action of touching the grey token with the appearance of the sweet-tasting black token (reward). This happens because, in our robots, the markers left on the synaptic sites of the learning neurons tend to disappear with the passage of time.

Another result of laboratory experiments on real animals is that an animal learns only if it is motivated to learn. If it is not hungry, a rat will not learn to push the lever

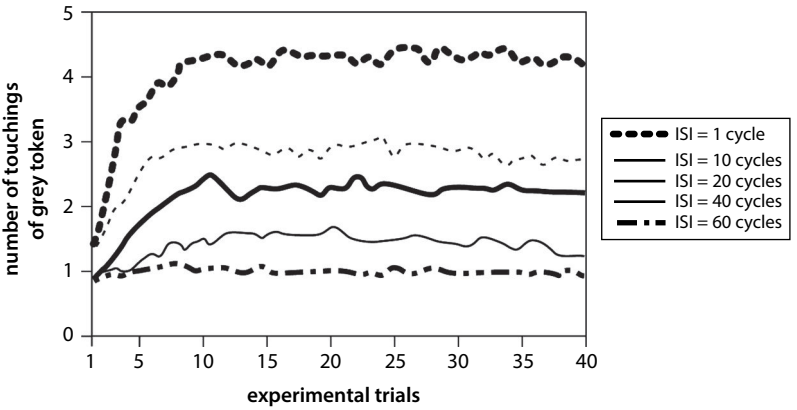


Figure 19. Learning curves for five inter-stimulus intervals (ISI) between touching the grey token and the appearance of the sweet-tasting black token.

to obtain food. To replicate this phenomenon we evolve a new type of robots which are identical to the preceding robots except that they have a bodily store of energy whose level cannot go to zero if the robot must remain alive and leave offspring. The level of energy is increased by some quantity when the robot eats a food token and it decreases by the same quantity when the robot eats a poisonous token. The robot's neural network has a hunger sensor and the activation level of the hunger sensor reflects the quantity of energy currently contained in the robot's body. These robots generate one offspring at regular intervals and, therefore, for these robots fitness is not the number of food tokens eaten minus the number of poisonous tokens eaten during their life but their ability not to die for lack of energy. In all other respects the robots are like the preceding robots.

We do with these robots the same instrumental learning experiment with the grey token and the results are that learning only takes place if, at the time of the experiment, the robots are hungry. If the robots are not hungry, they do not learn to approach and touch the grey token. Figure 20 shows that learning depends on how hungry is the robot at the beginning of the experiment.

Like real animals, these robots learn if they are hungry and they do not learn if they are not hungry. But what happens if, after a robot has learned to touch the grey token because it was hungry, we test the robot when it is not hungry? The results are that the robot ignores the grey token. If the robot is not hungry, the grey token loses value for the robot. (This phenomenon is called devaluation by psychologists.) Figure 21 shows that, after the robot has learned to approach and touch the grey token in the first 40 trials of the experiment, the number of times the robot touches the grey token rapidly decreases as a function of the degree of satiation of the robot – which is manipulated by us – and it goes to zero if the robot is completely satiated.

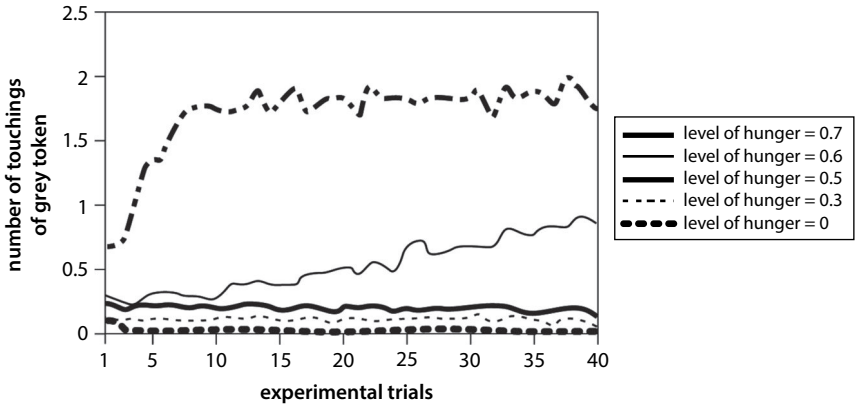


Figure 20. Learning curves for five different levels of hunger at the beginning of the experiment. The robots which are hungrier learn more than the robots which are less hungry and the robots with zero hunger simply do not learn.

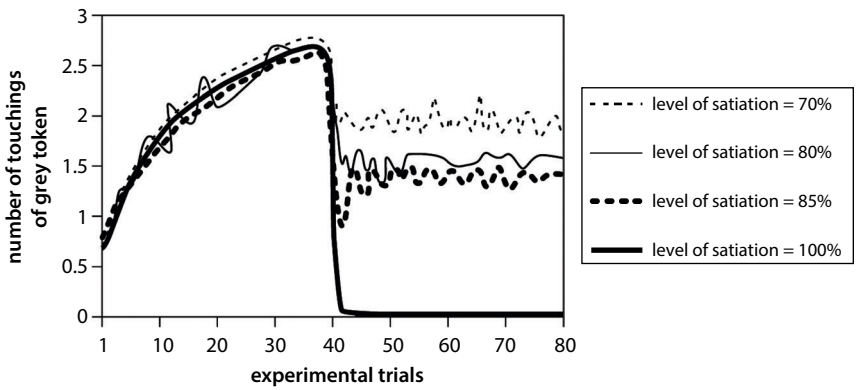


Figure 21. After the robots have learned to reach and touch the grey token, we manipulate their level of satiation (trial 40) and we find that the number of times that the robots reach and touch the grey token depends on their satiation level.

These robots replicate various results obtained in experiments on instrumental learning. The robots learn to do X if doing X is followed by something that evolution has made desirable – and, in fact, our robotic experiments resemble real experiments on rats that learn to press a lever to obtain food. But our robots can also replicate other experimental results, those on classical or Pavlovian learning: a dog learns to salivate in response to the sound of a bell if the sound of the bell is followed by the arrival of food in the dog's mouth.

Our experiment is the following. The robot sees a grey token which it has never seen before and, after a certain interval, a sweet-tasting black token makes its appearance. While in the preceding experiments the black token only appeared if the robot touched the grey token, in this experiment the black token automatically appears a certain number of cycles after the appearance of the grey token – and this is what distinguishes instrumental learning from classical conditioning. In this experiment the robot cannot move and we do not actually reproduce the behaviour of salivating but we assume that the robot salivates when the learning neurons of the robot's neural network are activated. A grey token (conditioned stimulus) appears, after a short time interval it disappears and, after another short interval, a black token (unconditioned stimulus) makes its appearance. What happens in the robot's brain? Initially, the learning neurons are activated (and, automatically, salivation occurs) only when the black token appears but, after a certain number of trials, the learning neurons tend to be activated as soon as the grey token appears. If we vary the interval between the disappearance of the grey token and the appearance of the black token (the inter-stimulus interval), we find that learning is faster when the inter-stimulus interval is small and it becomes progressively slower with increasing inter-stimulus intervals until, with very large inter-stimulus interval, there is no learning (see Figure 22). This is another result which is obtained in real experiments on real animals.

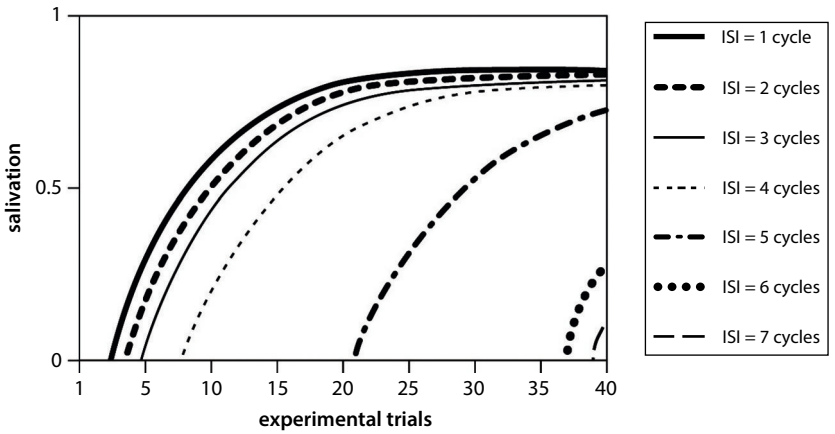


Figure 22. Classical conditioning. Increase in the probability of responding to the conditioned stimulus (salivating) as a function of the temporal interval (ISI) between the conditioned stimulus and the unconditioned stimulus.

7. Imprinting and learning from one's mother

The robots described in the preceding section replicate various experimental results on real animals but learning can take other forms and, given our principle "one robot/many phenomena", we must apply our model of learning to these other forms of learning. This is what we do in this section. But this section also has the goal to explore other ways in which evolution and learning interact together. Our robots reproduce three different ways in which evolution and learning interact together. First, evolution explains why animals learn: learning allows an individual to adapt to an environment which cannot be predicted by its genes. Second, evolution gives reward or punishment value to stimuli, and the reward or punishment value of stimuli is crucial for learning. Third, evolution does not only create the starting point for learning but it continues to play a role in learning because the inherited properties of the brain – the weights of the connections of the robots' neural network – interact with what changes in the brain during learning – the learning neurons. But evolution and learning interact together in other ways and two of these other ways are learning by getting "imprinted" with another specific individual – usually one's mother – and following another individual in order to learn from the other individual. In this Section we describe robots that get imprinted with their mother and robots that follow their mother to learn from her.

Some animals become imprinted with the first moving "object" they perceive after birth and human beings treat this "object" differently from all other objects during their entire life. Konrad Lorenz' gees followed him because he was the first moving object that they had seen after they were born. Human infants become attached to the first person who takes care of them, usually their mother. Imprinting is a form of learning which illustrates another way in which evolution and learning interact together: the tendency to become attached to one's mother is encoded in the genes but infants must learn *who* their mother is. Can we reproduce imprinting with our robots?

A newborn robot lives in an environment that contains two adult robots. One of the two adult robots is its mother and, to increase its fitness, the newborn robot must approach and remain in the vicinity of its mother – clearly, because its mother helps the newborn robot to remain alive. The mother and the other adult robot are not really robots but they are two objects of different colours – black and grey. When the newborn robot touches an adult robot, either its mother or the other adult robot, the adult robot disappears and it reappears in another part of the environment. But, to remain alive, the newborn robot must approach and reach its mother, not the other adult robot. The problem is that for each newborn robot the colour of the adult robot that the newborn robot must approach and reach – its mother – can be black or grey, randomly chosen. So, when it is born, a robot does not know if it must approach the black adult robot or the grey adult robot. At birth, the robot is placed near to its mother so that the robot can learn the colour of its mother and, during its life, it can approach and reach its mother and avoid the other adult robot (see Figure 23).

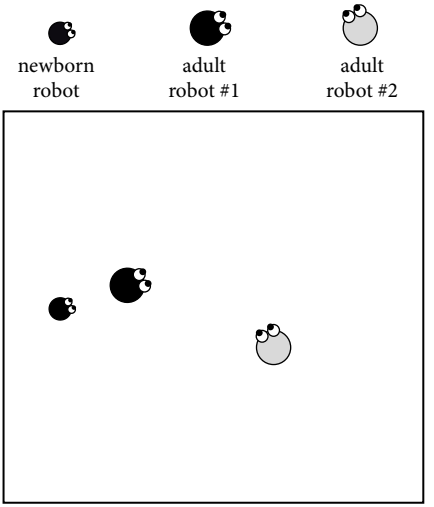


Figure 23. A robot is born in an environment with two adult robots with different colours, black and grey. The newborn must learn to approach the adult robot which it sees at birth (its mother) – in this case the black robot – and to avoid the other adult robot.

We compare two different populations of robots. In one population the infant robots can move as soon as they are born. In the other population the infant robots are unable to move for 100 cycles after birth and during these 100 cycles an infant robot continues to see the same adult robot – its mother. Figure 24 shows the result of this comparison.

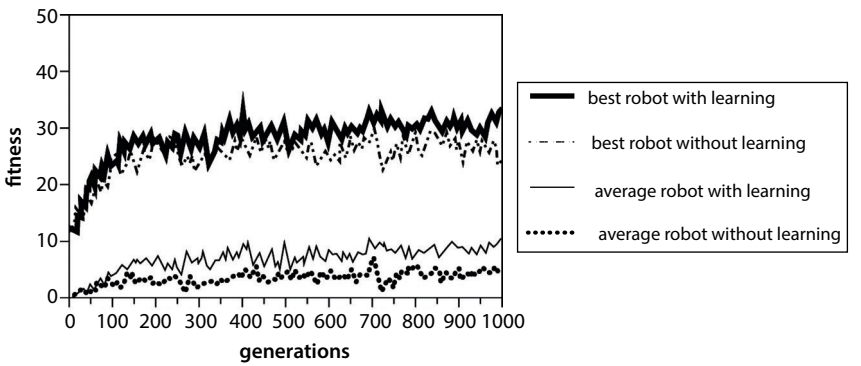


Figure 24. Fitness of robots that learn the colour of the adult robot they see when they are born (their mother) because they canoot move for 100 cycles after birth and of robots that do not learn because they are able to move as soon as they are born.

The robots that at birth cannot move so that they have 100 cycles for learning the colour of their mother have more fitness than the robots that can move as soon as they are born and, therefore, do not have the opportunity to learn the colour of their mother.

The newborn robots of both populations could learn which adult robot is their mother because they have a neural network with learning neurons. But a newborn robot needs 100 cycles in which it constantly sees its mother in order to get imprinted with her. The robots that can move as soon as they are born do not have this opportunity to get imprinted with their mother and, therefore, they approach indifferently their mother and the other adult robot, with a significant reduction of their fitness. The importance of imprinting is shown by the following fact. If we pick up a newborn robot from the population in which the newborn robots have 100 cycles to get imprinted with their mother, and we lesion (eliminate) the robot's learning neurons as soon as the robot is born, the newborn robot is unable to learn what is the adult robot which it must approach during its life – its mother – and, therefore, its fitness is drastically reduced.

For these robots we decide that they cannot move for 100 cycles after their birth so that they can get imprinted with their mother. But this characteristic of the robots should spontaneously evolve in the robots. We predict that, if we add to the robots' genotype a gene which specifies the age in which an infant robot can move in the environment, the necessity for newborns to get imprinted with their mother becomes a selective pressure for evolving an initial period of life in which newborns cannot move. We have not constructed these robots but the relation between the age in which a newborn robot can move in the environment and the necessity to get imprinted with its mother is another interesting type of interaction between evolution and learning.

The newborns of many animals get imprinted with their mother in order to stay near to her and to be taken care of by her. But human newborns, when they become able to move, follow their mother not only to be taken care of by her but also to learn from her. This is shown by our next robots.

The new robots live in an environment which contains only two tokens, a food token and a poisonous token. Both the food token and the poisonous token disappear when a robot touches (eats) them and another token of the same type appears in another position in the environment. The problem is that the colour of the food tokens and the colour of the poisonous tokens change unpredictably from one generation to the next and, therefore, a newborn robot must learn the colour of the food tokens and the colour of the poisonous tokens in the environment in which it happens to be born. The environment of a newborn robot also contains the robot's mother, and the robot's mother already knows which tokens are food and which are poison and, therefore, she approaches and eats the food tokens and avoids the poisonous tokens – whatever their colour. The robot's mother is not really a robot and it is hardwired by us to do what she does but she might also have learned from *her* mother which tokens are food and which poison.

What must a newborn robot do to remain alive? The newborn robot must follow its mother in order to learn what is the colour of the tokens that she eats (the food tokens) and what is the colour of the tokens that she avoids (the poisonous tokens) (see Figure 25).

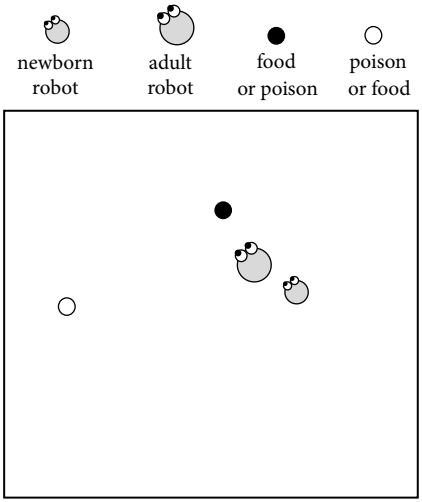


Figure 25. The environment contains one food token and one poisonous token whose colour varies from one generation to the next. The newborn robots evolve the behaviour of following their mother in order to learn from her which is food and which is poison.

Unlike our preceding robots, these robots do not have the taste sensors and, therefore, they cannot learn to associate one colour with food (which tastes sweet) and the other colour with poison (which tastes bitter). The only way for them to distinguish between the food tokens and the poisonous tokens is to follow their mother in order to learn from her to eat the tokens that she eats and to avoid the tokens that she avoids (see Figure 26).

This is especially important because a robot's life lasts 8000 cycles but the robot's mother dies after 2000 cycles. Therefore, when the robots are born, they must follow their mother so as to learn from her what they must do when she dies. If we examine the robots of the last generation, we see that, when they become adult and their mother has died, the robots eat the tokens that their mother ate and to avoid the tokens that their mother avoided (see Figure 27).

This is another way in which evolution and learning interact together. The robots *evolve* the behaviour of following their mother so that they can *learn* from her the colour of the food tokens and the colour of the poisonous tokens. These robots learn from their mother which tokens are food and which are poison but they already know that they must approach the food tokens and avoid the poisonous token because these behaviours are evolved and, therefore, innate. In Chapter 8 we describe other robots whose behaviour is entirely learned by imitating other robots.

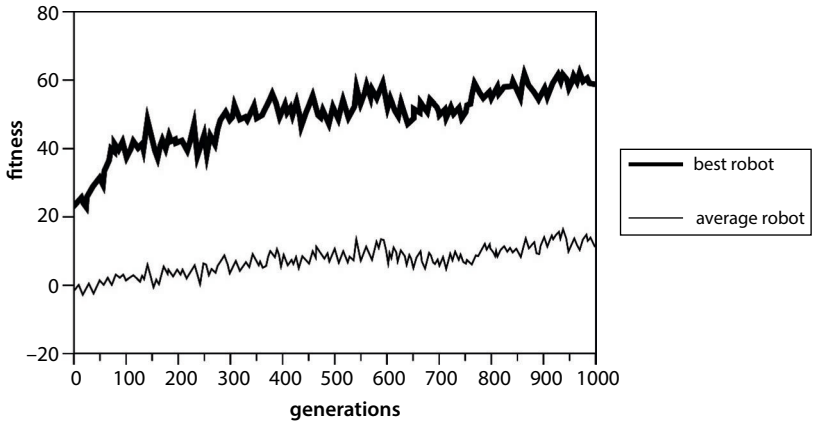


Figure 26. Fitness growth for robots that evolve the behaviour of following their mother so as to learn from her to eat the tokens which she eats and to avoid the poisonous tokens that she avoids.

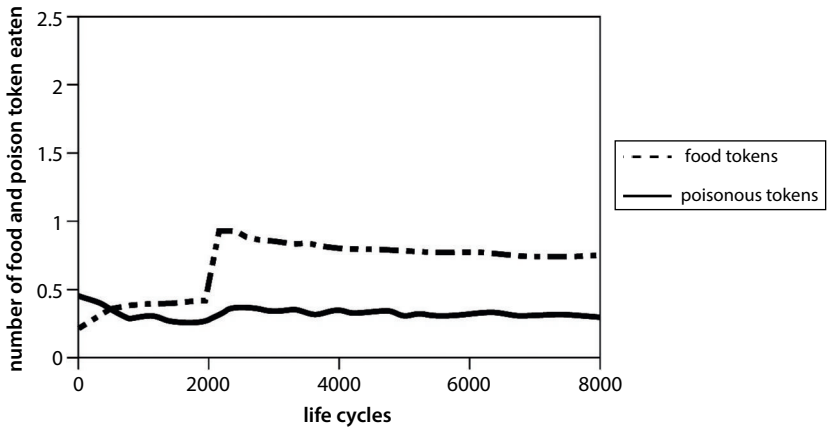


Figure 27. Food and poisonous tokens eaten by the robots of the last generation. In the first 2000 cycles of life the follow their mother so as to learn from her what is the colour of the food tokens and of the poison tokens and then they use this knowledge when she dies.

8. The influence of learning on learning

Learning necessarily begins with what is encoded in the genes and, therefore, evolution necessarily influences learning. But learning is also influenced by what an animal has already learned. Animals learn many different things during their life and what they have already learned inevitably creates the initial conditions for learning new behaviours. Psychologists call this phenomenon transfer of learning. The influence can be positive – learning one behaviour facilitates the learning of a second behaviour – or it can be negative – learning one behaviour makes it more difficult to learn a second behaviour. Today's robots start their learning from zero not only because evolution does not create the initial conditions for their learning and does not help them to learn, but also because they usually learn only one single behaviour and, therefore, it is impossible to study how learning one behaviour influences learning another behaviour.

Our next robots learn two different behaviours, one after the other, and they reproduce both how learning one behaviour may facilitate the learning of a second behaviour and how learning the second behaviour does not destroy the knowledge of the first behaviour. The robots are those that evolve in the environment in which the colour of the food tokens and the colour of the poisonous tokens change from one generation to the next and, therefore, the robots must learn which tokens are food and which are poison in the particular environment in which they happen to be born. At the end of evolution we test the robots in an experimental laboratory in which for a certain number of cycles the food tokens are black and the poisonous tokens are white and then the situation is reversed and the food tokens are white and the poisonous tokens black. This is like what happens in the robots' natural environment but, while in the natural environment the food tokens and the poisonous tokens change their colour from one generation to the next, in the laboratory the tokens change their colour in the course the experiment. Therefore, a robot must first learn to eat the tokens of one colour and avoid the tokens of the other colour and then the same robot must learn to do the opposite because the colour of the tokens has changed. The robots begin by learning that the food tokens are black and the poisonous tokens are white (one behaviour) but then the situation changes and the robots must learn that the black tokens are poison and the black tokens are food (another behaviour). The results are shown in Figure 28. The robots begin their learning of the second behaviour from a higher point compared to their initial learning of the first behaviour. This is positive transfer of learning. Learning one behaviour facilitates the learning of a second behaviour.

But what happens if the situation changes again and the food tokens return to be black and the poisonous tokens white? When they learn two successive behaviours, today's robots exhibit what is called catastrophic forgetting: learning the second behaviour destroys their knowledge of the first behaviour. This is due to the fact that today's robots learn because the connection weights of their neural network change and, since what they have learned is encoded in their connection weights, learning a second task

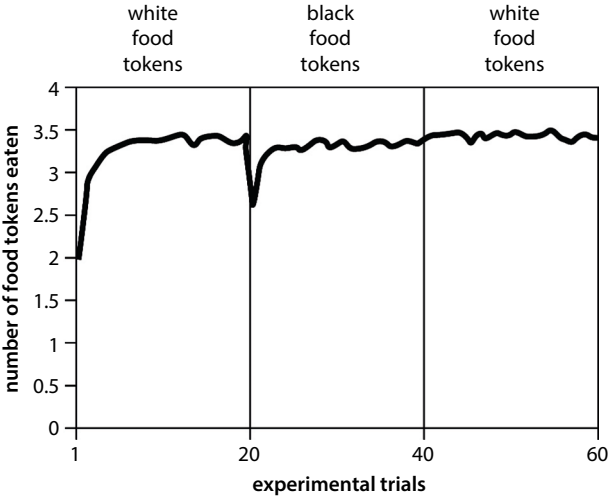


Figure 28. Learning one behaviour helps the robots to learn a second behaviour (cycle 20) and learning the second behaviour does not destroy their knowledge of the first behaviour (cycle 40).

changes these weights so that their knowledge of the first task is destroyed. This is not true for real animals. When animals learn one behaviour and then they learn a second behaviour, their knowledge of the first behaviour remains more or less intact. Does catastrophic forgetting occurs in our robots that learn by changing not the connection weights of their neural network but the receptivity of their learning neurons? The answer is No. As Figure 28 also shows, when the colour of the food tokens returns to be white, the robots do not have forgotten what they have learned at the beginning of the experiment and there is no catastrophic forgetting – and this is an argument in favour of our model of learning.

9. Learning with an evolved neural architecture

The neural network of the robots described in this Chapter and in the rest of the book is made of a certain number of neurons connected together in certain ways, and this neural architecture is decided by us, is the same for all the robots of a given population, and it never changes. But the architecture of the brain evolves like the rest of the body and evolution finds the best architecture for the brain of each species of animals. The robots of this section have a neural network whose architecture is not decided by us, is different from individual to individual, and it evolves in a succession of generations.

The robots are not our usual robots. They don't have a body and they do not move but they have a retina and, in each input/output cycle, they see an object in one specific part of their retina. The robots must learn to recognize *what* is the object and *where* the object is in their retina. The robots have visual neurons which encode the entire content of the retina and two separate sets of output neurons – the "what" neurons and the "where "neurons. The "what" neurons recognize the *shape* of the object and the "where" neurons the *location* of the object in the retina. For example, if a T-shaped object appears in the lower left part of the retina, the "what" neurons must recognize that it is a T-shaped object and the "where" neurons that the object is in the lower left part of the retina (see Figure 29). The output neurons of these robots are not motor neurons because recognizing what is the shape of an object and where is the object in the retina is only a preliminary step to decide what to do with respect to the object. But we ignore what the robots will do with respect to the object and the neural network is only one part of their total neural network.

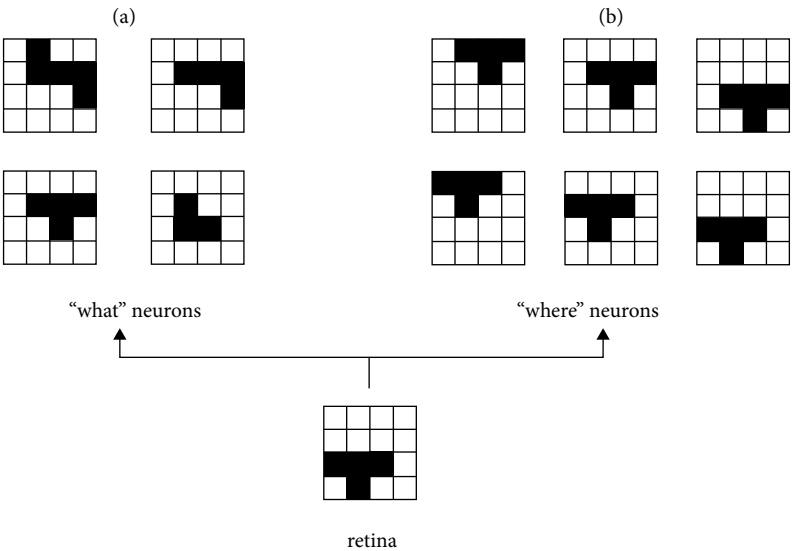


Figure 29. In each cycle the robot sees one of four objects (a) located in six different positions in its retina (b). The robot is currently seeing a T-shaped object in the left lower part of its retina and it must recognize the shape of the object and its position in the retina.

To make the robots learn we do not use our model of learning but we use another model of learning: the backpropagation learning algorithm. (We explain this choice in the last section of this chapter.) When a robot is born, the connection weights of its neural network are random and, therefore, the robot is unable to recognize what is the shape of the object and where the object is located in its retina. But then the robot

learns. In each cycle the robot sees an object and its neural network responds by "saying" what is the shape of the object and where is the object in its retina. The robots learn because we "tell" them how they should respond. The robots compare their response with this "teaching input" provided by us and, as a result of this comparison, the connection weights of their neural network gradually change until the robots become able to respond correctly.

But what interests us is not how the robots learn but what is the architecture of their neural network. The neural architecture of our robots is not decided by us but it encoded in their genes and it evolves in a succession of generations. All robots have the same number of visual neurons which encode the content of their retina, the same number of internal neurons, and the same number of output neurons which "say" what is the object and where is the object in the retina. And in all robots the visual neurons send their connections to all the internal neurons (Figure 30a).

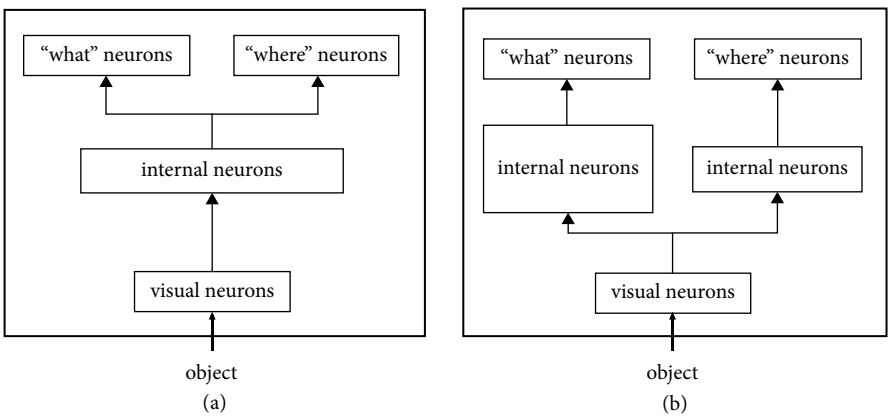


Figure 30. The robots begin with a non-modular neural network (a) but, at the end of evolution, their neural network is made of two modules, with more neurons dedicated to recognizing what object they see and fewer neurons to where is the object in their retina (b).

This does not change in the course of evolution. What changes during evolution is how the internal neurons are connected to the "what" and "where" neurons. In the robots of the first generation, the connections can be more or less numerous and they are random. In one robot an internal neuron can be connected to a "where" neuron or to a "what" neuron or to both neurons and, in another robot, the connections can be different. The robots that reproduce are those which are better at recognizing the shape of the objects and their location in the retina. The offspring inherit the neural architecture of their parents with the addition of random variations that can add or cancel one or more of the connections that link the internal neurons to the "what" and "where" neurons.

At the end of evolution we find that the robots have a neural network which is made up of two separate modules. Some internal neurons send their connections only to the "what" neurons and the other internal neurons send their connections only to the "where" neurons (see Figure 30b). One module is dedicated to recognizing what is the object and the other module is dedicated to recognizing where is the object in the retina. Why this modular architecture? Knowing what is an object and knowing where the object is located in one's retina are two different capacities and the robots' neural networks keep these two capacities separate by dedicating one module to one capacity and another module to the other capacity. A non-modular architecture in which all the internal neurons are connected to all the output neurons would mix together the two capacities and the two capacities would interfere with each other. These robots have what neuroscientists call the "what" neural pathway and the "where" neural pathway.

But, as Figure 30b shows, there is another interesting result: the "what" module is larger than the "where" module. More internal neurons are dedicated to recognizing what is an object and fewer internal neurons to recognizing where is the object in the retina. Recognizing *what* is an object is a more complex task than recognizing *where* the object is located in one's retina.

What these robots tell us about the relation between evolution and learning? The architecture of the robots' neural network is evolved but the robots learn during their life what is an object and where is the object in their retina. What happens if we completely eliminate learning and we let evolution evolve both the architecture of the neural network and its connection weights? The results are not very good. The genes of the robots encode both the architecture and the connection weights of their neural network, and the connection weights do not change during a robot's life because the robots do not learn. This is a problem because, if some random mutation changes the architecture of an offspring robot, the mutated architecture may need connection weights which are different from the connection weights that the robot inherits from its parent, and the robot may be unable to do the "where" and "what" tasks. With learning the problem is solved. The offspring robots inherit the neural architecture of their parent but it is learning which finds the appropriate connection weights for their neural architecture. Evolution and learning are both important and they must collaborate together.

10. On the limits of the robots described in this Chapter

This Chapter has been dedicated to how behaviours are acquired. Behaviours can be acquired in a succession of generations – evolution – and they can be acquired during an individual's life as a result of the individual's specific experiences – learning. In constructing our robots we have applied the principle "one robot/many phenomena" in many ways. The same robots both evolve and learn. The robot learn both in their

natural environment and in the experimental laboratory. We have used the same model of learning to reproduce different types of learning – instrumental learning, Pavlovian conditioning, imprinting, and learning from one's mother – and different factors that influence learning – the timing of events and the individual's motivations. And we have reproduced the positive effects of learning one behaviour on the successive learning of another behaviour and how learning another behaviour after having learned one behaviour does not cancel the knowledge of the first behaviour.

But our robots have many limitations and many questions remain open, and in this last section we list these limitations and these questions.

One first question is if the model of learning described in Section 2 can reproduce all forms of learning. Is all learning based on the same basic mechanism? The robots that learn to recognize where an object is located in their retina and what is the shape of the object learn by using another model of learning – the backpropagation algorithm. And we will use the backpropagation algorithm to construct other robots that learn: the robots that learn to predict future events described in Chapter 5 and the robots that learn by imitating other robots described in Chapter 8. Can our model of learning replace the backpropagation algorithm so that we are entitled to say that all learning is based on the same basic mechanism? According to the backpropagation algorithm, the neural network receives an external "teaching input", compares its output with this "teaching input", and changes its connection weights in such a way that its outputs progressively become the same as the "teaching inputs". If we assume that evolution assigns reward value to a match between the neural network's output and the "teaching input" and punishment value to a mismatch, our model of learning could be extended to the various forms of learning which today are reproduced by using the backpropagation algorithm.

Other forms of learning that our model of learning should be able to reproduce are sensitization and habituation. Sensitization is the progressive amplification of a response as a consequence of the repeated presentation of the same stimulus. Habituation is the opposite of sensitization. It is the decrease in the response to a stimulus after repeated presentations of the stimulus. Sensitization and habituation can be seen as forms of classical conditioning and, since we have reproduced classical conditioning by using our model of learning (Section 6), perhaps our model of learning can also reproduce sensitization and habituation.

Can our model of learning reproduce other aspects of learning and other changes in behaviour? Can we construct robots that, when they are learning a complex behaviour – for example, the robotic equivalent of driving a car – are unable to respond to other, irrelevant, stimuli, but when this complex behaviour has been learned and it becomes automatic, they are able to respond to these irrelevant stimuli while executing the complex behaviour – the robotic equivalent of talking with another person while driving the car? When a behaviour has become automatic, the control of the behaviour is to a large extent transferred from external sensory inputs to internal sensory inputs.

Very young infants look at their hand – an external sensory input – when they move their arm but then the control of their arm is transferred to the proprioceptive input that arrives to their brain from the muscles of the arm – an internal sensory input. The infants cease to look at their hand because they know where is their hand with no need to see it. Internal sensory input tends to be simpler than external sensory input, and this contributes to make a behaviour automatic. Can our model of learning reproduce these phenomena?

There are other phenomena of learning and of changes in behaviour that future robots should also be able to reproduce. Learned behaviours tend to be forgotten if one does not continue to practice them – for example, playing piano. Or what has been learned often needs to be adapted to new circumstances. Or behaviour can change for other reasons which are not learning. Imagine that, at some point of its life, a robot becomes blind or one of its two wheels ceases to function properly. How will the robot's behaviour change as a consequence of blindness or a damaged wheel? Or a robot's behaviour becomes psychopathological because of the robot's experiences during its life and, then, this pathological behaviour is reduced or eliminated through some form of psychotherapy. These are all forms of changes in behaviour that do happen during the life of both nonhuman animals and human beings and we must find out if the model of learning described in Section 2 can reproduce all these changes in behaviour, or these changes in behaviour require something else.

But if we are interested in how behaviour is acquired and how it changes, our robots have two more serious limitations. Behaviour depends on the body – on its size, its shape, and its sensory and motor organs. Our robots have a body which is decided by us, does not evolve in a succession of generations, and is identical for all the robots of a population. Future robots should have a body which is different in different individuals, is evolved, and is well adapted to the environment in which the robots live. A simple example is a population of robots whose bodies have different sizes. Our robots eat a food token when their body touches the food token and, therefore, robots with a larger body will eat more than robots with a smaller body. If the robots have bodies of different sizes and this size is encoded in their genes, the size of the robots' body will probably increase in a succession of generations. But if having a larger body is more costly in terms of energy, we predict that at the end of evolution the body of the robots will have an intermediate size. It will be large enough to eat a sufficient number of food tokens but not so large to avoid too high energy costs. We have not constructed these robots but they can be easily constructed.

The other limitation of our robots is that they reproduce evolution and learning but they do not reproduce development. Behaviour changes during the life of an individual because of the individual's experiences – learning – but it also changes because of genetic reasons – development. All animals are born with a set of genes which specify not only what they are at birth but also how they will change during the course of their life. Learning and development are different forms of change but they influence each

other and one cannot understand them if one does not understand how they interact together. What an individual learns depends on its experiences but also on its age. An individual's genes may dictate certain changes at certain ages but the manner in which the changes occur and the exact timing of their occurrence are influenced by the individual's experiences.

What develops is, first of all, the body. The body becomes larger and its shape also changes. And, since the brain is part of the body, the brain also develops like the rest of the body. In human beings the number of neurons is already rather large at birth but the connections between the neurons continue to increase in number for a considerable time after birth. If the body and the brain change, behaviour also changes. Human beings acquire the capacity to reach objects at a certain age, they walk at another age, they learn to speak at still another age, and all these ages are encoded in their genes. Some of the robots described in this Chapter develop during their life because they see different colours at different ages but this development is decided by us and is not evolved. In real animals development is the result of evolution, and evolution's really important task is to create not the initial state of an organism but a program of development – a life history – for the organism. The reproduction of this crucial aspect of behaviour – how behaviour changes during an individual's life because of the program of changes encoded in the individual's genes – remains a task for the future.

Development is also important for the demography of a population. In many of our robots population size is fixed but, even when population size changes because some robots die before they reach their maximum age and new robots are born, the robots are just "robots". They have no age at which they can have offspring, no age at which they cease to be able procreate, no age at which they become old. Some of our robots – for example the robots described in Chapter 7 dedicated to robotic families – have some of these properties but, again, these properties are decided by us and are not the result of a spontaneous process of evolution and development. Robots with a demography are future robots.

Another important limitation of our robots is that they reproduce how evolution influences learning but they ignore how learning influences evolution. Evolution is Darwinian, not Lamarckian. The changes in brain and behaviour which are due to learning do not influence the genes and are not transmitted to the next generation. However, the fitness of an individual is the fitness not of the individual's genotype but of the actual individual – the phenotype – and the phenotype is changed by learning. An individual remains alive and reproduces not only because of its genes but also because of what the individual learns during its life. Therefore, learning indirectly influences evolution and can channel evolution in specific directions. And, for today's human beings, it is even less clear what determines the reproductive chances of an individual.

Still another limitation concerns evolution itself. The robots described in this Chapter do not start their *learning* from zero but – like real animals – they start their learning from what they inherit at birth. However, they start their *evolution* from zero

because the robots of the first generation have random genes and, therefore, random neural networks – and this is not true for real animals. Real animals evolve by starting from the results of past evolution and this explains many important phenomena such as the origin of new species and what evolutionary biologists call pre-adaptation – an already evolved behaviour may be a favourable pre-condition for the evolution of another behaviour – and ex-aptation – an evolved behaviour may acquire a new adaptive function. If we want to construct robots that reproduce these phenomena, future robots will have to evolve starting not from zero but from the results of past evolution.

And a final – and obvious – limitation of the robots of this Chapter is that, like the robots of the preceding chapter, they resemble nonhuman animals rather than human beings. Human beings acquire their behaviours in more different ways than nonhuman animals. They acquire their behaviours because they are told or taught by other human beings. They receive many of their rewards and punishments for learning not from nature but from other human beings. And they do not only change their behaviour in the course of their life but their cultures and technologies also change and these changes have a crucial impact on their behaviour. We have begun this Chapter by saying that we cannot really know and understand human beings unless we know and understand how they become what they are. We conclude the chapter by saying that much of the work is still to be done.

CHAPTER 4

Robots that have language

When ME compares human beings with nonhuman animals, ME is struck by how much time human beings spend talking with each other on all sorts of things and it is convinced that the complexity of their social life and of their societies is due to the complexity of their language. ME knows very well that language is primarily a communication tool and it agrees with the ancient philosopher Gorgias that "with the smallest and almost unperceivable bodies words succeed in accomplishing the most divine enterprises". But what interests ME is not language as a communication tool but how language changes the model of the world which is in the brain of human beings and guides their behaviour. Language divides non-linguistic experiences into separate pieces that can be recombined together and, in this manner, it makes the model of the world which human beings have in their brain richer and more articulated than the model of the world of nonhuman animals and their behaviour more effective. In addition, language makes human beings live not only in the concrete world of their sensory-motor experiences but also in an world of abstract entities because a word ignores the specific properties of the different things one refers to when one uses the word – this object is a computer but a "computer" is not this particular object. This – ME thinks – has many advantages but its also disadvantages, and one disadvantage is that language induces human beings to think that there are "essences" and to look for these "essences", while reality is not made of "essences" and to look for "essences" is a waste of time.

Given the centrality of language in the adaptive pattern of human beings, ME considers constructing robots that have language a high priority in its program of research. But human language is so complicated that the robots constructed by ME only reproduce some very basic aspects of language and most of the work lies in the future.

1. The cognitive consequences of having language

Why do human beings have language? What are the adaptive advantages of having language that explain why human beings have language? With language human beings inform each other, discuss with others, convince others to do or not to do this or that thing, make plans with others, and coordinate their behaviour with the behaviour of

others. In fact, with language human beings do more things than with their hands and their tools. It is because human beings have language that their social life and their societies are so complex. To convince oneself of the importance of language, one has only to imagine what would be the social life of human beings if they did not have language. It would not be very different from the social life of nonhuman animals.

This is language as a communications tool and these are the *social* advantages of possessing a language. But language is also a cognitive tool and it also has *cognitive* advantages. Language isolates specific parts of human beings' experiences and creates a richer and more articulated "model of the world" in their mind which makes them behave more effectively with respect to the world. This is, perhaps, the most important difference between human language and animal communication. Animal communication is only a tool for social interaction. Human language is both a tool for social interaction and a tool for having a better model of the world in one's brain. The cognitive advantages of language are so important that language may have first evolved in human beings not (only) because it made their social life more effective but also because it made them more intelligent. Language helps human beings to focus their attention on specific aspects of their experiences, to recombine parts of these experiences in new ways, to sub-divide problems into sub-problems, to construct plans for solving complex tasks. And while nonhuman animals only communicate with one another, human beings use language not only to communicate with other human beings but also to communicate with themselves. They talk to themselves. (We discuss talking to oneself in the next Chapter which is dedicated to robots that have a mental life.)

Language is such an important characteristic of human beings that, if we want to construct robots that can really be said to be human, the robots must necessarily possess a human language and this language must not only allow the robots to interact together in complex ways but it must also create in their brain a better representation of the world in which they live. There is much work on language in robots but most of this work is dedicated to language as communication, not on language as a cognitive tool. The robots of this Chapter are constructed to better understand the cognitive consequences of having language.

2. Meaning as co-variation between sounds and nonlinguistic experiences

Language is made of sounds which have "meaning". What is for a linguistic sound to have meaning? Linguistic sounds have meaning because specific sounds co-vary with specific nonlinguistic experiences and these co-variations are captured by the individual's brain. The sound "ball" has a meaning for an English-speaking person because it co-varies with the nonlinguistic experiences of seeing a ball, touching a ball, or playing with a ball. The same is true for our robots. If a robot hears the same sound when the robot has certain nonlinguistic experiences, this co-variation is captured by the robot's brain and the robot is able to respond appropriately when it hears the sound

(language understanding) and to produce the sound in the appropriate circumstances (speaking). We know that a sound has meaning for a robot and which meaning it has for the robot not only because we see how the robot reacts to the sound but also because we can examine the robot's brain and see how the co-variation between the sound and certain nonlinguistic experiences has been incorporated in the robot's brain. For example, when the robot hears the sound "ball" for the first time, the sound causes a certain pattern of activation in the internal neurons of the robot's neural network and this pattern of activation reflects only the sound "ball". When the sound "ball" has acquired a meaning for the robot, the activation pattern changes and it reflects not only the sound "ball" but also the robot's nonlinguistic experiences which have accompanied its hearing the sound "ball".

But what is important about human language is that linguistic sounds co-vary not with *whole* nonlinguistic experiences but with *specific parts* of nonlinguistic experiences. The experiences of nonhuman animals are wholes. The experiences of human beings are made of parts because human beings have language. If a linguistic sound co-varies with one specific component of a nonlinguistic experience but not with the rest of the nonlinguistic experience, this specific component is isolated from the rest of the nonlinguistic experience and so is represented in the brain.

The same is true for our robots. A robot sees a ball fall on the ground and it hears the sound "ball". Then, in another occasion, the robot sees a ball *roll* on the ground and it hears the same sound "ball". The two experiences are different but they have something in common. The sound "ball" co-varies with one specific part of the robot's nonlinguistic experiences – the ball – and not with other parts – the falling or the rolling of the ball. Hearing the sound "ball" isolates the ball from the rest of the robot's nonlinguistic experiences and the ball is separately represented in the robot's brain. The same is true for the sound "fall" and the sound "roll". If the robot sees a ball or a box fall on the ground and in both cases it hears the same sound "fall", the robot's brain captures the co-variation of the sound "fall" with the event of falling and not with the other components of its nonlinguistic experiences: the ball and the box. If a ball is white, the robot can hear the sound "ball" but it can also hear the sound "white" and, if the sound "white" is also heard by the robot when it sees a white *box*, the robot isolates the colour of the ball from the ball and the white colour is separately represented in the robot's brain. In this manner linguistic sounds cause the robot's brain to represent the world as made up of separate components and this has important consequences for the robot's behaviour. The robot can look for a ball or for a box independently from their colour or it can look for things which are white independently of whether they are balls or boxes. And what is important, the robot can respond appropriately to *new* experiences. The experiences are new but their parts are known to the robot because the robot's brain has incorporated in its structure and way of functioning the co-variation between different linguistic sounds with these different parts. The robot has never seen a white box fall on the ground but it responds appropriately to the sequence of sounds "white box fall" – and this makes the robot's behaviour more creative and more adaptive.

Language helps to segment a nonlinguistic experience into separate pieces and it also helps to segment a nonlinguistic experience into a *succession* of separate nonlinguistic experiences. Here is an example of how this can make a robot's behaviour more effective.

The robots are not our usual robots but they have a human-like (simulated) body. They have an arm made up of two parts – arm and forearm – which they can move separately and at the end of the arm there is a hand with fingers that they can also move so as to grasp an object. The robots' neural network has three different types of sensory neurons: visual neurons, proprioceptive neurons, and touch neurons. The visual neurons tell the robots where is the object, the proprioceptive neurons which are located in the muscles of the arm encode the current position of the arm so that the robots know where is their hand even if they do not see their hand, and the touch neurons are activated when their fingers touch the object. The motor neurons control the movements of the robots' arm and of the fingers of the robots' hand (see Figure 1).

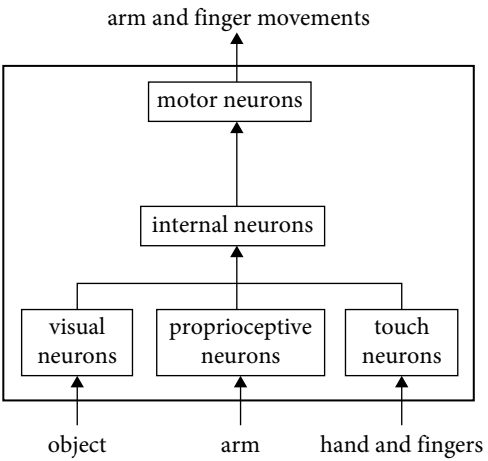


Figure 1. The neural network of a robot that receives a visual input from an object, a proprioceptive input from its arm, and a touch input from its hand and fingers, and responds by moving its arm and fingers.

The connection weights of the robots' neural network evolve in a succession of generations. The life of a robot is a succession of episodes and the position of the object and the initial position of the arm are different in different episodes. The robots that reproduce are those that are better at moving their arm so as to reach the object with their hand, grasping the object with their fingers, and then moving the object to another place (see Figure 2).

We compare two populations of robots, one with language and the other one without language. The robots without language evolve the appropriate behaviours by using only the nonlinguistic sensory input from their eyes, their arm, and their fingers. The neural network of the robots with language has an additional set of auditory neurons

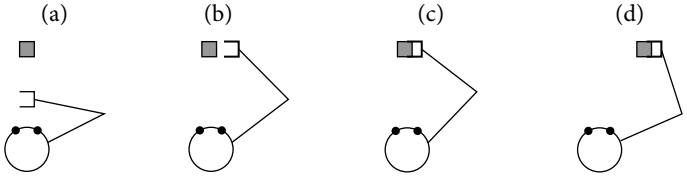


Figure 2. The robot sees an object (a), reaches the object with its arm (b), grasps the object with its fingers (c), and moves the object to another place (d).

(see Figure 3) and the robots hear one sound ("reach") when they must move their arm and reach the object with their hand, another sound ("grasp") when they must grasp the object with their fingers, and a third sound ("move") when they must move the object to another place (see Figure 4).

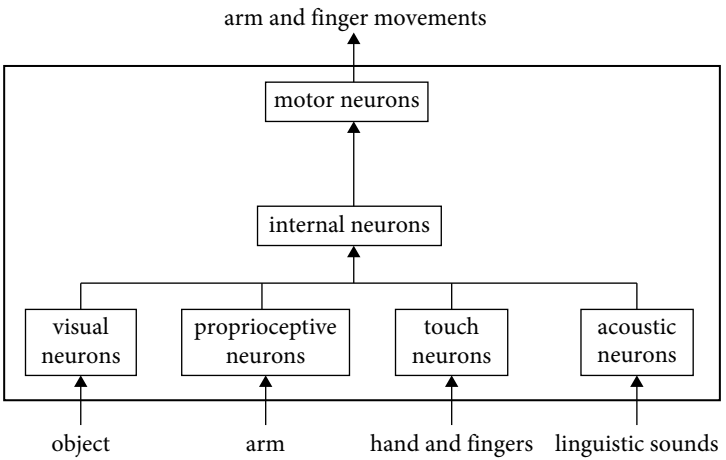


Figure 3. The neural network of a robot which, is addition to the sensory input from the object and from its arm and fingers, hears a succession of three linguistic sounds: "reach", "grasp", and "move".

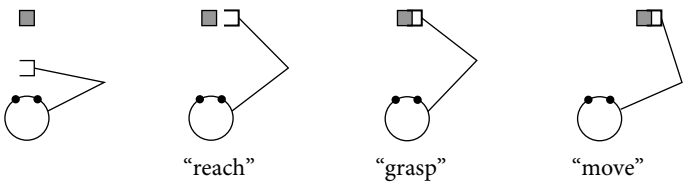


Figure 4. The robot hears the sound "reach" when it must reach the object with its hand, the sound "grasp" when it must grasp the object with its fingers, and the sound "move" when, after having grasped the object, it must move the object to another place.

What are the results? At the end of evolution the robots with language have more fitness than the robots without language. They are better than the robots without language at doing what they must do, and this is especially true for the action of moving the object to another place when they have grasped the object with their fingers. Why? Language segments the robots' sensory-motor experience into a succession of separate pieces and it helps the robots to do what they must do at any given time. This is especially important for the action of moving the object to another place. When the robot reaches the object with its hand, it receives a touch sensory input from its fingers but it is only when the robot has actually grasped the object with its fingers that it can move the object to another place. Hearing the sound "move" helps the robot to know when it is time to move the object. This is a very simple demonstration of how language segments sensory-motor experiences – and reality – into distinct pieces and makes behaviour more effective.

3. Classes of linguistic sounds

Now we ask a question which is almost inevitable to ask when one talks about language. Are there different classes of linguistic sounds? Can we construct robots which can be said to have verbs, nouns, adjectives, and adverbs? Our next robots try to answer these questions.

The new robots have an arm and a hand with fingers like our preceding robots but, at the beginning of each episode, they are already grasping the object with their fingers. The robots can hear one of two sounds: the sound "push" or the sound "pull". The two sounds are commands that the robots must obey if they want to survive and have offspring (see Figure 5).

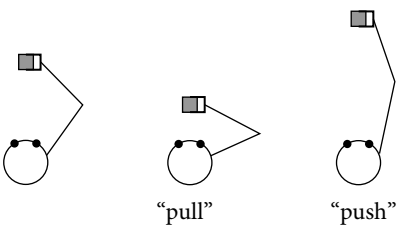


Figure 5. The robot is grasping an object with its hand and it must respond to the sound "pull" by pulling the object towards itself and to the sound "push" by pushing the object away from itself.

The neural network of these robots does not have visual neurons and, therefore, the robots do not see the object or their arm. But they know the position of their arm because on the muscles of their arm there are proprioceptive neurons which tell the robots' brain which is the present position of their arm. In addition, their neural network has

auditory neurons with which the robots hear either the sound "push" or the sound "pull", and a robot's fitness depends on how they respond to these two sounds. When a robot hears the sound "push", it must move its arm so as to push the object away from its body and when it hears the sound "pull", it must pull the object towards its body. The neural network of the robots has motor neurons which encode the movements of the robot's arm and, in a succession of generations, the robots evolve the capacity to respond appropriately to the two sounds "push" and "pull". The two sounds co-vary with the action that the robots must do, not with the object on which the action must be done – and they cannot co-vary with the object on which the action must be done because there is only one object in the robot's world. Therefore, for the robots the two sounds "push" and pull" are verbs.

That verbs are sounds that co-vary with the robot's actions and not with the objects on which the actions are done, can be seen more clearly if we slightly modify our robotic scenario. The object which a robot is grasping with its hand and which it must either push or pull is not always the same object but it can be either a ball or a box – two objects with different shapes (see Figure 6).

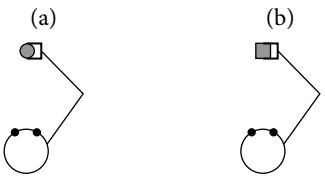


Figure 6. The robot is grasping with its hand either a ball (a) or a box (b).

Like the preceding robots, when they hear the sound "push", the robots must push the ball or the box away from themselves and, when they hear the sound "pull", they must pull the ball or the box towards themselves. But, unlike the preceding robots, the neural network of these robots has visual neurons so that the robots can see if what they are gasping with their hand is a ball or a box. The robots see that the ball and the box are different but they must ignore what makes them different, and hearing only the sounds "push" or "pull" helps them to ignore – abstract from – what makes the ball and the box different. This is why we call the two sounds "push" and "pull" verbs.

If verbs co-vary with the actions of the robot but not with the objects on which the actions are done, what are nouns? To construct robots that have nouns, we change the preceding scenario. At the beginning of each episode the robots see both a ball and a box but the robots are not grasping anything with their hand (see Figure 7). The robots' arm is in a randomly chosen position and what the robots must do is move their arm so as to reach either the ball or the box. Again, the robots hear one of two sounds, the sound "ball" and the sound "box" and, at the end of evolution, when they hear the sound "ball", they reach the ball and, when they hear the sound "box", they reach the

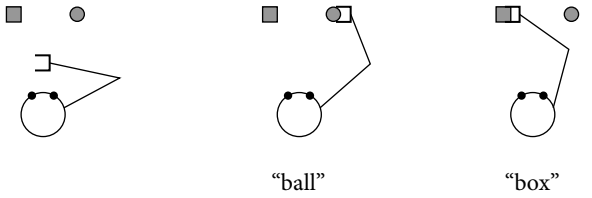


Figure 7. The robot reaches the ball with its hand when it hears the sound "ball" and it reaches the box when it hears the sound "box".

box. The action of reaching is the same but the objects on which the action is done are different. For the robots the two sounds "ball" and "box" are nouns.

To better understand what makes nouns different from verbs we construct a more complex scenario and a more complex language. The preceding robots were capable of doing only one action, the action of reaching an object, either the ball or the box, by moving their arm. The new robots can do two different and more complex actions: they can reach and then push the ball or the box away from their body and they can reach and pull the ball or the box towards their body. In each episode the robots see both the ball and the box and, therefore, for these robots hearing one sound is not sufficient to know what they must do. The robots need to hear *two* sounds. (Sounds happen in time and, therefore, the two sounds form a temporal sequence but, for these robots, the temporal order of the two sounds is irrelevant.) They must know what they must do – push or pull – and they must also know which object must be involved in their action – the ball or the box. In different episodes the robots hear one of four possible sequences of two sounds: "push ball", "pull ball", "push box", "pull box". These sequences are verb/noun sentences. The verb co-varies with the action that the robot must do and the noun co-varies with the object on which the action must be done. Verb-noun sequences cause a more articulated representation of behaviour in the robot's brain because they segment behaviour into two parts: what the robot does and which object is involved in what the robot does.

We have said that, for these robots, "push" and "pull" are verbs and "ball" and "box" are nouns. But are the sounds "ball" and "box" really nouns for the robots? The two sounds might be nouns but they might also be adjectives which co-vary not with the objects but with the shape of the objects, which is round for the ball and square for the box. So what are adjectives and how can we distinguish between nouns and adjectives?

The next robots have a language which includes adjectives and the sequences of sounds which they hear are made of an adjective and a noun. The robot's environment contains both a ball and a box but now both the ball and the box can be of two different colours: grey and white. There is a grey ball and a white ball and there is a grey box and a white box. The robot's language is made of four sounds. The sound "ball" co-varies with balls, independently of their colour, and the sound "box" co-varies with boxes, also independently of their colour. Of the other two sounds, one – the sound "grey" – co-varies with the colour grey and the other one – the sound "white" – with

the colour white, independently of whether what is grey or white is a ball or a box. In each episode the robot sees all *four* objects at the same time – the grey ball and the white ball, the grey box and the white box – and the robot must reach one of the four objects with its arm by responding to a linguistic command. (These robots do not need verbs because they must always do the same action: reaching an object.) In this case too, the command cannot be made of a single sound because one single sound can specify the object (ball or box) to be reached but not its colour (grey or white), or it can specify the colour of the object but not the object and, therefore, the robots would not know which object to reach with their hand. The command must be a sequence of two sounds, with one sound co-varying with the object to be reached and the other sound co-varying with its colour (see Figure 8).

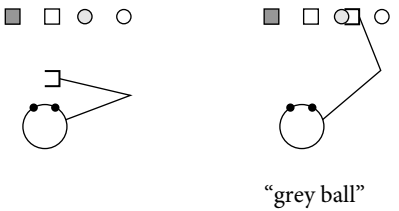


Figure 8. The robot sees two balls, one grey and one white, and two boxes, one grey and one white. The robot hears the sounds "grey ball" and it reaches the grey ball with its hand.

For English speakers one of the two sounds is a noun ("ball" or "box") and the other sound is an adjective ("grey" and "white"). But is this true for our robots? Balls and boxes have different shapes (round or square) and, therefore, instead of two nouns – "ball" and "box" – the two sounds might be two adjectives – "round" and "square". Given the very simple environment and the very simple behavioural repertoire of these robots, the robots' neural network cannot distinguish between sounds which are nouns and sounds which are adjectives. So, why do we call some sounds nouns and other sounds adjectives? If we look at the nonlinguistic experiences with which nouns and adjectives co-vary (which is our ultimate criterion for everything that concerns language), we find that nouns tend to co-vary with nonlinguistic experiences to which human beings tend to respond with specific actions while this is not true for adjectives. For English speakers, the sound "ball" is a noun because it co-varies with things to which they normally respond by doing the action of making the thing roll or playing with it, and the sound "box" also is a noun because it co-varies with things to which they normally respond by doing the action of opening the thing or putting something inside it. The sound "grey" and the sound "white" are adjectives because they do *not* co-vary with anything to which English speakers respond with specific actions. Many different things can be grey or white but English speakers do not respond in the same way to all grey or white things. This implies that, if we want to construct robots that can be said to distinguish between nouns and adjectives, the environment and the

behavioural repertoire of the robots must be more complex. The robots must be able to do many different types of actions and the environment must contain many types of objects with different shapes and different colours.

Adverbs are for verbs what adjectives are for nouns. At beginning of each episode, a robot is grasping an object with its hand and, if it hears the sound "push", it responds by pushing the object away from its body while, if hears the sound "pull", it responds by pulling the object towards its body. But now the robot can push or pull an object in two different ways. It can push or pull the object slowly or it can push or pull the object more quickly. In each episode the robot hears one of four possible sequences of two sounds: "push slowly", "push quickly", "pull slowly", "pull quickly". One of the two sounds co-varies with the action that the robot must do to obey the command – pushing or pulling the object – and the other sound co-varies with the manner in which the action must be done – slowly or quickly. To obey the command – and to survive and have offspring – if the second sound is "slowly", the robot must execute the action of pushing or pulling the object slowly while, if the second sound is "quickly", the robot must execute the action more quickly. The two sounds are adverbs.

We construct a final type of robots whose language is more complex because it contains all the eight sounds of the preceding robots – "push", "pull", "ball", "box", "white", "grey", "slowly", "quickly" – and each command is a sequence of four sounds: "push white ball slowly", "pull white ball slowly, "push black ball slowly", "push white box slowly", "push white ball quickly", etc. The language contains only eight sounds but the robots are able to respond correctly to a very large number of different commands. This is typical of human language. Human beings have linguistic sounds (sentences) which are made of shorter linguistic sounds (words) and the meaning of the sequence of sounds is a combination of the meanings of the sounds that make up the sequence. This is a very important characteristic of language. With only a limited number of words – tens of thousands – human beings are able to produce and understand a practically infinite number of sentences. And, what is even more important, with their compositional language human beings are able to produce and to understand sequences of sounds they have never produced or understood before. This is what we find with our robots. At the end of evolution the robots hear a command which is a sequence of sounds they have never heard before and they respond appropriately to the command.

We have called the different classes of linguistic sounds of our robots nouns, verbs, adjectives, and adverbs. Linguists call verbs, nouns, adjectives, and adverbs "grammatical classes" but a language can have grammatical classes only if it has syntax. The language of our robots does not have syntax and, therefore, it does not have grammatical classes. Not having syntax implies that our robots would not be able to distinguish between the sequence of sounds "put ball on box" and the sequence of sounds "put box on ball" because they are unable to take the order of sounds into consideration. And they would not be able to recognize that some sequences of sounds – for example, the sequence "push pull white slowly" or the sequence "push ball box quickly" – are

"wrong" in the sense that the robots are unable to respond appropriately to them. Linguists say that these sequences of sounds violate syntactic rules, but our robots do not have syntactic rules. Constructing robots that have a language with syntactic rules is an important task for the future because syntactic rules are an important characteristic of human language which distinguishes human language from animal communication systems.

4. Language helps human beings to categorize their environment

In the real world everything is different from everything else but, to survive and reproduce, animals must take into consideration some differences and ignore other differences. Each individual plant is different from all other plants but an animal must respond to some plants by eating them and to other plants by avoiding them. The animal must ignore the differences among the first type of plants and the differences among the second type of plants and it must take into consideration what makes the first type of plants different from the second type of plants. All the different things to which an animal responds in the same way are said to be a *category* while the things to which the animal responds in different ways are different categories. This definition of categories is in terms of behaviour but, as we will see, categories can also be defined in neural terms.

An important advantage of possessing language is that language helps human beings to better categorize the entities that make up the environment in which they live and, therefore, to behave better. If a human being hears the same sound when he or she is experiencing two different things, the sound make it easier for the human being to know that the different things belong to the same category and, therefore, that he or she must respond in the same way to both things. When a human being hears different sounds when he or she is experiencing different things, the different sounds make it easier to know that the things belong to different categories and, therefore, that he or she must respond in different ways to them. This is what we show with our next robots. We compare robots which have language with robots which do not have language and we find that the robots which have language behave better than the robots without language because language improves the way in which the robots' categories are represented in their brain.

The robots live in an environment which contains food tokens and poisonous tokens. All the food tokens have a roundish shape and all the poisonous tokens have an angular shape but the problem is that the tokens that belong to the same category do not have exactly the same shape. The food tokens have a roundish shape but no two food tokens have exactly the same roundish shape and are perceptually identical, and the same for the angular shape of the poisonous tokens (see Figure 9). Therefore, when a robot sees a token, the robot must first categorize or recognize the token as food or as poison in order to know what to do with the token.

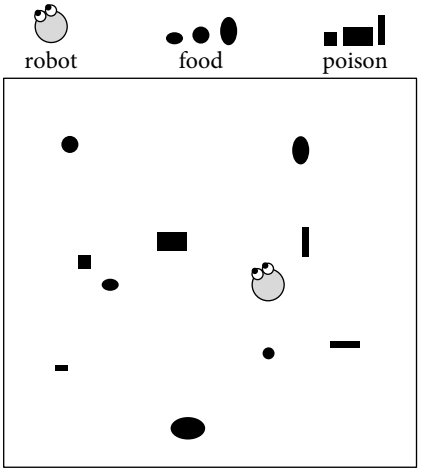


Figure 9. The environment contains roundish tokens which are food and angular tokens which are poison but the food tokens have different roundish shapes and the poisonous tokens different angular shapes.

The neural network of the robots that do not have language is our basic neural network. It has visual neurons which encode both the shape and the location of the tokens, the visual neurons are connected to the internal neurons which are connected to the motor neurons. The motor neurons control the speed of the two wheels of the robot and, therefore, the robot's movements in the environment (see Figure 10).

Eating a food token increases a robot's energy while eating a poisonous token decreases the robot's energy and, since the robots that reproduce are those that have more energy, the robots must eat the food tokens and avoid the poisonous ones. In fact, in a succession of generations the robots evolve the appropriate behaviours – which means that they are able to distinguish between food tokens and poisonous tokens based on their roundish or angular form.

Now we construct another population of robots which are identical to the preceding robots except that, when a robot sees a token, the robot also hears one of two sounds, the sound "food" or the sound "poison", and the two sounds co-vary with the category of the token. (These "sounds" are produced by a virtual robot which wants to help the robots to eat the edible tokens and avoid the poisonous tokens.) When a robot sees a food token, whatever the shape of the token, it hears the sound "food", and when it sees a poisonous token, whatever the shape of the token, it hears the sound "poison". The neural network of these robots is more complex because it is made of two modules: a nonlinguistic module and a linguistic module (see Figure 11).

The nonlinguistic module is identical to the entire neural network of the robots without language (see Figure 10). The linguistic module is made of acoustic neurons which encode the sounds which the robots hear and these acoustic neurons send their

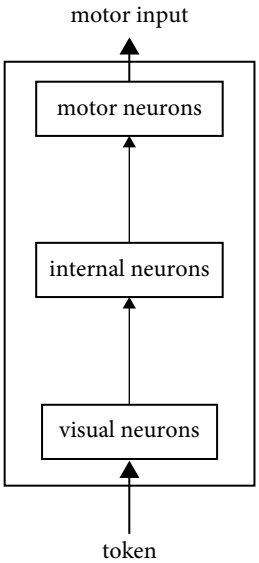


Figure 10. The neural network of the robots that do not have language. The robots see a roundish token (food) or an angular token (poison) and they must respond by either approaching or avoiding the token.

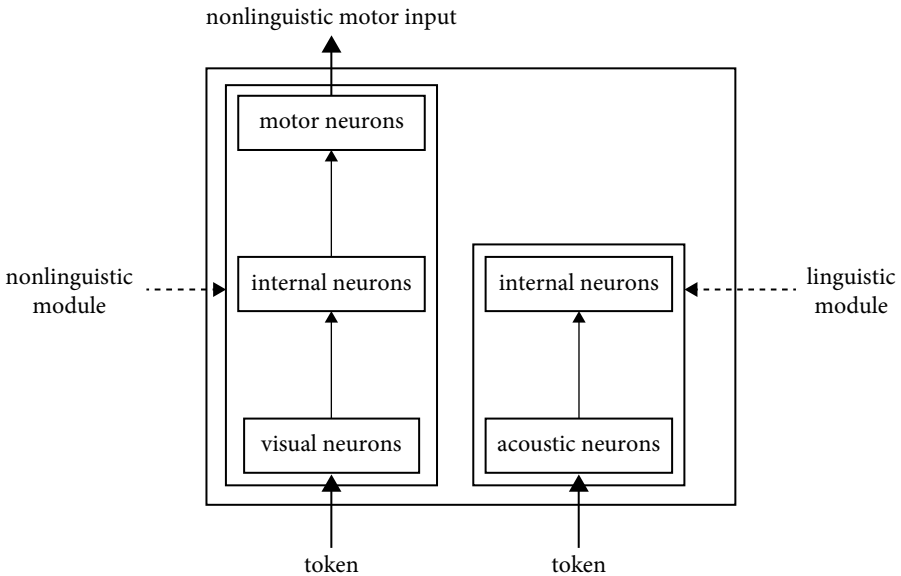


Figure 11. The neural network of the robots which have language is made of two modules, a non-linguistic module and a linguistic module.

connections to a separate set of internal neurons. But since the internal neurons of the linguistic module send their connections to the internal neurons of the nonlinguistic module, the sound which a robot hears when it sees a token influences how the robot responds to the token. As always, the robots of the first generation have random connection weights and, therefore, the sounds have no significant effect on the robots' behaviour. Then, in a succession of generations, the inherited connection weights change and, when they have changed, we find that hearing the two sounds helps the robots to recognize the category of the tokens and to respond appropriately to them. The robots with language have more fitness than the robots without language.

Why are the robots with language better at eating the food tokens and avoiding the poisonous tokens than the robots without language? To behave appropriately, the robots must take into consideration the differences *between* the tokens that belong to the two different categories and they must ignore the differences *among* the tokens that belong to the same category. The differences among the tokens that belong to the same category are only sensory "noise" that the brain of these robots must be able to ignore, and language helps them to eliminate this sensory "noise". The proof is that, if we evolve the robots in another environment in which all the food tokens have exactly the same round shape and all the poisonous tokens have exactly the same angular shape – and, therefore, there is no sensory "noise" – having or not having language does not make any difference.

We have defined a category in terms of behaviour. A category is all the tokens to which a robot responds in the same way. But our robots are neuro-robots and we may better understand how language helps the robots to categorize the different tokens if we look at the robots' brain. When the robot sees a token and the token activates the internal neurons of the robot's neural network, each internal neuron has a certain level of activation and we call the pattern of activation of all the internal neurons the internal representation of the token. Each token has a different shape and, therefore, each token produces a different internal representation in the robot's neural network. This is true for both the robots without language and the robots with language but with a difference. In the robots without language the internal representation of a token depends only on the shape of the token. In the robots with language the internal representation of a token depends on *both* the token's shape *and* the sound which the robots hear when they see the token because the sound influences the internal representation of the token. And this is why language makes a difference. If we examine the internal representations of the different tokens in the robots' brains, we find two things. The internal representations of the tokens which belong to the *same* category are *more similar* in the robots with language than in the robots without language and the internal representations of the tokens that belong to the two *different* categories are *more different* in the robots with language than in the robots without language. Language reduces the sensory "noise" which makes one token of one category different from another token of the same category and it amplifies the difference between the tokens of the two categories.

This can be measured precisely – quantitatively – in the following way. Since the neural network of these robots has only two internal neurons, we can imagine an abstract space with two dimensions, one for one internal neuron and the other one for the other internal neuron. The internal representation of a token is a point in this abstract bi-dimensional space. For each of the two dimensions of the space, the point is located in a position which reflects the level of activation of the corresponding internal neuron. If two tokens have similar internal representations, the two points are near to one another in the abstract space of internal representations. If their internal representations are different, the two points are more distant from one another.

The robots of the initial generation have random connection weights in their neural network and, therefore, the points that correspond to the different tokens are randomly distributed in the entire space of internal representations (see Figure 12a). The robots still do not know how to represent neurally the two categories of tokens, they confound the food tokens with the poisonous tokens and, therefore, they are unable to respond appropriately to them. Then, in a succession of generations, the connection weights of the robots' neural networks change and the robots evolve the capacity to distinguish between the food tokens and the poisonous tokens. If we examine the robots' neural networks at the end of evolution, we find that the internal representations of the tokens form two separate "clouds" of points in the abstract space of internal representations. The internal representations of the food tokens have become more similar and they form one "cloud" of points which are near to one another. The internal representations of the poisonous tokens also have become more similar and they form another, separate, "cloud" of points. And the internal representations of the two categories of tokens have become more different and this is reflected in the distance between the two "clouds" of points (see Figure 12b and 12c).

A "cloud" of points is a category in neural terms. The robots respond in the same way to all the tokens whose internal representations are part of the same "cloud" and they respond in another way to the tokens whose internal representations are part of the other "cloud". They approach the food tokens and they avoid the poisonous tokens. In the early stages of evolution the two "clouds" are very large because the tokens of the same category still have very different internal representations, and there is some overlap between the two "clouds" – and this means that the robots may confound a food token with a poisonous token, or vice versa. At the end of evolution, the two "clouds" have become smaller and they are well separated one from the other, and this explains why the robots respond appropriately to the two categories of tokens. This is true for both the robots with language and the robots without language but, if we compare the two "clouds" of points of the two types of robots, we find that in the brain of the robots with language the two "clouds" are smaller and the distance between the two "clouds" larger than in the brain of the robots without language.

This explains why language helps the robots to categorize the different entities that constitute their environment and allows them to behave better with respect to these entities. We have shown this for nouns – "food" and "poison" – but it could also be shown

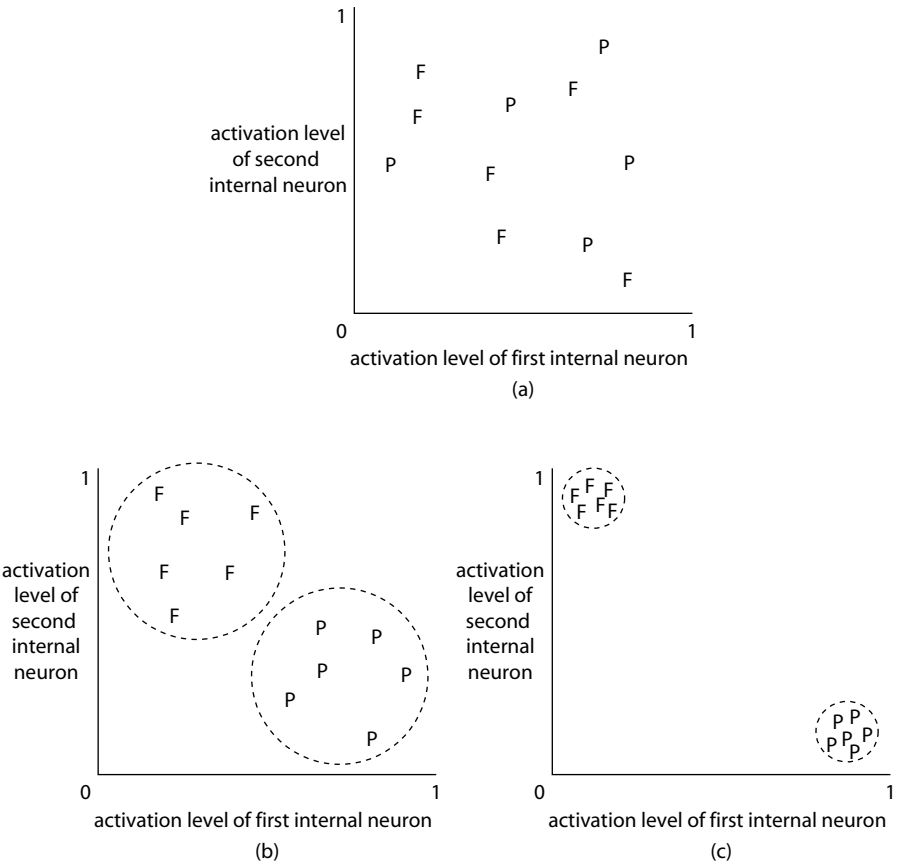


Figure 12. At the beginning of evolution, the internal representations of the food (F) and poisonous (P) tokens are mixed together (a). At the end, they form two separate "clouds" which, for the robots with language, are smaller and more distant than for the robots without language ((c) vs (b)).

for verbs, adjectives, and adverbs. One can *grasp* an object in different ways (verbs). Two *red* objects may not have exactly the same colour (adjectives). Two individuals can both walk *slowly* but they may not walk at exactly the same speed (adverbs). Language reduces the sensory "noise" represented by the differences among nonlinguistic experiences that belong to the *same* category and it emphasizes what makes one category *different* from another category.

But language does not only help the robots to divide their environment into categories. It can also create a *hierarchy* of categories in the robots' brain, and this hierarchical representation of the environment is another advantage of possessing language. This is what we show with our next robots.

The new scenario is a complication of the preceding scenario. The robots' environment still contains roundish food tokens and angular poisonous tokens but, while the poisonous tokens all belong to the same category, the food tokens are divided into two different categories because they contain two different types of energy. This is signalled by the colour of the food tokens: the food tokens that contain one type of energy are white (meat) and the tokens that contain the other type of energy are black (vegetables) (see Figure 13).

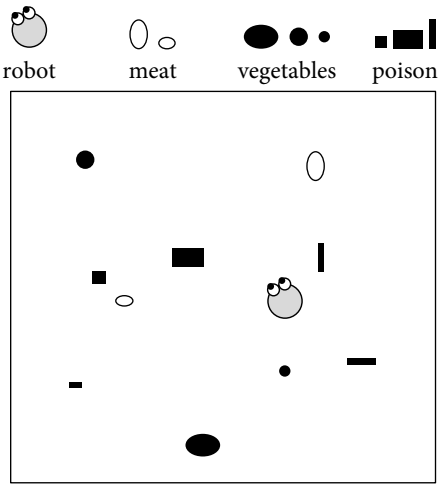


Figure 13. The environment contains poisonous tokens with an angular shape and food tokens with a roundish shape but, while all the poisonous are black, some food tokens are white (meat) and other food tokens are black (vegetables).

The robots must avoid the poisonous tokens but, since their body needs both types of energy, they must eat both the white food tokens and the black food tokens. A robot knows what its body needs because its neural network has two different types of hunger sensors whose activation level reflects the quantity of the two different types of energy currently contained in the robot's body, and the robot must look for roundish white tokens when its body needs the first type of energy and it must look for black roundish tokens when its body needs the second type of energy. Clearly, the life of these robots is more complicated. They must not only distinguish the food tokens from the poisonous tokens based on their shape but they must also distinguish between the two types of food tokens based on their colour.

What is the language which is most useful for these robots? The language certainly must include one sound, "poison", which co-varies with all poisonous tokens and which helps the robots to recognize and avoid the poisonous tokens. But what about the food tokens? There are three possible languages. In addition to the sound "poison", the first

language has only another sound – the sound "food" – which co-varies with all food tokens, both the white food tokens and the grey food tokens. In the second language, one sound – the sound "meat" – co-varies with the white food tokens and another sound – the sound "vegetables" – co-varies with the black food tokens. In the third and most complex language, in addition to the three sounds "poison", "meat", and "vegetables", there is a fourth sound, "food", which co-varies with *both* the white and the black food tokens. What are the results when we compare the fitness of the robots which have the first, the second, or the third language?

The first language, which is made only of the two sounds "food" and "poison", is not very useful to the robots because it does not help the robots to distinguish between the white food tokens and the black food tokens. The robots must distinguish between these two types of food because, depending on the state of their body, in one occasion they may need to eat a white food token and in another occasion they may need to eat a black food token, and they are not helped by the language to know what they must eat. The second language, which is made of three sounds – "poison", "meat", and "vegetables" – is more useful than the first language because it helps the robots to solve this problem. But the language which leads to highest level of fitness is the language which, in addition to the sounds "poison", "meat", and "vegetables", also has the sound "food".

Why the more complex language which also includes the sound "food" is better than the second language which does not have the sound "food"? Imagine that the robots are too young to respond appropriately to the state of their body. In some occasions their body needs the energy contained in the white food tokens while in another occasions it needs the energy contained in the black food tokens. The sounds that they hear are produced by their parents that want to help them. (As always, the sounds are produced by virtual robots hardwired by us.) A parent knows what is the state of the body of its child and, if the parent knows that its child needs only one type of energy, it produces the sound "meat" or the sound "vegetables". But if the parent knows that its child needs *both* types of energy, the parent produces the sound "food" which co-varies with all food tokens, independently from their colour. This is advantageous for the children robots. If they hear the sound "meat", they look for the white food tokens and, if they hear the sound "vegetables", they look for the black food tokens. But if they hear the sound "food", they look for any food token, independently from its colour.

These robots illustrate how linguistic sounds help the robots to organize their nonlinguistic experiences into a hierarchy of categories. There is the category of food and there are the two sub-categories of meat and vegetables. If a category is a "cloud" of points in the abstract space of internal representations of the robots' neural network, the internal representations of the white food tokens and the internal representations of the black food tokens form two different "clouds", the "cloud" of meat and the "cloud" of vegetables, but there is a third larger "cloud", the "cloud" of food, and this third "cloud" includes the two smaller "clouds" (see Figure 14). When they hear the sound "food", the robots ignore the colour of the food tokens and they concentrate on their shape

because it is their shape that distinguishes the food tokens from the poisonous tokens. This hierarchical organization of categories also exists in the brain of the robots which do not have language but language makes the hierarchy of categories more explicit – with positive consequences for the robots' behaviour.

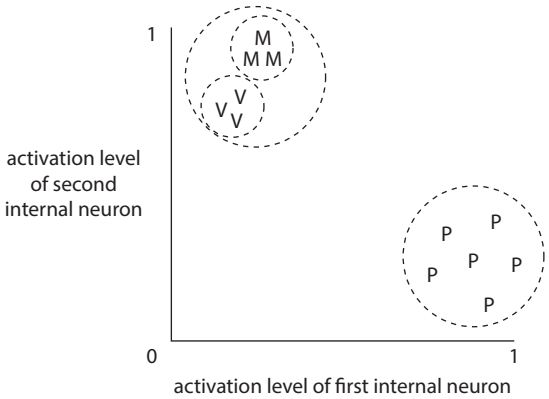


Figure 14. In the abstract space of internal representations the food tokens and the poisonous tokens form two separate "clouds" of points and the "cloud" of points of the food tokens includes two "sub-clouds" of meat tokens (M) and vegetables tokens (V).

All these robots tell us something important about words: the abstract nature of their meanings. We are not talking of the abstractness of abstract words but of the abstractness of all words – both concrete and abstract, both "table" and "justice". Categorization, by itself, requires abstraction. To respond in the same way to different things, all animals need to abstract from the differences among these things. To respond in different ways to similar things, they have to abstract from their similarities. But language help human beings to categorize their world and it makes them live not only in the concrete world of their nonlinguistic experiences but also in a world of abstract internal representations. Our definition of the abstractness of all words is in neural terms. Each individual token has a different shape and, therefore, when a robot sees a token, the token inevitably causes an internal representation in the robot's neural network which is different from the internal representations of all the other tokens that belong to the same category. This internal representation corresponds to a specific point in the abstract space of internal representations, and the internal representations caused by the tokens which belong to the same category – which are responded to with the same action – form a "cloud" of points close to one other and distant from the other "clouds" of points that correspond to the tokens which belong to other categories. The internal representations of the tokens which are more typical of a given category are points located more centrally in the "cloud" of points of the category while those which

are less typical correspond to points located more marginally, but no two tokens have internal representations that correspond exactly to the same point. This is true for both the robots which have language and the robots which do not have language. But if the robots are human robots and they have language, when a robot sees one token and at the same time it hears the linguistic sound which in its experience co-varies with the token, the internal representation of the token is a point which is nearer to the centre of the "cloud" and, therefore, it *abstracts* from what makes that particular token different from all the other tokens that belong to the same category. This is why language causes human beings to look at the world as made up of abstract entities. And this is true for all words, not only nouns such as "food", "poison", "meat", and "vegetables" – because all nonlinguistic experiences result in the creation of "clouds" of points in the abstract space of internal representations and all words push the internal representations of these nonlinguistic experiences towards the centre of their "clouds". It is one thing to see that one event happens after another event and a different thing to be able to use the word "after". It is one thing to see that an event causes another event and a different thing to be able to use the word "cause" or the word "because". It is one thing to have a sense of self and a different thing to be able to use the word "I". Words cause internal representations in the brain which are abstract in the sense that they tend to be at the centre of the "clouds" of points of different experiences and, therefore, ignore what makes these experiences different. This is true even when language is used to talk about things which are physically present – as is true for our robots – but the phenomenon is more pronounced when language is used to talk about things which are not physically present. When language is used to talk about things which are not physically present, the meaning of a word is abstract because it is a point situated at the centre of the space of internal representations and no specific non-linguistic experience may correspond to this point – although all words are used in some context, even only the context of other words, and therefore no word ever activates the exact centre of the cloud of "points" that constitutes its meaning.

Before we conclude this section we must ask another question on the influence of language on behaviour. Our robots respond exactly in the same way to all the food tokens: they "eat" the food tokens. But since the food tokens have different sizes and different shapes, the behaviour of the robot should be adapted to the particular size and shape of the token they are seeing. If the robots had a mouth, they would need to open more widely their mouth when they eat a large food token than when they eat a smaller food token. Categories are important to decide what to do with respect to what one sees, or perceives in some other way, but behaviour should be adapted to the specific member of the category one is currently perceiving. Does language help in adapting one's behaviour to the specific member of one category because it makes it easier to recognize the category?

5. The invention of language

The robots we have described so far evolve the capacity to respond appropriately to the linguistic sounds that they hear – language understanding – but they do not evolve the capacity to produce the appropriate sounds when they see the tokens – speaking. The language that they hear already exists and there is a virtual robot hardwired by us to produce the sound "food" when they see a food token and the sound "poison" when they see a poisonous token. But there was a time – tens of thousands of years ago – in which language did not exist and then human beings invented it. Can we construct robots that invent language? At the beginning the robots do not have language and then they acquire *both* the capacity to produce linguistic sounds in the appropriate circumstances *and* the capacity to respond appropriately to these linguistic sounds. To know how language was invented is a very complex and difficult problem and our next robots only try to shed some light on this problem.

The robots are the same robots that live in the environment with roundish food tokens and angular poisonous tokens but now there are no virtual robots hardwired by us to help the robots by producing the sound "food" and the sound "poison" in the appropriate circumstances. The neural network of the new robots is more complex (see Figure 15). Like the neural network of the preceding robots, it is made of a nonlinguistic module and a linguistic module but now the linguistic module is more complete and it has its own motor neurons which encode the sounds that the robots

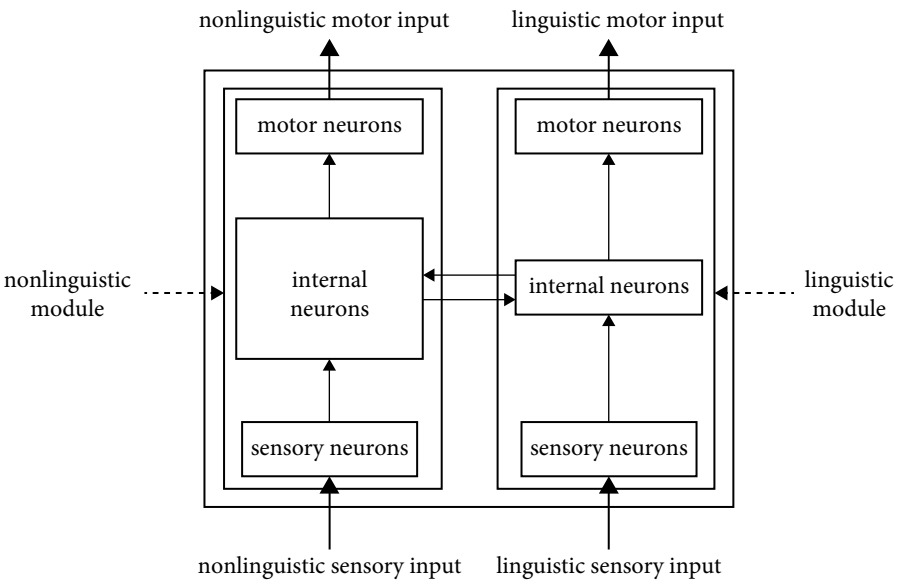


Figure 15. The neural network of robots which both produce and understand language.

produce with their phono-articulatory organs. Furthermore, the internal neurons of the two modules now have bi-directional connections. Not only the internal neurons of the linguistic module send their connections to the internal neurons of the non-linguistic module – language understanding – but, also, the internal neurons of the nonlinguistic module send their connections to the internal neurons of the linguistic module – language production.

When a robot sees a token, we randomly select another robot and we place this robot near to the first robot so that both robots see the same token. This second robot is the robot that speaks. It sees a token and it responds by producing a sound (see Figure 16a). The sound is heard by the first robot which responds by approaching or avoiding the token (see Figure 16b). Given this arrangement, all robots function as speaking or hearing robots in different occasions.

These robots evolve both the capacity to respond appropriately to the linguistic sounds that they hear – understanding language – and the capacity to produce the linguistic sounds in the appropriate circumstances – speaking. We already know what is to understand language. To understand language is to incorporate in one's brain the co-variations between specific heard sounds and specific nonlinguistic experiences so as to respond better to these nonlinguistic experiences. If this is understanding language, what is speaking? For our robots, speaking is using their phono-articulatory organs to produce one sound when they see a food token and a different sound when they see a poisonous token. But, since the robots speak to be understood, the *same* co-variations between sounds and specific nonlinguistic experiences must be incorporated in the brain of the robot that produces the sound and in the brain of the robot that understands the sound. And since all robots can function as speakers or hearers in different occasions, the same co-variations should be incorporated in the brains of all the robots.

Clearly, these robots have a more complex problem to solve. The preceding robots began with an already existing language. The new robots start from zero. The robots of the initial generation do not have language because they lack both the capacity to speak and the capacity to understand language. The connection weights of their neural network are random and, therefore, when they act as speakers, they respond to the sight of a token by producing one random sound, and when they hear a sound, they respond to the sound with a random behaviour. In a succession of generations the robots must evolve a language – where a "language" means that the *same* co-variations between sounds and specific tokens are incorporated in the brains of the robots that produce the sounds and of the robots that hear and understand the sounds, and the same co-variations are incorporated in the brains of all the robots of the population because all robots can act as speakers and hearers in different occasions.

The problem is that, in our scenario, while for a robot that hears a sound responding appropriately to the sound increases the robot's survival and reproductive chances, it is not clear why the speaker robot should produce one sound when it sees a food token and another sound when it sees a poisonous token. What is the adaptive advantage of speaking for our robots? For our robots speaking is an altruistic behaviour

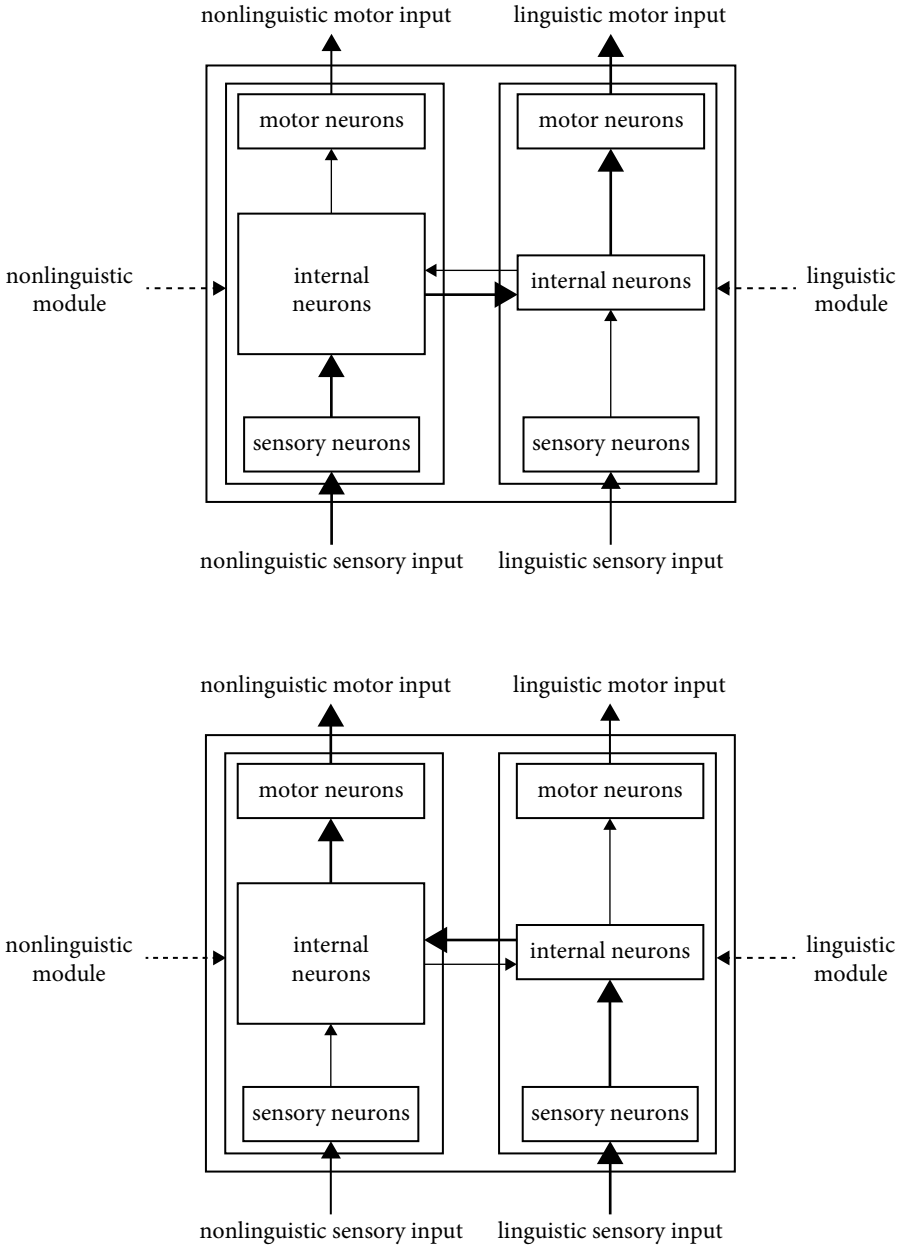


Figure 16. The neural network of Figure 15 when it is used to produce linguistic sounds (a) and when it is used to understand linguistic sounds (b).

which increases the survival/reproductive chances of the hearer but not those of the speaker. Since the robots compete with each other for survival and reproduction, why should they evolve the capacity to speak? But if there are no robots that speak, there are no robots that understand and, therefore, there is no language. This is exactly what we find with our robots. At the end, the robots produce different sounds in response to the tokens that belong to the same category and perhaps the same sound in response to a food token and to a poisonous token, and each robot produces its own sounds. The robots do not have evolved a language and, as a consequence, their fitness remains as low as the fitness of the robots without language.

So, how can language evolve? If speaking is altruistic behaviour, a robot may speak if the hearer has its same genes. As we will see in Chapter 7 dedicated to robotic families, robots that have the same or similar genes (parents and offspring, siblings, etc.) tend to act altruistically towards each other. Speaking may not affect or it can even reduce the reproductive chances of the speaker while it increases the reproductive chances of the hearer but, if the two robots have the same or similar genes, the genes which encode the altruistic behaviour of speaking will remain in the genetic pool of the population and the robots will speak to benefit other robots.

Is this how language has evolved? We construct another population of robots in which the robot that speaks is not randomly selected from the population of robots but is a sibling of the hearer robot and, therefore, the two robots have similar genes. What we find is that, in these conditions, a language does emerge. The robots evolve both the behaviour of producing the appropriate sounds in the appropriate circumstances – speaking – and the behaviour of responding appropriately to the heard sounds – understanding. So, perhaps, language has first evolved in groups of genetically related human beings.

Of course, the problem is more complex because language is more complex and it is used in a variety of different ways. For our robots speaking is an altruistic behaviour which benefits the hearer but not the speaker but language has many other uses and, in many of these uses, the speaker speaks for its own benefit – for example, the speaker asks or commands the hearer to do something which benefits the speaker. Or language may benefit both the speaker and the hearer.

But the real problem is that, unlike the language of our robots, the language of human beings is not innate and encoded in their genes but it is learned in the early stages of their life. What is encoded in the genes of human beings – and not in the genes of nonhuman animals – is only the capacity to learn a language but the specific language which they speak is learned during the first years of their life by imitating the language of those who already speak the language – initially their parents and then other people. The economist and psychologist Herbert Simon has proposed that human beings are born with a genetically inherited tendency to learn from others, and this hypothesis makes sense since, as we have said in Chapter 2, most human behaviours are learned from others. (On learning from others, see Chapter 8 dedicated

to robots that have culture.) To test Simon's hypothesis, we construct a new type of robots which are born with random connections weights in the linguistic module of their network and which, therefore, at birth do neither speak or understand language. These robots learn language by imitating the language of the robots of the preceding generation. The life of a robot is divided up into two periods: infancy and adulthood. During infancy a robot has the only problem to learn language because its parents feed the infant robot with the food tokens they find in the environment and keep them distant from the poisonous tokens. But, when it becomes an adult, the robot must autonomously look for the food tokens and avoid the poisonous tokens. We make the robots learn language by using the backpropagation algorithm. In each cycle, an adult robot and an infant robot see the same token and they both produce a sound. The infant robot compares its sound with the sound produced by the adult robot – the "teaching input" – and the connection weights of its neural network gradually change so that at the end the infant robot has learned to speak the language spoken by the adult robot. The same happens for learning to understand language. The adult robot and the infant robot see the same token and they hear the same sound – produced by a virtual robot – and, while the adult robot responds to the sound "food" by eating the token and to the sound "poison" by avoiding the token, the infant responds to both sounds randomly. But the infant robot compares how it responds to the tokens with how the adult robot responds to the same tokens, the connection weights of its neural network gradually change, and the infant robot learns to respond to the two sounds in the same way as the adult robot. The infant robot has learned to understand language.

Why the infant robot imitates the behaviour of the adult robot? Following Simon's hypothesis, we endow each robot with a "docility gene", where "docility" is a genetically inherited tendency to learn from others. The "docility gene" has a value that can vary from 0 to 1, where a higher value of the gene means a greater tendency to learn from others and a lower value a more limited tendency to learn from others. The robots of the initial generation have a random value for their "docility gene", which means that the average value of the gene is 0.5. Many robots do not learn language and, as a consequence, they find it difficult to distinguish between the food tokens and the poisonous tokens. The robots reproduce on the basis of their capacity to eat the food tokens and avoid the poisonous tokens and not on the basis of the quality of their language but, since a good language helps them to distinguish the food tokens from the poisonous tokens, there is an evolutionary pressure to learn language during infancy. In fact, after a certain number of generations we find that the average value of the "docility gene" has almost reached its maximum level of 1, and this means that the robots have genetically evolved a tendency to learn language during their infancy.

This may also explain why human languages are different and why they change culturally, not genetically. Human beings have a genetically inherited capacity to learn a language but the specific language that they learn is the language spoken in their community. And this language can change even from one generation to the next. As we

will see in Chapter 8 dedicated to robots that have culture, what is learned from others changes because some behaviours are more likely to be copied than other behaviours and because new behaviours are invented or copied from other cultures. Language is one behaviour among others and the cultural transmission of language explains why languages change and why they become progressively different from other languages if the communities in which they are spoken become separate and cease to interact together. (On this, see the simulation of the evolution of Indo-European languages in Chapter 8.)

6. Asymmetries between language production and language understanding

Human beings both speak and understand language but speaking and understanding language are not symmetrical capacities. Children learn to understand words before they learn to produce those same words. Old people have problems remembering proper names but they do not have problems understanding these same names when they hear them. Another phenomenon which shows that speaking and understanding are not symmetrical capacities is the "tip-of-the-tongue" phenomenon. People may not find the word they want to say but they immediately recognize the word they wanted to say as soon as they hear it. Can we construct robots that reproduce these asymmetries between speaking and understanding language?

The robots described in the preceding Section have a neural network with the same number of internal neurons in its nonlinguistic module and in its linguistic module but the brain of human beings is different because more of their brain is dedicated to nonlinguistic behaviour than to language. Our next robots reflect this asymmetry because they have a nonlinguistic module which has more internal neurons than the linguistic module (see Figure 17).

What are the consequences of this asymmetry? The environment of the new robots contains many different types of tokens and, therefore, many different linguistic sounds. The language of the robots does not evolve in a succession of generations but, more realistically, the language already exists and a robot learns the language in the early stages of its life. To make a robot learn the language, we use a new type of learning algorithm which is called the Hebbian algorithm. The Hebbian algorithm requires that the internal neurons of a module are all reciprocally connected together so that their activation level depends not only on what arrives from the sensory neurons and from the internal neurons of the other module but also on what arrives from the other internal neurons of the same module. Learning happens in the following way. In each cycle the value of the weight of the reciprocal connections between two internal neurons increases if the two neurons have similar activation levels and it decreases if they have different activation levels. After a certain of number of cycles the neural network of the robots reaches a stable state. There are no more changes in the values of the connection

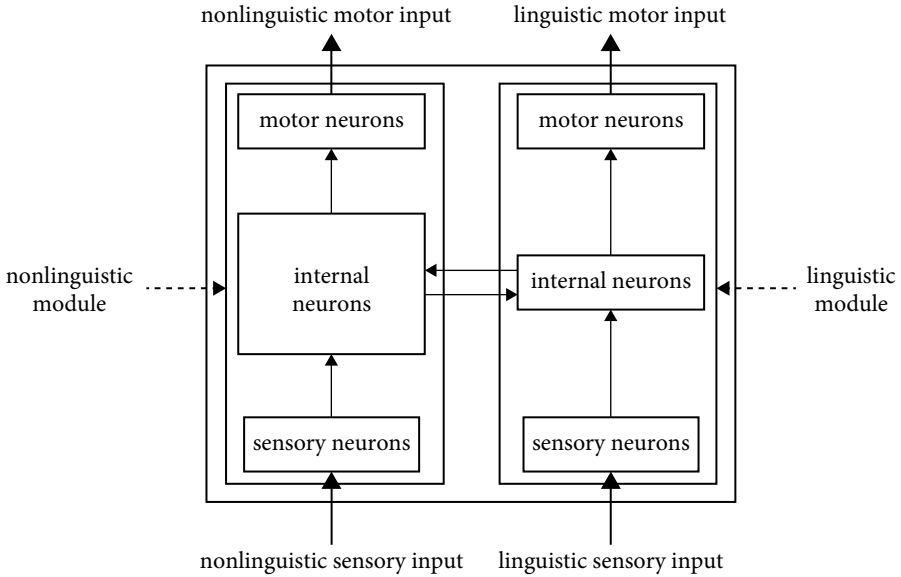


Figure 17. A neural network in which the nonlinguistic module has more internal neurons than the linguistic module.

weights and we find that the robots have learned the language. When a robot hears a sound, the sound evokes the appropriate internal representation of the corresponding token in its nonlinguistic module – the robot understands the sound – and, when the robot sees a token, the sight of the token evokes the internal representation of the appropriate sound in its linguistic module – the robot speaks.

But, although the robots have learned both to speak and to understand the language, they are better at understanding the linguistic sounds than at producing them. This is due to the fact that the nonlinguistic module has more internal neurons than the linguistic module (see Figure 18). Given the reciprocal connections between the internal neurons of the same module, it is sufficient that a heard sound evokes *one part* of the internal representation of a token in the nonlinguistic module for the entire internal representation of the token to appear in the nonlinguistic module – the robot understands the sound. On the contrary, the linguistic module has fewer internal neurons that can reciprocally activate each other, and, therefore, finding the appropriate sound that must be produced to name a specific token is more difficult. The robots understand a word but they are less good at producing the word.

We also find that the asymmetry between language production and language understanding is greater if there are fewer connections *between* the nonlinguistic module and the linguistic module, and this may explain the phenomena mentioned at the beginning of this section (see Figure 18).

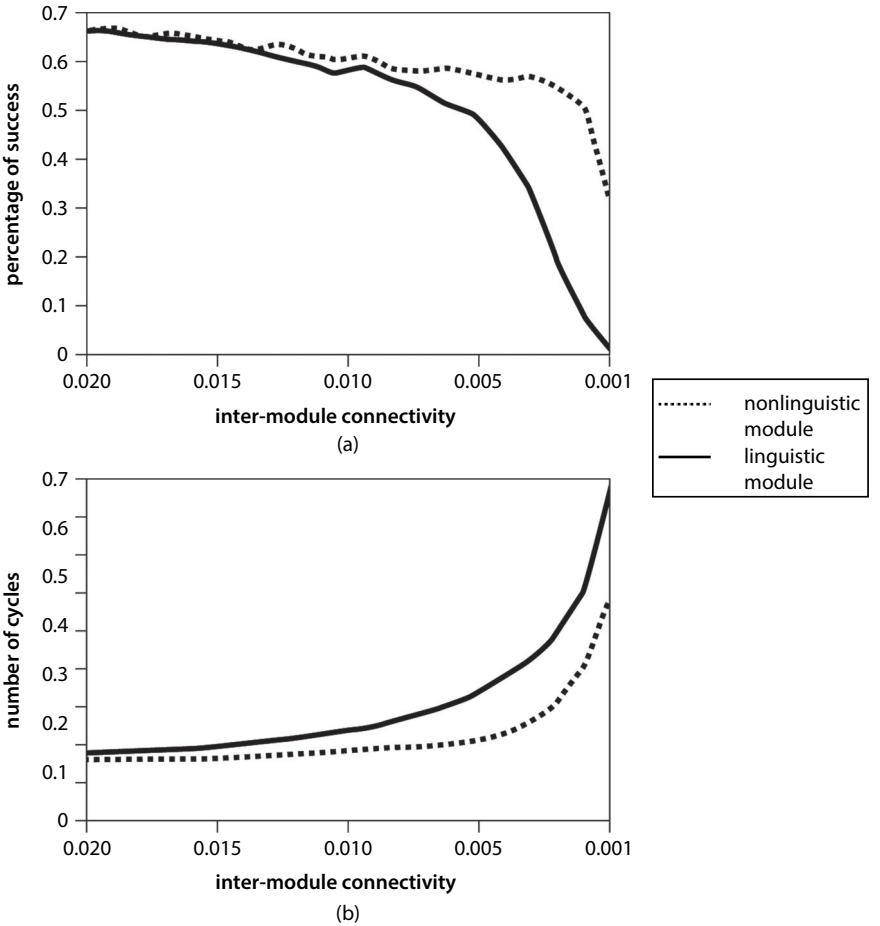


Figure 18. If there is less connectivity between the nonlinguistic module and the linguistic module, the percentage of success decreases (a) while the number of cycles needed to reach a stable state increases (a).

If at birth there are fewer connections between the two modules and then their number increases during the period in which the child is learning language, this may explain why children learn to understand language before they learn to speak. If in old age some of the brain connections cease to function properly, there may be fewer functional connections between the nonlinguistic module and the linguistic module and, since this is a problem for language production more than for language understanding, the robots may explain why old people have more difficulty producing proper names than understanding proper names. And the asymmetry between the two modules may also explain the "tip-of-the tongue" phenomenon.

7. Robots that count and measure

Another cognitive advantage of having language is that language allows human beings to count and measure, and being able to count and measure makes the model of the world they have in their brain different from – and more powerful than – the model of the world which exists in the brain of nonhuman animals. We have not actually constructed robots that count and measure but we will outline how to construct them.

A robot that counts sees a food token and it responds by producing the sound "one". The robot sees two food tokens and it responds by producing the sound "two". The robot sees three food tokens and it responds by producing the noun "three", and so on. The same happens if, instead of food tokens, the robot sees poisonous tokens or predators or any other thing. The sounds co-vary with the number of what is counted, not with the specific things which are counted.

A robot is not born with the capacity to count and it acquires this capacity in two successive stages. In the first stage, it simply learns to produce a sequence of sounds by imitating the sounds produced by other robots that already know the sequence of sounds. In this stage the robot is not counting anything but it is simply learning to produce a fixed sequence of arbitrary sounds. In one cycle the robot produces the sound "one". In the next cycle it hears the sound "one" which it has produced in the preceding cycle and its neural network responds by producing the sound "two". The robot hears the sound "two" and its neural network produces the sound "three", and so on (see Figure 19).

It is in the second stage that the robot learns to count. If the tokens, or whatever must be counted, are presented to the robot one after the other, counting the tokens is relatively easy. The robot learns to produce each sound of the sequence not only in response to the preceding sound but also in response to the token which is currently seeing (see Figure 20).

If the tokens to be counted are presented all together to the robot, the robot's task is more difficult because the robot must also learn to sequentialize the tokens by producing attentional actions on the tokens (Figure 21).

Learning to count a group of tokens can be divided into two sub-stages. In the first sub-stage the attentional actions are actions which physically involve the tokens. The robot grasps with its hand the token which it has just counted and it moves the token to a separate place. In the second sub-stage the attentional actions do not involve the tokens but only the robot. The robot moves its eyes from one token to the next or it points with its finger to the next token to be counted. This is a more difficult task because the robots must also learn to remember which token has already been counted so as to know when all the tokens have been counted, and this is relatively easy if the number of tokens is small but it is much more difficult if the tokens are more numerous. If the tokens are numerous but they are disposed with some regularity in space, for example, in successive rows, the robot learns to exploit these regularities, for example

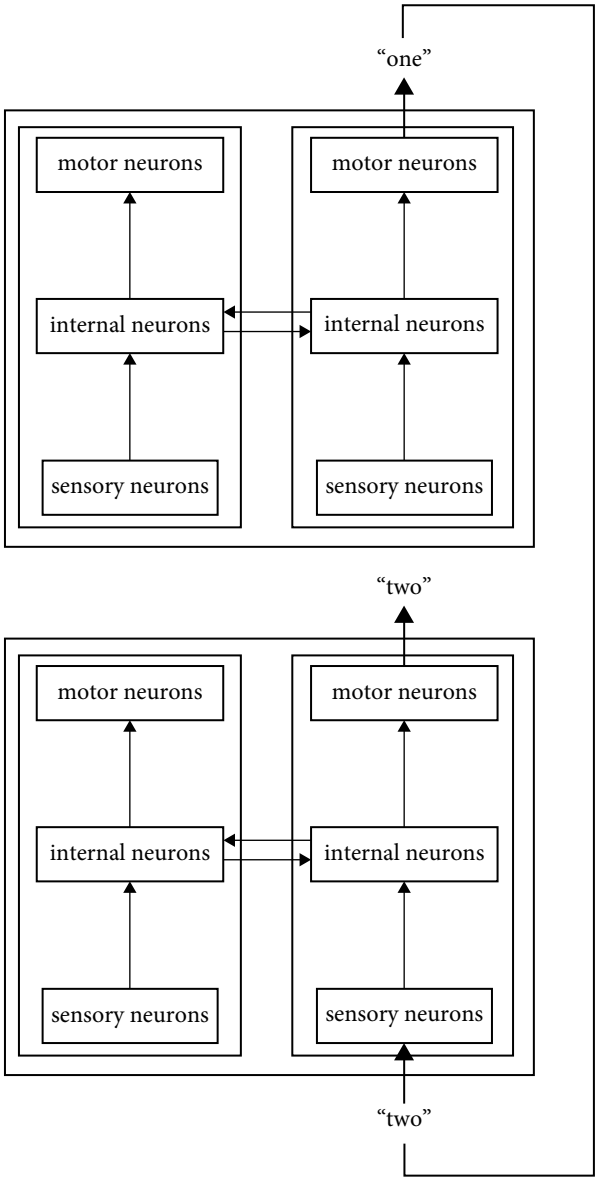


Figure 19. Learning the succession of count words. The robot produces the sound "one", then it hears the sound "one" it has just produced, and it responds by producing the sound "two", etc.

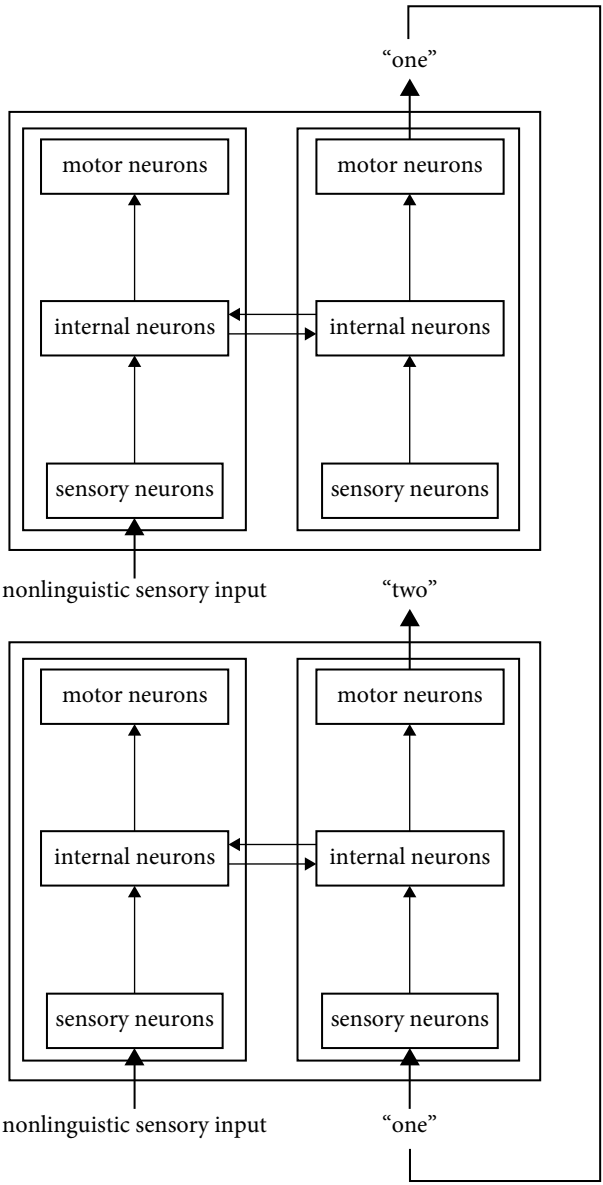


Figure 20. Learning to count a sequence of tokens. The robot produces the sound "one" in response to one token, then it produces the sound "two" in response to both the sound "one" it has just produced and the next token in the sequence, and so on.

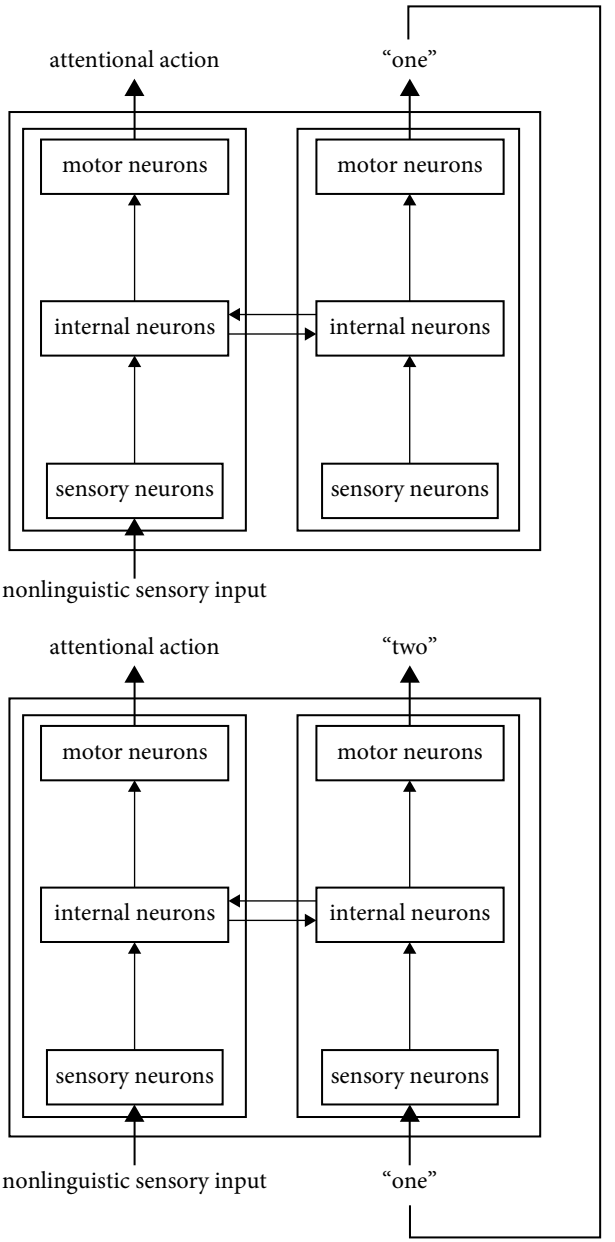


Figure 21. Learning to count a group of tokens. The robot responds to one token with an attentional action on the token and with the sound "one". The robot hears this sound and it responds with an attentional action on another token and with the sound "two", and so on.

by first counting the tokens of the first row, then the tokens of the second row, and so until there are no more rows. But if the tokens are very numerous and are disposed randomly in space, for most robots the task is impossible. What the robot can do is only produce a sound which approximates the number of tokens.

The number sounds of our robots are completely arbitrary: "one", "two", "three". In human languages the sounds which designate the number of large sets of objects are not arbitrary but compositional. Like sentences, they are sequences of smaller sounds. Take English. From "one" to "ten" the sounds are arbitrary. But if one has to count more than ten objects, the sounds become compositional – with some exceptions such as "eleven" and, to some extent, "twelve" or "twenty". For example the sound "twenty-three" is composed by the sound "twenty" followed by the sound "three". We predict that a robot will find it easier to learn a long succession of numerical sounds if, after an initial sequence of arbitrary sounds, the succession of sounds becomes compositional. This raises two interesting questions concerning the nature of the robot's neural network. Do robots that learn long sequences of arbitrary sounds need a larger neural network compared to robots that learn sequences of compositional sounds? Does the neural network of robots that learn sequences of compositional sounds become a modular neural network, with a module for units, a module for tens, a module for hundreds, a module for thousands, and so on?

But, as we have said, in addition to producing the number sounds in the appropriate sequence, a robot that counts must possess other capacities such as making attentional actions that isolate something from all the rest of its experience and "create" sets of things so as to know which token is a member of the set and which is not a member of the set – what should be counted and what should not be counted. These capacities are much helped by the possession of language, and this shown by the fact that it is easier to count the number of pens on a table than the number of "things" in a room without assigning a name or a linguistic descriptions to these "things".

Another thing that our robots tell us about counting is that, whatever one counts, one always counts one's actions and, since one's actions happen in time, counting is strictly linked to time. A robot can count seven food tokens or seven poisonous tokens but in both cases it counts the same thing: the number of its actions of looking or pointing with its finger at each of the tokens to be counted. Counting implies the capacity to abstract from the specific properties of what one is counting because one does not count what is out there in the world but one always counts one's actions.

Once we have constructed robots that count, the robots should be able to reproduce the great variety of phenomena that psychologists and neuroscientists study in the experimental laboratory. For example, the robots must explain why, if the things that must be counted are small in number – say, less than seven – experimental subjects can count them even if they see the things to be counted all together and for such a short time that it is impossible for them to execute a temporal sequence of actions on them. On the contrary, if the things to be counted are more numerous, they are

unable to count them and they can provide only an approximate answer. Is this due to their ability to keep the image of a small set of things in their memory so that they can execute the "attentional actions" on the image of the set of objects rather than on the objects themselves? Another example is why some nonhuman animals exhibit behaviours which take into account the number of a small set of things even if they cannot be said to count them. And a final example has to do with neuroscience rather than psychology. To what extent the neural network of a robot that counts resembles the brain of a human being who counts?

Constructing robots that count is a first step towards constructing robots that look at their environment as made up of quantities that can be measured. Measuring implies that the robots learn from other robots what are the different units of measurement (meters, hours, litres, kilos, etc.), apply these units to spatial distances, temporal lengths, liquids, solids, etc., and count the number of units of whatever they are measuring. The next step is to construct robots that do arithmetical operations, for example, additions. The robots' neural network receives two sounds, say "three" and "two", plus the sound "add" as input and it produces the sound "five" as output. The same for subtractions, multiplications, and divisions. Being able to do arithmetical operations is important because arithmetical operations can be applied to measuring, and applying arithmetical operations to measuring explains much of human societies and human history. Counting, measuring, and doing arithmetical operations create a more articulated model of the world in the brain of the robots but they are also communication tools that play a very important role in their social life, for example, in the coordination of their behaviour in space and time and in the exchange of goods. (For robots that exchange goods, see Chapter 12 on robotic economies.) And, of course, counting, measuring, and doing arithmetical operations are the basis of science. (For robots that do science, see Chapter 11.)

8. On the limits of our robots that have language

Human language is much more complex than the language of the robots described in this Chapter. Some words co-vary with private and not with public nonlinguistic experiences, and this is especially true for words linked to the motivational/emotional half of the mind such as "pain", "sad" or "happy", which have an emotional meaning and not only a cognitive meaning. Some words acquire their meaning not because they co-vary with the hearer's nonlinguistic experiences but because they co-vary with other words which already have a meaning for the hearer. This is what happens when a new word is defined by using other words but it is a more general phenomenon because all words have a meaning which depends on the other words with which they tend to be used. Many words do not have only one meaning but they have many meanings which can be related or unrelated to each other, and the specific meaning which they

evoke in the mind of the hearer depends on the verbal and non-verbal context in which they are used. Our robots can be said to have nouns, verbs, adjectives, and adverbs but human languages also have articles, pronouns, prepositions, and conjunctions – and languages different from English may have different classes of words. The robots must be able to distinguish among the sounds "reaches", "is reaching", "will reach", "reached", "has reached", "to reach", they must be able to produce and understand words made of parts (morphology), demonstratives such as "this" and "that", pronouns, including personal pronouns such as "I" and "you", proper and common names, mass and count nouns, concrete and abstract words, singular and plural nouns, and verbal metaphors.

Another important phenomenon linked to language has to do with space and time. All organisms live in space and time but it is not clear if they can be said to "know" space and time. Animals may "know" space and time more than plants because, unlike plants, they move the different parts of their body and they go from one place to another place and because moving one's body and going from one place to another place takes time. But human beings "know" space and time in a different way because they have language. Their "knowledge" of space and time can be understood by constructing robots that produce and understand linguistic sounds such as "under", "above", "on the left", "on the right", "inside", "outside", or sounds like "now", "before", "after", "in the meantime". And space and time become more articulated in the head of the robots if the robots know how to measure space and time by developing temporal and spatial units of measurement. (For robots that count and measure, see the preceding section.)

Another, very serious, limitation of our robots is that hearing a linguistic sound causes a static entity in their brain: the pattern of activation of the internal neurons of the robots' nonlinguistic module. This corresponds to the idea that the meaning of a word is a well-defined and static entity. But the meaning of a word is not a well-defined and static entity, as linguists and psychologists tend to think. Hearing a word should cause not a static pattern of activation in the nonlinguistic module of our robots but a succession of patterns of activation, and this succession of patterns of activation should be different in different conditions in which the word is produced or understood and it should not always have the same end. Furthermore, unlike the sounds of our robots, words have a meaning which has both cognitive and emotional components because a word activates not only the regular (cognitive) internal neurons but also the emotional neurons that we have discussed in Chapter 2. (This is why human beings write and read poems and novels.)

The idea that the meaning of a word is a well-defined and static entity has an important implication. Language allows human beings to have a more articulated model of the world in their brain but it also causes their mental world to be populated by distinct entities with clear-cut boundaries – which may not be true for the mental world of nonhuman animals. The nonlinguistic experience of seeing a ball while hearing the sound "ball" does not only isolate the ball from the rest of the nonlinguistic experience but it also makes the ball an entity with clear-cut boundaries and with an identity

which is different from the identity of all other entities. Philosophers think that the world is populated by "essences". But the world is not populated by "essences" and it is only language which makes us think that the world is populated by "essences". The meaning of a word is not an "essence" or a "concept" but, as the Austrian philosopher Ludwig Wittgenstein thought, it is the use that we make of the word – and, we add, what happens in our brain when we hear or produce the word.

We have already mentioned another limitation of our robots. Unlike the language of most of the robots described in this Chapter, human language is not encoded in the human genes but it is learned in the early stages of life. The *capacity to learn* a human language is encoded in the human genes – and not in the genes of other animals – but human beings *learn* their specific language by imitating the language which is spoken by the people around them. Language learning is a sequence of steps: (1) infants spontaneously produce language-like sounds, (2) they imitate the sounds that they themselves produce (babbling); (3) they imitate the sounds produced by the people around them; (4) they incorporate in their brain the co-variations of specific sounds with specific aspects of their nonlinguistic experience; (5) they learn to put words together and they develop syntax. The construction of robots that learn language by going through this sequence of steps is an important task for the future.

Human beings learn language but, as we have said, the capacity to learn a language is encoded in their genes. Therefore, language is one important way in which evolution and learning interact together. (See the preceding chapter.) In this Chapter we have described robots that either evolve or learn language but it is not clear how a population of animals without language became a population of animals with language. The "words" of our robots are sounds arbitrarily linked to their meanings – and this is typical of human language. The sound "food" which is used to refer to food in English might be any other sound and, in fact, it is another sound, "cibo", in Italian. Human beings also communicate by using manual gestures which are not – or were initially not – arbitrarily linked to their meaning, and this poses the problem if human language first evolved as a gestural language and then it became an arbitrary verbal language or it was an arbitrary verbal language from its beginning. This is a problem which is much debated among paleoanthropologists, linguists, and psychologists, and robots should help us to better understand the problem and to find the correct answer.

But the most important property of human language which is absent from the language of our robots is syntax. Sequences of words have a meaning which depends on the meanings of the component words but the meanings of the component words must be put together according to what linguists call "syntactic rules" – although there are sequences of words, called idioms, whose meaning cannot be derived from the meanings of the components words plus the rules of syntax. And syntactic rules are recursive. A sequence of words can take the place of one single word in a sentence – and this makes the number of sequences of words that human beings can produce and understand theoretically infinite.

A final limitation of the robots described in this Chapter is that they reproduce some very basic phenomena studied by linguists and psycholinguists but they say nothing about other aspects of language which are studied by other disciplines. The brain of our robots which have language is extremely simple compared to the actual human brain, and future human robots should have a brain more similar to the human brain especially for the parts of the brain which are dedicated to language (neurolinguistics). We should also be able to reproduce the different pathologies of language, for example, the different forms of aphasia, by appropriately lesioning the different parts of the robots' neural network. Future robots should also reproduce the different uses of language – to inform, to ask for information, to comment, to convince, to discuss, to make plans, to coordinate the behaviour of many individuals (linguistic pragmatics). We must construct communities of robots which have language and we must examine the role of language in their social interactions and in the life of their communities (social psychology and sociolinguistics). We must construct robots that have a written language and we must study the role of written language in the political, economic, and juridical organization of their societies (political, economic, and juridical disciplines). Language is learned by imitating other people who already speak the language but with the constant addition of new words, often imported from other languages, changes in the meaning of existing words and in the rules of syntax, and the emergence of new languages. Therefore, the language of our robots should change in a succession of generations and should be different in different communities of robots (historical and comparative linguistics). And, finally, people who speak different languages may have a different model of the world in their mind – a hypothesis which is especially important given the cognitive consequences of language which we have explored in this Chapter – and we should test this hypothesis by constructing robots which speak different languages (anthropological linguistics).

In this Chapter we have been interested in only one aspect of language – the cognitive consequences of having language – and we hope we have shown that human beings have language not because their mind is so complex but that the human mind is so complex because human beings have language. But robots that have language are, to a large extent, future robots.

CHAPTER 5

Robots with a mental life

Another thing which strikes ME about human beings is that, unlike nonhuman animals, human beings live in two worlds, a real world and a mental world. The sensory inputs which arrive to their brain from outside their brain are the real world. The mental world is the sensory inputs that are self-generated by their brain. Their brain self-generates images of both existing and non-existing things, memories of past events, and predictions of future events. And they react to these self-generated sensory inputs by self-generating other sensory inputs, so that they can become lost in their mental world. All aspects of mental life are important but predicting future events is especially important because, if their brain is able to predict the consequences of their actions before they execute the actions, human beings can evaluate these consequences and, on the basis of theis evaluation, decide whether or not to actually execute an action. In this sense ME agrees that, unlike nonhuman animals, human beings have "free will" and they may be said to "want" to do what they do.

ME knows that human beings have a mental life because they talk about their mental life, about what they remember, imagine, and predict. But language is important for mental life because human beings' mental world is often a linguistic mental world. Human beings use language not only to talk to other human beings but also to talk to themselves. The words that they hear are not produced by someone else but they are self-generated by their brain. And since language makes the model of the world which they have in their brain more articulated and sophisticated, talking to oneself also makes their mental world more articulated and sophisticated. They describe to themselves what they see, they tell to themselves what has happened in the past and what they predict will happen in the future, and they put into words the consequences of what they plan to do so that they can better evaluate these consequences and decide if to do or not to do what they would like to do. Another consequence of talking to oneself is that human beings do with themselves the same things that they do when they talk with other human beings. They comment with themselves their real and mental life, they discuss with themselves, they generate new ideas starting from already existing ideas, and they try to convince themselves to do or not to do this or that thing. The advantages of talking to oneself are so great that ME believes that talking to oneself is a very ancient human behaviour and that it may have been one of the causes of the emergence of language.

1. Mental life as the self-generation of sensory inputs

If we are interested in what makes human beings different from nonhuman animals, an inevitable candidate is the possession of a mental life. Nonhuman animals have only behaviours. Human beings have both behaviours and a mental life. And, given the importance of mental life for human beings, the construction of robots that have a mental life is a crucial requirement for calling our robots human.

But what is mental life? The expression "mental life" may be used to refer to a variety of elusive and ill-defined phenomena and robots should help us to define and operationalize the meaning of this expression. This is our robotic definition of mental life. Most of the sensory inputs which arrive to the brain originate in the external environment or inside the body. Something happens in the external environment or inside the body which causes the arrival of sensory input to the brain and the brain responds to these sensory inputs with some behaviour. Mental life is the self-generation of sensory inputs by the brain. It is the brain itself that produces its sensory inputs, and the brain responds to these self-generated sensory inputs.

How is the neural network of a robot which has mental life? The neural network of our robots has sensory neurons, internal neurons, and motor neurons. An event outside the robot's neural network causes a pattern of activation in the sensory neurons, this pattern of activation causes a pattern of activation in the internal neurons which, in turn, causes a pattern of activation in the motor neurons. This is the neural network of a robot which has behaviours but does not have a mental life. The neural network of a robot which has mental life has an additional set of internal neurons and the pattern of activation of these internal neurons is like the pattern of activation of the sensory neurons which might be caused by an event outside the robot's neural network but is not caused by an event outside the robot's neural network. The pattern of activation is self-generated by the neural network (see Figure 1). It is what the robot imagines, remembers, thinks, predicts, plans, hallucinates, or dreams. The self-generation of sensory inputs inside the brain must have some adaptive value, and human robots should help us to discover this adaptive value.

How do we know that human beings have a mental life? We know that human beings have a mental life because they talk about their mental life. So what we must do is construct robots that talk about their images, memories, thoughts, predictions, plans, hallucinations, and dreams. And human beings do not only self-generate sensory inputs inside their brain and have an "internal" mental life but they also have what can be called an "external" mental life. They create all sorts of physical artefacts, from written documents to artistic and symbolic artefacts, which send sensory inputs not only to the brain of other human beings but also to their brain. By studying (reproducing) these artefacts we can know that human beings have a mental life and what is this mental life.

In this Chapter we describe robots that have some very simple forms of mental life and we try to understand the adaptive value of having a mental life. We begin with

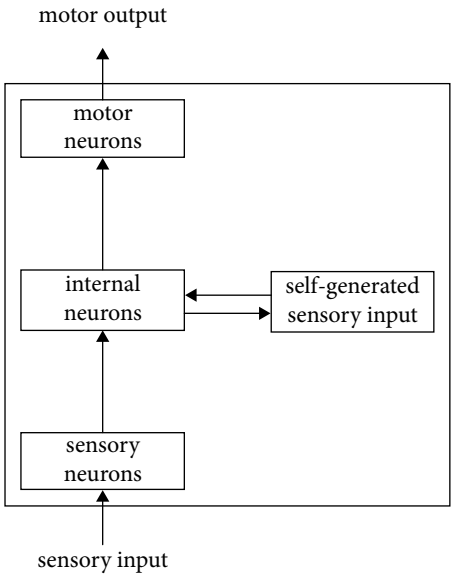


Figure 1. Neural network of a robot whose brain self-generates its own sensory input.

robots that have mental images but most of the Chapter is dedicated to robots that predict future sensory inputs. In the last section of the chapter we discuss the mental life of robots that have language because most of human mental life is a linguistic mental life: human beings talk to themselves.

2. Mental images

Mental images are the most typical manifestation of mental life. The robot "sees" or "hears" something but what in the external environment normally causes the seeing or hearing is absent. The sensory inputs are self-generated inside the robot's brain. If the mental image corresponds to some sensory inputs that the robot has experienced in the past, having the mental image is remembering. If the robot's brain self-generates mental images which are not like anything which the robot has experienced in the past or which put together in novel ways different parts of past sensory inputs, they are imaginations and inventions.

The adaptive value of remembering, imagining, and inventing can be demonstrated by constructing robots that evolve the capacity to remember, imagine, and invent because doing these things increases their fitness and their well-being. For example, mental images can be "search images" that help the robots when they are looking for something because explicitly knowing what they are looking for makes

their search behaviour more effective. (Some nonhuman animals are also said to have "search images".) Mental images as memories can be adaptive because they allow the robots to remember things to do, to avoid past mistakes, and to recreate in their mind past events associated with positive emotional states. Mental images as imaginations can cause a robot to invent useful things. And mental images play a critical role in producing artistic artefacts and in exposing oneself to artistic artefacts. (For robots that have art, see Chapter 13.) In other cases the adaptive value of mental images is less clear. Mental images can occur when a robot is asleep – robotic dreams – and the adaptive value of dreams is unclear. Remembering some past events can cause negative emotional states, and it is not clear why a robot should remember these events. And mental images can be pathological. A robot can confuse a mental image with reality. These would be robotic hallucinations – and hallucinations are maladaptive.

In fact, the adaptive nature of mental images is not always clear and this can be easily seen if we try to construct robots that can be happy or unhappy. Nonhuman animals cannot be happy or unhappy in the sense in which human beings are happy or unhappy. Both animals and human beings can have good and bad experiences (see Chapter 2 on motivations and emotions) but animals cannot be said to be happy or unhappy – although, as always, the difference between human beings and nonhuman animals is of degree. What explains the difference is that human beings have a mental life. Human beings do not only have experiences that make them have positive and negative emotional states but they also *remember* past experiences and they *imagine* future experiences that make them have positive and negative emotional states. Happiness is not just having one good experience and unhappiness is not just having one bad experience. Happiness is having good experiences *most of the time* and unhappiness is having bad experiences *most of the time*. And this is why mental life as remembering and imagining – and, more generally, as the self-generation of sensory inputs – is a pre-condition for being happy or unhappy. Like non-human animals, human beings tend to have only a limited number of *actual* good and bad experiences but they have a mental life so that they may continually remember or imagine good or bad experiences – and this is what makes them happy or unhappy.

If we construct robots that are happy or unhappy, we can reproduce with the robots many interesting phenomena linked to happiness and unhappiness. One is that there are inter-individual differences in the tendency to feel happy or unhappy, and these inter-individual differences may have a genetic basis. Some robots may be happier than other robots not because they have good experiences more often than less happy robots but because they tend to remember or imagine good experiences more often than less happy robots or because they succeed in *not* remembering and *not* imagining bad experiences while unhappy robots are unable to do so. Another phenomenon is the relation between happiness/unhappiness and psychopathology and psychotherapy. Some psychopathologies consist in being unable *not* to remember or imagine bad experiences and, by using different methods, psychotherapists try to limit this tendency.

Bad experiences can be remembered without being conscious of them – where consciousness is talking to oneself (see Section 10) – and many forms of psychotherapy consists in bringing bad experiences to consciousness by talking about them so as to learn not to remember or imagine them. Another form of psychotherapy is meditation. Meditation is abolishing mental life. An individual learns not to remember or imagine anything by concentrating on one single thing which is not good or bad. Meditation helps the individual not to remember or imagine bad experiences and therefore not to be unhappy, and not being unhappy is a form of happiness.

We have not constructed robots that have mental images and, therefore, we cannot reproduce these phenomena with robots. But we have a hypothesis on the neural network of robots that self-generate their sensory inputs, and this is why we think that constructing robots that have mental images is not impossible. In the next section we describe robots which have a neural network that self-generates its sensory inputs and which use this network to predict future sensory inputs – and we have constructed these robots.

3. Robots that predict

A very important form of mental life is the ability to predict future sensory inputs. The sensory input has not arrived yet to the robot's sensory neurons but the robot's brain self-generates a sensory input which is similar to the future sensory input. The neural network of a robot that predicts is made up of two modules: a sensory-motor module and a prediction module. The sensory-motor module is the basic neural network of our robots: sensory neurons, internal neurons, and motor neurons. The prediction module is new. It consists of a set of internal neurons, the prediction neurons, which receive connections from the sensory-motor module and send their connections to the sensory-motor module. If the activation pattern of the prediction neurons is the same as the activation pattern that *will* appear in the sensory neurons in the next input/output cycle – or, more generally, in the future – the robot is able to predict.

There are two types of predictions. A robot can predict a future sensory input which is *independent* from the robot's behaviour. The robot predicts that it will rain given the dark clouds in the sky. The environment changes independently of the robot and these changes produce a succession of sensory inputs for the robot's neural network. If there is some regularity in the succession of sensory inputs, the robot can predict the next sensory input on the basis of the current sensory input. Figure 3 shows the neural network of a robot which does this first type of predictions. The pattern of activation of the prediction neurons is caused by the current sensory input (dark clouds in the sky) and we say that the robot is able to predict if the pattern of activation of the prediction neurons corresponds to the pattern of activation which will appear in the robot's sensory neurons in the next cycle (rain).

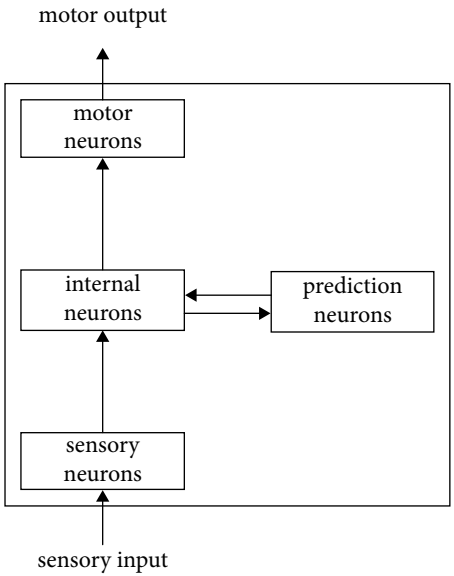


Figure 2. Neural network of robots that predict.

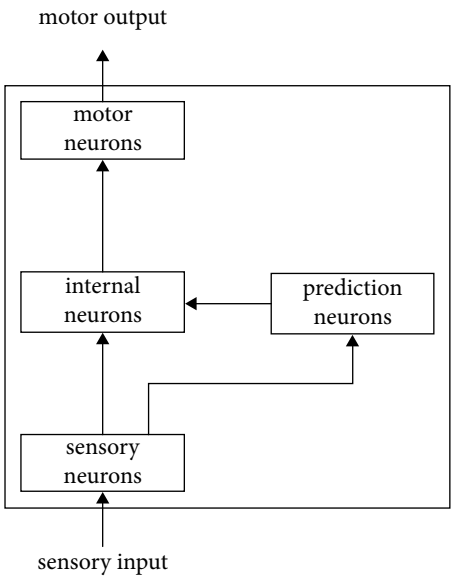


Figure 3. Neural network of a robot that predicts the next sensory input on the only basis of the current sensory input.

The second type of predictions are predictions of future sensory inputs which *depend* on the robot's behaviour. Examples are predicting the sound that the robot will hear if the robot moves its phono-articulatory organs in one specific way or predicting that a glass will break if the robot lets the glass fall to the ground. The robot's behaviour changes the external environment – a sound is produced, the glass breaks – and the next sensory input depends on the changes due to the robot's behaviour. To say that the robots predicts, the prediction neurons must be activated *before* the robot physically does anything. The movements of the robot's body (letting the glass fall to the ground) are encoded in the motor neurons but they are not physically executed because, when the movements are physically executed, it is too late for predicting a sensory input which is already there. For all the robots described so far, when an activation pattern appears in the motor neurons, the corresponding movement of the robot's body is, immediately and automatically, physically executed. The robots that predict the consequences of their behaviours are different. The movement is encoded in the motor neurons but its physical execution is blocked. It is on the basis of this already defined but not physically executed movement that the robot's neural network generates a prediction of the next sensory input by using its prediction neurons. After the prediction has been generated, the robot physically executes the movement, the movement changes the environment, and these changes cause the appearance of a new activation pattern in the sensory neurons. And the predicted sensory pattern and the actual sensory pattern are the same.

The predicted sensory input can be a result only of the robot's behaviour or it can be the result of both the robot's behaviour and the environment. If the robot moves its phono-articulatory organs so that they produce one particular sound, the nature of the sound is entirely determined by the movements of the robot's phono-articulatory organs – aside from wind and other sounds in the environment. But more frequently the next sensory input depends on *both* the robot's behaviour *and* the environment. The robot grasps a glass with its hand and it plans to open its hand. What happens to the glass depends not only on the robot's behaviour but also on the glass. If the robot lets a rubber ball, rather than a glass, fall to the ground, the robot will not predict that the rubber ball will break. The robot's prediction that the glass will break must take into consideration both the planned movement and the properties of the glass.

The neural network of robots that predict future sensory inputs which depend on their behaviour is shown in Figure 4.

Figure 4(a) is the neural network for predicting the next sensory input when the next sensory input depends only on the robot's behaviour – the sounds produced by the robot's phono-articulatory organs. Figure 4(b) is the neural network for predicting the next sensory input when the next sensory input depends on both the robot's behaviour and the environment – what happens to the glass if the robot lets it fall on the ground. In both cases the prediction module is activated by the motor neurons which encode the planned movement but, while in the first case this is sufficient to predict the next

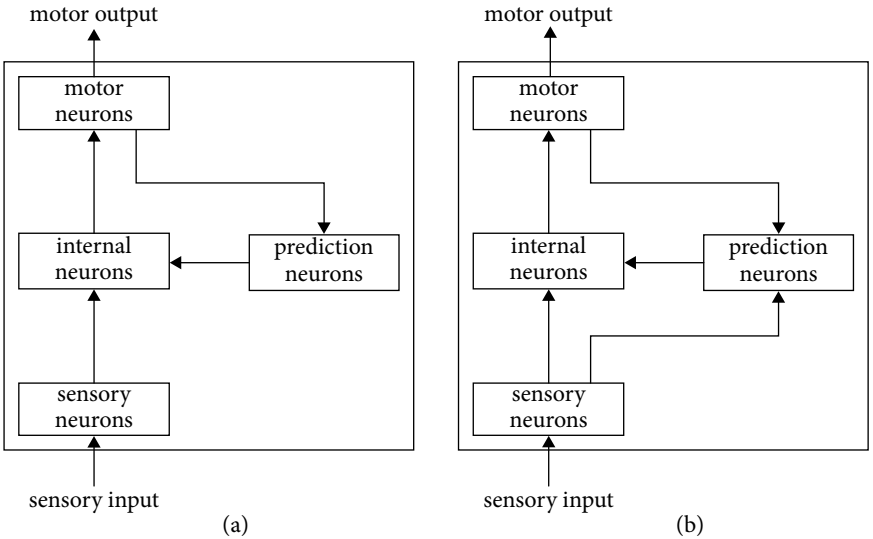


Figure 4. Neural networks for making predictions that depend on the robot's planned behaviour. The prediction depends only on the robot's planned behaviour (a) or on both the planned behaviour and the environment (b).

sensory input, in the second case the prediction module must also receive connections from the sensory neurons (from the glass) in order to predict the next sensory.

The capacity to predict – and, in particular the capacity to predict the consequences of one's behaviour – is typically human, and this may derive from the fact that, unlike non-human animals, human beings learn almost all their behaviours during their life and, therefore, they cannot already know the consequences of their behaviour. Human beings have a genetically evolved capacity to learn to predict the consequences of their actions, and it is probably the evolutionary emergence of the human hands – a cause and a consequence of bipedalism – which has been a critical adaptive pressure for evolving an ability to predict the consequences of one's actions. Nonhuman animals have a very limited repertoire of behaviours and they can produce a small number of effects in the environment with their behaviours. Therefore, for them being able to predict the consequences of their actions is less important than for human beings. With their hands human beings can cause an enormous variety of direct and indirect effects in the environment and they find it very useful to be able to predict these effects in order to decide what to do and what not to do.

This hypothesis can be tested by constructing robots whose genotype includes a "prediction gene" with a value that can go from 0 to 1, where 0 means zero probability to have a prediction module in one's neural network and 1 means 100% probability to have this module. If the robots have an adaptive pattern which is made of a small

repertoire of different behaviours, most of these behaviours would be innate and not learned during life and, therefore, there would be little pressure on the "prediction gene" to increase its value because the consequences of the robots' behaviours are already "known" to the robots' genes and to the robots' brain. This is what happens with nonhuman animals. On the contrary, if the robots have a very large repertoire of behaviours which are mostly learned during their life and whose consequences cannot be already known by their brain, the robots would evolve a value for their "prediction gene" near to 1. This is what has happened to human beings who all have a prediction module in their brain.

We have not actually constructed robots to test this hypothesis but we have constructed robots that already have a prediction module in their neural network and, using this module, they learn to predict the consequences of their behaviours. Their neural network is that of Figure 4(b). The activation pattern which appears in the prediction neurons when a robot is planning some behaviour depends on the weights of the connections that link both the sensory neurons and the motor neurons to the prediction module. When a robot is born, these weights are random, and this means that the robot is unable to predict the consequences of its behaviours. The newborn robot lets a glass fall to the ground but it does not predict that the glass will break – and it may even be surprised that the glass breaks. Then the robot learns. To make the robot learn we use the backpropagation algorithm. The robot's neural network compares the activation pattern of its prediction neurons with the activation pattern that appears in its sensory neurons in the *next* cycle and, based on the results of this comparison, the weights of the connections linking the sensory-motor module to the prediction module change. They change in such a way that the difference between a predicted sensory input and the actual sensory input of the next cycle is gradually reduced and, at the end, the robot is able to predict the sensory inputs that will arrive to its sensory organs from the environment in the next cycle.

4. Predicting and anticipating

Before we discuss the advantages of possessing an ability to predict, it is necessary to make a distinction between predicting and anticipating. If the succession of sensory inputs which arrive from the environment to an animal's brain has some intrinsic regularity – which is true in many cases – the animal's brain changes and it incorporates these regularities in its structure and manner of functioning, with the consequence that the animal responds to a sensory input in ways which take into consideration the next sensory input that will arrive from the environment. This is anticipating the environment, and anticipating the environment does not require the self-generation of future sensory inputs. A robot can anticipate the environment even if its neural network consists only in the basic sensory-motor module. This is what is shown by our next robots.

The robots have an eye that they can move. The environment contains a mobile object and this object moves in front of the robot from left to right. What the robot must do is move its eye so as to keep the object in the centre of its visual field (fovea). The robot's eye movements are controlled by a neural network with visual neurons that encode the current position of the object and motor neurons that encode the movements of the robot's eye. But the robot's eye must not *follow* the movements of the object but they must *anticipate* these movements. When the object is in a certain position, the robot must respond by moving its eye in such a way that the eye fixates the point in space in which the object will be in the *next* input/output cycle. At birth the connection weights of the robot's neural network are random and therefore the robot is unable to anticipate the movements of the object with its eye. Then the connection weights of the robot's network change (the robots learn by using the backpropagation learning algorithm) and, after a certain number of cycles, we find that the robot is able to anticipate the movements of the object with its eye.

In the environment of these robots objects move from left to right always at the same speed. We construct another type of robots which live in an environment in which the objects can move at different speeds and this, clearly, makes the robots' task more difficult. To help the robots to anticipate the position of an object that can move at different speeds we add to their neural network a set of memory neurons that keep a record of the past. In each input/output cycle the activation pattern of the internal neurons is copied in the memory neurons and, in the next cycle, the memory neurons return this information to the internal neurons. In this way, the memory neurons determine the movement of the robots' eyes together with the visual input from the object. The memory neurons capture the speed with which the object moves and the results are that the robots are able to anticipate the movement of the object whatever the speed with which the object moves (within limits). And the robots can do even more. At the end, their eye is able to anticipate the movements of the object even if the object moves at a speed which is different from the speeds experienced during learning and even if the speed of movement of the object is not constant but it increases or decreases.

These robots tell us something interesting about the relation between behaviour and the environment. If the environment changes and if there are regularities in how the environment changes, it is adaptive for both animals and human beings to incorporate these regularities in their brain so as to exploit them in their behaviour. (The environment of plants changes less frequently and more slowly, and this may be a reason why plants – and very simple animals – do not have a brain.) Our moving object is an example of these regularities. The object provides the robot with a succession of sensory inputs – the successive spatial positions occupied by the object as the object moves – and the robot's neural network incorporates in its connection weights the regularities which exist in the succession of sensory inputs and uses these regularities to anticipate with the movement of the eyes the position that the object will occupy in the next cycle. All animals to some extent anticipate the next sensory input from the environment given the preceding inputs and they use this ability to behave more

effectively with respect to the environment. Two examples are anticipating the next position of a moving prey to catch the prey or the next position of a predator to avoid being caught by the predator. And some behaviours of even simple animals appear to be motivated by the acquisition of this capacity to anticipate the environment – a form of intrinsic motivation – rather than by more practical motivations such as eating or avoiding dangers.

But predicting is different from anticipating. Predicting is the ability to *explicitly* self-generate in one's brain the sensory inputs that will arrive from the environment and this requires the possession of a neural network with a prediction module. This is why predicting is an aspect of mental life and why, as we will see in the next sections, being able to predict has so many important consequences. Both animals and human beings anticipate but only human beings predict.

5. Evaluating the predicted consequences of one's actions

A crucial advantage of being able to predict the consequences of a planned action is that one can evaluate these consequences before actually executing the action and, on the basis of this evaluation, decide whether to execute or not execute the action. This is shown by our next robots.

The robots have an arm and a hand and their life is a succession of episodes. At the beginning of each episode they have in their hand a stone which they must throw in order to hit – and automatically kill – a prey and the robots which have more offspring are those which are better able to hit the prey with the stone. The stones can be of different weights and the prey can be at different distances and, therefore, the robots must decide with what force they must throw the stone in each particular occasion. We compare two populations of robots. The robots of one population have a neural network made up only of the sensory-motor module. The robots are informed by the sensory neurons in their arm and hand of the weight of the stone which they have in their hand and by their visual neurons of the distance of the prey and they respond by throwing the stone with a force which is encoded in their motor neurons. The neural network of the robots of the other population is more complex. In addition to the sensory-motor module, it includes a prediction module for predicting where the stone will end up given its weight and given the force with which the robot will throw the stone. But there is more. The prediction module is connected with a success/failure module whose neurons can have one of two activation patterns, a "success" activation pattern or a "failure" activation pattern. The success/failure module is connected with the motor neurons and the throwing action is executed only if the activation pattern of the success/failure module is "success" (see Figure 5). What are the results? The results are that the robots that are able to predict if a planned throwing action will actually result in hitting the prey have more fitness than the robots that have a simpler, purely reactive, neural network which does not allow them to make predictions.

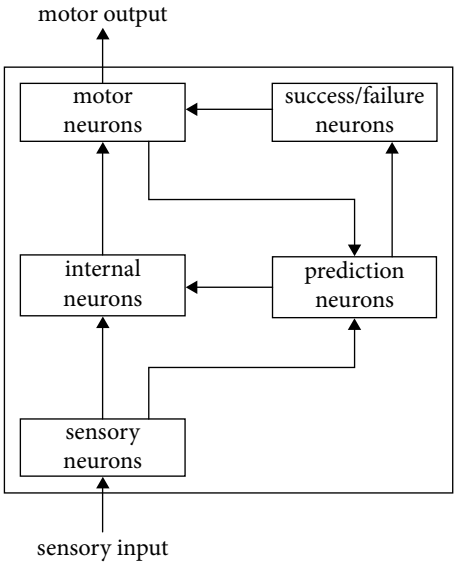


Figure 5. Neural network with a prediction module and a success/failure module which evaluates the predicted consequences of a planned action and lets the sensory/motor module physically execute the action only if the evaluation is "success".

All the weights of the connections of the robots' neural network are genetically evolved and inherited at birth except those of the connections that link the sensory-motor module to the prediction module. At birth these weights are random but then they change during a robot's life in such a way that the robot gradually learns to predict if a planned throwing action will actually result in hitting the prey.

These robots tell us what might the role of infancy in the life history of human beings. Infancy is a protected period during which an individual can learn a number of capacities that will be useful when the individual becomes an adult. To test this idea, we compare two populations of robots which both learn during their life to predict if a planned throwing action will or will not be successful. However, the robots of the two populations have two different life histories. Adult life has the same length in both populations but in one population adult life is preceded by a period of infancy. In the population without infancy, a newborn robot must already be able to hit the prey with stones in order to eat. In the other population there is a period after birth in which the robots are not required to hit a prey in order to eat because during this period they are fed by their (virtual) parents. In both populations the robots learn to make predictions after birth but the robots with infancy have an advantage compared to the robots without infancy. During the initial period of their life in which they are learning to predict if a planned throwing action will be a success or a failure, the robots inevitably make many unsuccessful throwing actions. However, while the robots without infancy

pay for these unsuccessful throwing actions because they eat less, this is not true for the robots with infancy. During infancy the robots do not pay any costs for learning an ability which will be useful when they become adults. When they become adults, the robots with infancy already know to predict if a throwing action will be successful or unsuccessful and, therefore, they can avoid doing unsuccessful throwing actions. This is why the robots that have an infancy have a higher level of fitness that the robots which do not have an infancy.

If this is true, there may be an evolutionary pressure to develop infancy as a "safe" period during which an individual can learn what will be useful to the individual when it becomes an adult. Imagine we add an "infancy gene" to the genotype of our robots and this gene has a value that can go from 0 to 1. If the value of the "infancy gene" is 0, the robot carrying the gene has no infancy and it must immediately be able to hit the prey to eat. If the value of the gene is not zero, the robot has an infancy whose length depends on the value of the inherited "infancy gene", although the value of gene cannot be 1 because a value of 1 would mean an infancy which lasts for the entire life. We have not constructed these robots but we expect that if we start with a population of robots with a randomly assigned value for their "infancy gene", the average value of the gene will increase in a succession of generations because of the advantages of possessing an infancy during which the robots can learn to avoid unsuccessful throwing actions. Human beings have a long infancy and our robots may suggest why.

In fact, human beings do things for two reasons. They do things in order to get practical results (extrinsic motivation) and they do things in order to know what they must do in order to obtain practical results (intrinsic motivation). Infancy is a period of life in which many of the things that human beings do are controlled by an intrinsic motivation to know the consequences of their actions, although many human beings – especially scientists and artists – continue to have this intrinsic motivation for the entire length of their life.

6. Freedom of the will

Being able to predict the consequences of one's actions is linked to the possession of what philosophers call "freedom of the will". Can robots have freedom of the will? Can robots not only do what they do but can they also *want* to do what they do? In some cases human beings are not free because there is some external condition which prevents them from doing what they want to do. They are in prison and they cannot go where they would like to go, they do not have money and they cannot buy what they would like to buy, they would like to do X but they cannot do X because they would be punished for doing X. This type of absence of freedom can be easily reproduced with robots but it does not really address the question of the freedom of the will. (For robots that do not do X because they would be punished for doing X, see Chapter 6 on social

robotics.) Freedom of the will poses a more basic – we might say, philosophical – question: Is freedom of the will compatible with robots? Robots are physical entities and all that takes place inside their brain and in the interactions of their brain with the external and internal environment are processes in which physical causes produce physical effects. This is obvious for physically realized robots but it is also true if the robots are simulated in a computer because simulated robots are simulated physical robots. So, if having freedom of the will requires something different from physical processes in which physical causes produce physical effects, robots cannot have freedom of the will. But human beings are said to have freedom of the will and to do what they *want* to do. If robots cannot have freedom of the will because they are physical entities, how will ever be possible to construct human robots?

If a robot simply responds to some sensory input with some motor output, we are not justified in saying that the robot wants to do what it does. But if the robot's neural network includes a prediction module and the robot predicts the consequences of its behaviour before executing the behaviour, we can construct robots that have freedom of the will and do what they want to do. An activation pattern appears in the motor neurons of the robots' neural network in response to some sensory input but, before the behaviour encoded in the motor neurons is physically executed, the robot's neural network predicts the consequences of this behaviour – and, so far, the robot is like our preceding robots. But the neural network of the new robot includes another module: a good/bad module (see Figure 6). The neurons of the prediction module send their connections to the neurons of the good/bad module and the neurons of the good/bad module can have one of two activation patterns, a "good" activation pattern or a "bad" activation pattern. If the activation pattern is "good", the planned movement is physically executed. If it is "bad", the planned movement is not executed. The weights of the connections to and from the good/bad module evolve in a succession of generations, and this means that, as we have said in Chapter 2, it is evolution which tells the robot if doing something is "good" or bad" – although, to be really like human beings, the robots should also learn during their life if the predicted consequences of their *learned* behaviours are "good" or "bad".

Robots that predict and judge the consequences of their actions before executing the actions may help us to clarify the distinction between *doing* X and *wanting* to do X. Current robots do things but they cannot be said to *want* to do the things that they do. Only if a robot predicts the consequences of an action and judges these consequences as good or bad before executing the action, the robot can be said to want to do what it does – and to be responsible for what it does.

This poses some interesting questions. Do only human beings want to do what they do? Are there nonhuman animals that want to do what they do? Is there a neat separation or a continuum between simply doing X (for example, in worms) and doing X because one wants to do X. To answer these questions we should construct robots with different neural networks and different adaptive patterns and see how they behave and what is their life history.

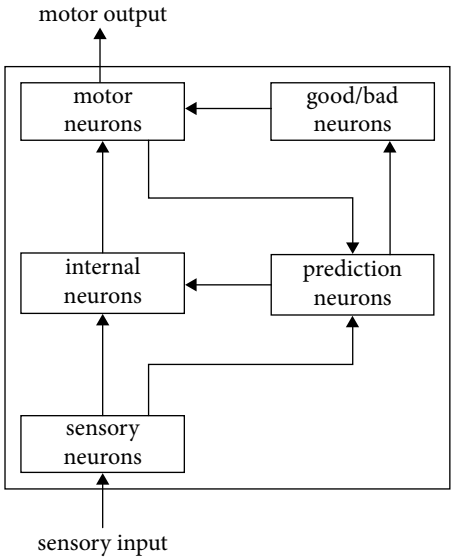


Figure 6. The neural network of a robot that predicts the consequences of its actions and it judges if these consequences are good or bad.

In any case, human beings are more complex than nonhuman animals. They predict the consequences of their actions (second type of predictions), they predict the consequences of these consequences (first type of predictions) and so on. A robot plans to let a glass fall on the ground and it predicts the consequences of its behaviour: the glass will break. If the good/bad module of the robot's neural network judges the broken glass as "good", the robot is ready to let the glass fall to the ground. But the robot does not stop here. The robot considers the broken glass as an already existing sensory input and it predicts the consequences of this (predicted) sensory input: another robot will blame the robot for having broken the glass. These consequences are judged as "bad" by the robot's good/bad module and the robot renounces to let the glass fall to the ground.

Human beings do not only predict the consequences of the consequences of their actions but they also make plans which consist in a succession of actions. A robot's motor neurons encode an action A and its prediction neurons predict the consequences of action A. Then, *before* executing action A, the robot's motor neurons encode another action B which takes into considerations the predicted consequences of action A, and the prediction neurons predict the consequences of action B, and so on until the robot is able to judge if the final result of what it plans to do is good or bad. (When human beings make plans, they do something more: they put into words both their actions and the consequences of their actions. On this, see Section 9.)

7. Predicted sensory inputs replace missing sensory input

Another advantage of the ability to predict future sensory inputs is that predicted sensory inputs can replace missing sensory inputs. Sensory inputs from the environment can be missing for a variety of reasons. For example, a robot is approaching a food token when an extraneous object interposes itself between the robot and the food token so that the robot ceases to see the food token. In these circumstances the predictive monitoring of the environment allows the robot to respond to the *predicted* sensory input from the food token rather than to the *actual* but irrelevant sensory input from the occluding object and the robot can maintain the effectiveness of its behaviour. This is shown by our next robots.

The robots live in an environment with food tokens which they must eat to remain alive and have offspring. The robots' visual neurons encode the location of the nearest food token and the robots respond by approaching and eating the food token. These are our basic robots and their environment contains food tokens and nothing else. Now we evolve another population of robots which also must eat the food tokens but these robots have a problem: once in a while an object interposes itself between the robot and the food token which the robot is trying to reach and this object prevents the robot from seeing the food token. Clearly, this is a more difficult environment because, when the sight of a food token is blocked by an occluding object, the robot does not know what to do. In fact, if we compare the robots which live in the environment with the occluding objects and the robots which live in the environment without the occluding objects, we find that the first type of robots eat less than the second type of robots (see Figure 7).

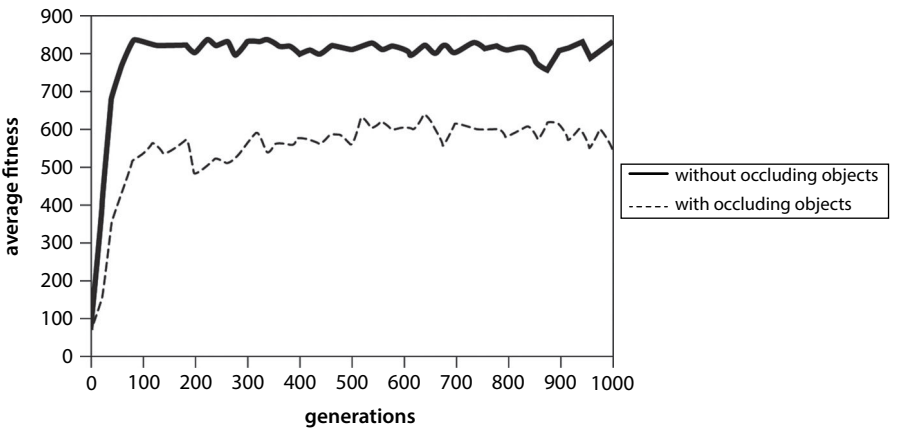


Figure 7. Average number of food tokens eaten by robots living in an environment without occluding objects and by robots living in an environment with occluding objects that can prevent the robots from seeing the food tokens.

For the robots that live in the environment with the occluding objects the ability to make predictions would be very advantageous. If the robots can predict what would be the visual input from a food token when they do not see the food token because of the occluding object, they can respond to the predicted visual input from the food token rather than to the irrelevant visual input from the occluding object. We evolve another population of robots in an environment that contains occluding objects but now the neural network of these robots has a prediction module. Using this module the robots learn to predict the visual input that will arrive from a food token given the current visual input from the food token and the movement with which they plan to respond to the visual input. When a robot is born, the robot is unable to make these predictions but then the robot rapidly learns. In the early phases of the robot's life the difference between predicted visual inputs and actual visual inputs is progressively reduced, and this means that in most of its life the robot knows how to predict the next visual input that will arrive from a food token given a planned and then executed movement of the robot. This capacity to predict is very useful to these robots. When the visual input from food is missing because of an occluding object, the robots respond to the predicted visual input rather than to the actual visual input, and the occluding object does not disrupt the robot's behaviour. These robots eat the same number of food tokens as the robots living in the environment without occluding objects (see Figure 8). This is another demonstration of the adaptive value of the ability to make predictions.

These robots are hardwired by us not to respond to the actual visual input when the visual input is not from food but from the occluding object and to respond instead to the *predicted* visual input from food. Can the robots be made more autonomous?

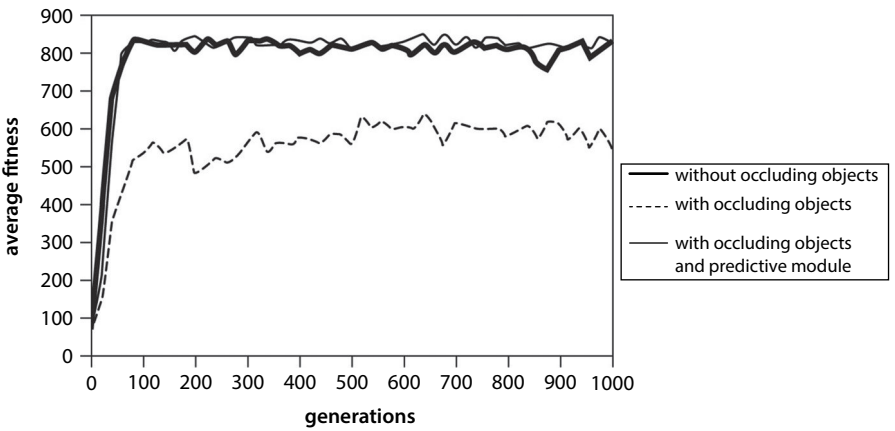


Figure 8. Average number of food tokens eaten by robots living in an environment without occluding objects, with occluding objects and no capacity to predict, and with occluding objects and a capacity to predict the position of an occluded food token.

Can they themselves judge when the current visual input is from an occluding object rather than from food and, therefore, they should respond to the predicted visual input rather than to the actual visual input?

The neural network of these more autonomous robots has two additional modules: a same/different module and an implementation module (see Figure 9).

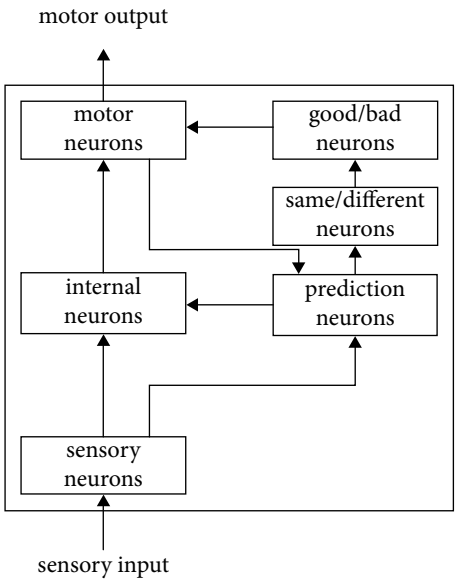


Figure 9. Neural network of robots that judge if the predicted input is "same" or "different" with respect to the actual input and respond to the actual input if the judgment is "same" and to the predicted input if the judgment is "different".

The neurons of the same/different module receive connections from both the visual neurons and the prediction neurons and they can have one of two activation patterns, a "same" activation pattern and a "different" activation pattern. They have the "same" activation pattern when the actual visual input and the predicted visual input are the same, and this means that there is no occluding object and the robot is seeing a food token. They have the "different" activation pattern when the actual visual input and the predicted visual input are different, and this means that there is an object which occludes the sight of the food token. This is the same/different module. Then there is the implementation module. The neurons of the implementation module are activated by the neurons of the same/different module and they cause the neural network to respond to the actual visual input if the activation pattern of the same/different module is "same" and to respond to the predicted visual input if the activation pattern is "different". In this way, when the sight of a food token is occluded by an object, the

robot autonomously responds to the *predicted* visual input from the food token rather than to the *actual* visual input from the occluding object – and this, as we have seen, is advantageous for the robot. The weights of the connections that link the evaluation module, the implementation module, and the sensory-motor module evolve in a succession of generations and they are genetically inherited together with the connections weights of the sensory-motor module. This means that the robots learn during their life to predict the consequences of their behaviour but they are born with a capacity to judge if the actual consequences are "same" or "different" with respect to the predicted consequences and to use this judgment to decide whether to respond to the actual visual input or to the predicted visual input.

The fitness of these more autonomous robots is shown in Figure 10. The robots eat somewhat less food than the robots which are hardwired by us to respond to the predicted visual input rather than to the actual visual input but they eat significantly more than the robots that do not make predictions.

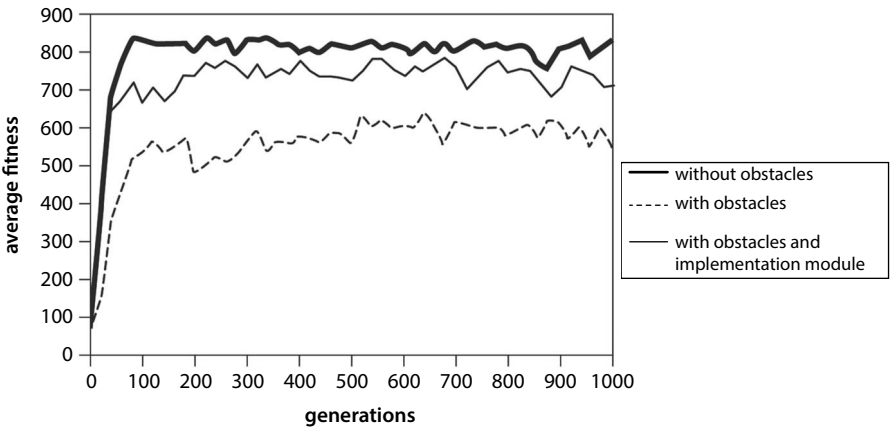


Figure 10. Average fitness of robots that judge if the actual sensory input from the environment is "same" or "different" with respect to the predicted sensory input and use this judgment to decide whether to respond to the actual input or to the predicted input.

These robots pose an interesting problem. If a robot is able to predict the next sensory input which will arrive from the environment, why should the robot ever respond to the *actual* sensory inputs from the environment instead of responding to the *predicted* sensory inputs? The ability to predict the next sensory input from the environment makes the robots more autonomous from the environment – and this autonomy is typical of mental life. But the autonomy from the environment – and mental life – may go too far. A robot which can predict the next sensory input given the current sensory input could take into consideration only the very first sensory

input from the environment and ignore all subsequent sensory inputs. The robot plans a behaviour in response to the first sensory input and it predicts the results of this planned behaviour. Then it does not execute the planned behaviour but it plans another behaviour in response to the predicted sensory of the first planned behaviour, and so on *ad infinitum*. A robot like this would live in a mental world rather than in the real world and, if the robot is really good at making these predictions, its behaviour would be as successful as that of a robot living in the real world and responding to the real world. Clearly, this is not plausible. Human beings cannot always live in their mental world, and this for two reasons. The first reason is that predictions are rarely completely correct but they are only approximately correct. If the next prediction is made on the basis of the preceding prediction rather than on the basis of the actual sensory input from the environment, prediction errors accumulate and the individual's behaviour becomes increasingly maladaptive. The second, crucial, reason is that the world of human beings is much more complex than the environment of our robots. It is intrinsically and variously unpredictable and, therefore, it cannot be entirely captured by their brain.

The first reason why it is impossible to live in a mental world – the accumulation of prediction errors – can be demonstrated with the robots that live in the environment with food tokens and occluding objects. After they have learned during the early stages of their life to predict the next visual input from food given a planned and then physically executed movement, the robots begin to live in a mental world. They begin by planning a behaviour in response to the actual visual input from the environment and they predict the consequences of this behaviour, but then they completely abandon the real world. They do not physically execute the planned behaviour but they plan another behaviour in response to the predicted consequences of the preceding planned behaviour, and so on for many cycles. The robots lose contact with the real world and they live in their mental world.

What are the results? Although the environment is very simple and predictable, the behaviour of the robots is not very effective. Predicted visual inputs are more or less correct but they are never entirely correct. Prediction errors accumulate and, after a while, the robots lose contact with the real environment and become unable to reach the food tokens. These robots exhibit a behaviour which might be called pathological and their reproductive chances are much reduced. This is why no robot (no human being) can live only in its mental world. Once in a while, the robot must look at the real world (the actual sensory input from the environment) and adapt its mental world to the real world.

8. Other consequences of the ability to predict

The capacity to predict has a central place in the adaptive pattern of human beings, and this is why, if we want to construct human robots, our robots should be able to predict. In this Section we briefly mention which other aspects of the adaptive pattern of human beings may be the result of their capacity to predict.

A capacity to predict is required if the robots must have a sense of *agency*. If a robot is able to predict the consequences of its behaviour, the robot will recognize that some events happen for independent causes and other events happen because of its behaviour. That human beings have a sense of agency is shown by experiments in which the participants in the experiment are asked to press one of two buttons and the pressing of one button is followed, after a fixed interval, by one sound while the pressing of the other button is followed by a different sound. When, in a test at the end of the experiment, the participants are asked if they are the causes of the sounds, they say Yes if their action is followed by the appropriate – predicted – sound and with the normal time interval and they say No if the wrong sound follows their pressing of the button or there is a too long temporal interval between the pressing of the button and the sound. Human beings have a sense of being the causal agents of events because they have learned to predict these events based on their actions. A robot which is able to predict the consequences of its actions may have the same sense of agency.

Human beings are more likely to be surprised and to look surprised than nonhuman animals. This also can be explained by linking surprise to the capacity to predict. When a robot has learned to predict that one event will be followed by another specific event, if in one occasion the prediction turns out to be wrong the robot should be surprised and it should look surprised. Being surprised has two components, a cognitive component and an emotional component. The cognitive component is that the robot tries to understand why its prediction was wrong. It looks at the sky if the dark clouds have not been followed by rain and it examines the glass if the glass has not broken. If its predictions are wrong, the robot wants to know why they are wrong in order to improve its capacity to make predictions. (And this may be the beginning of science. On robots that have science, see Chapter 13.) The emotional component of the robot's surprise is due to the fact that a prediction which turns out to be incorrect means that the robot is not so good at making predictions and that there are aspects of reality which it does not know. And this can make the robot preoccupied and anxious. Of course, only high-probability predictions cause surprise if they are falsified. If I expect that John will join me for lunch, knowing John I am not surprised if he does not, without telling me.

But there is one aspect of the human capacity to be surprised which our present robots would not be able to reproduce. Human beings may be surprised if something which they predict does not happen but they may also be surprised if something which they do *not* predict does happen. If a glass falls on the ground, they are surprised if

the glass does not break. But they are also surprised if a glass suddenly breaks with no apparent cause. Our robots would be surprised in the first case but not in the second case – unless we assume that they are always explicitly predicting everything which happens – which does not sound very plausible. Therefore, constructing robots that are surprised if something happens which they did not predict, is a task for the future.

Being able to predict the consequences of one's actions may also underlie *understanding* what one does. Current robots, like nonhuman animals, do things but they cannot be said to understand what they do. Psychologist Jean Piaget has distinguished between succeeding in doing X and understanding the reasons for one's success and he has shown that children move from a stage in which they can do things to produce some result but they do not understand why what they do produces that result, to a stage in which they understand why what they do produces certain results – and, often, they do things with the only purpose to know their results. Constructing robots that predict the consequences of their behaviour is a first step towards translating Piaget's distinction into robotic terms. A robot which predicts the consequences of its behaviour may not only do things that have certain consequences but it can also understand why what it has done has had those consequences – where understanding is varying the actions to identify the specific properties of one's actions which cause those consequences. (This is another step towards science.)

The capacity to predict may underlie another typically human ability, the ability to construct and use technological artefacts. *Constructing* an artefact requires a capacity to predict the consequences of one's actions on the artefact – for example, how the shape of a stone will change as a consequence of one's actions. *Using* an artefact presupposes the capacity to predict what are the effects of one's actions on the environment when these actions are mediated by the artefact (killing a prey with a modified stone). And the capacity to predict may also play a role in constructing an artefact by copying an existing artefact. (For robots that construct artefacts by copying existing artefacts, see Chapter 8.)

Another consequence of the ability to predict may be learning by imitating others. Many human behaviours are learned by imitating other human beings, and being able to predict can be a pre-condition for learning by imitating other human beings or at least a condition that results in more sophisticated imitation. A robot can imitate another robot because the robot compares the *predicted* consequences of its actions with the *observed* consequences of the actions of the *other* robot and uses the resulting discrepancy to modify the connection weights of its neural network. At the end the robot's behaviour will produce the same consequences that are produced by the behaviour of the other robot and, therefore, its behaviour will be like the behaviour of the other robot. (For robots that learn from other robots, see Chapter 8.)

Learning by imitating others is typical of human beings and, for example, learning language is learning to imitate the language of others. In the first 4–5 months of life, infants produce all sorts of sounds so that they can learn to predict the acoustic

consequences of their phono-articulatory movements. From 5 to 7 months, babbling emerges: they learn to imitate – to reproduce – their own sounds (self-imitation). From 7 to12 months, the sounds produced by the child tend to resemble the sounds of the particular language spoken in the child's environment: the child is learning to imitate the sounds produced by others. From 1 year on, true language begins: the child is learning to produce the same sounds produced by other individuals in the same circumstances – in response to the same objects or actions. (For robots that have language, see Chapter 4.) If learning by imitating others requires a capacity to predict, this capacity may play an important role in learning language.

But the most important consequence of possessing a capacity to predict may be having a causal model of world in one's brain. If a robot is able to predict that A will be followed by B, for the robot A becomes the cause of B. Of course, to say that A is the cause of B, the robot must control various things. It must observe that A is followed by B many times and it must find out why in some cases A is *not* followed by B and what property A must have to be followed by B. And it must find out if there is a third thing X such that A is the cause of X and X is the cause of B. In this way the robot's brain will possesses a complete, interconnected, model of reality as made up of causes that produce effects. Doing all these things and doing them in precise, quantitative, and socially shared ways is another human adaption: science. (For robots that have science, see Chapter 13.)

The capacity or tendency to predict the consequences of one's actions has many advantages but it may also push human beings to behave in ways which appear to be less clearly advantageous for them. As we have said in Section 5, if a robot plans an action which has consequences which are judged as "bad" by the robot's neural network and evoke negative emotional states such as fear, the robot will not execute the action. But if in the past an action has *not* been followed by negative consequences – such as dying or becoming ill or being punished by some "father" figure – the robot's brain may establish a nonexistent causal link between the action and the *lack* of negative consequences and the robot will continue to do that action to "prevent" the occurrence of negative consequences. And the robot's behaviour becomes self-reinforcing because it continues *not* to be followed by negative consequences. Examples of such behaviours are obsessive behaviours and ritualistic behaviours. Obsessive behaviours are individual behaviours: they are "invented" by one specific individual and are not socially shared or culturally transmitted. Examples are washing one's hands repeatedly, avoiding putting one's feet on lines on the pavement, and doing apparently useless things always in the same way and in the same order. Obsessive behaviours help the individual not to fear "bad" future events because the individual "predicts" that these behaviours will not be followed by the occurrence of "bad" events and these predictions turn out to be true. Ritualistic behaviours are social behaviours and they are social in two senses. They are behaviours of groups of individuals and they are subject to explicit and socially enforced rules. They do not only help not to fear "bad" future events – like obsessive behaviours – but

they also establish and maintain group cohesion and coordination. From this point of view ritualistic behaviours are similar to singing and dancing together.

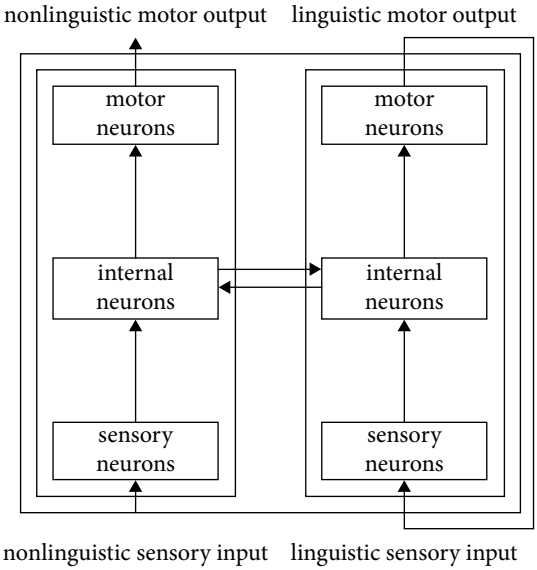
There is a final but very important consequence of being able to predict. Being able to predict may create a sense of time. All animals live in time but it is not clear if nonhuman animals have a *sense of time*. The activation pattern observed in a robot's prediction neurons refers to some future event and, therefore, robots that predict – and do not only anticipate – are robots that have time in their brain – although this is future time, not past time. This may be true for predicting in general but a sense of time may become stronger when what is predicted is described in words addressed to oneself. We discuss producing words to talk to oneself in the next Section.

9. Talking to oneself

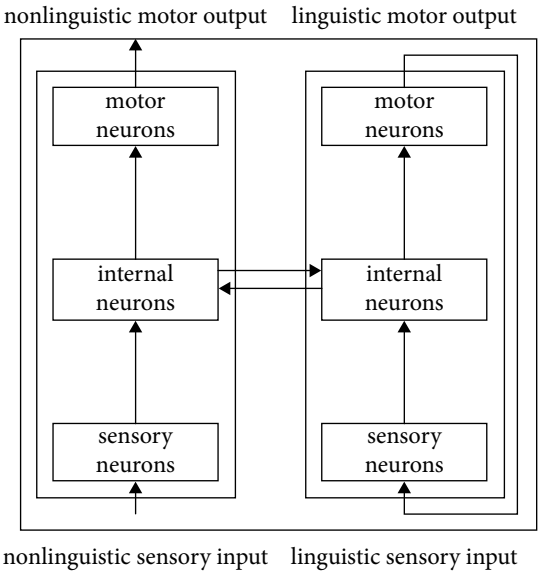
Mental life as the self-generation of sensory inputs is typical of human beings, although some primitive forms of mental life do probably exist in at least some nonhuman animals. But mental life becomes uniquely human if the self-produced sensory inputs are linguistic sounds because only human beings have language. Language is primarily social behaviour and the linguistic sounds which arrive to the acoustic sensors of one individual are caused by the movements of the phono-articulatory organs of another individual. When one talks with another individual, the first individual hears his or her own linguistic sounds but these sounds are addressed to the other individual and it is the other individual who responds to the sounds.

But human beings also talk to themselves. They produce linguistic sounds which are not addressed to other individuals but to themselves, and they respond to the linguistic sounds that they themselves produce. In some cases the linguistic sounds are physically produced so that they can also be heard by other individuals. The individual "thinks aloud" (see Figure 11a). But in most cases the phono-articulatory movements are not actually executed and, therefore, there are no physical sounds which can be heard by other individuals. Only the individual who produces the sounds "hears" the sounds and reacts appropriately to them. The individual thinks (see Figure 11b).

If mental life is the self-generation of sensory inputs, talking to oneself is by definition part of mental life. As we have seen in the chapter dedicated to language, language influences how robots experience the world. Words segment the robots' sensory-motor experiences into separate pieces and these separate pieces can be recombined together in novel ways. Words create different types of entities in the robots' brain (objects, actions, properties of objects and actions), they help the robots to categorize these entities in ways which are useful for their behaviour, and they cause the robots to have an abstract model of the world in their brain. These consequences of possessing a language extend to mental life and make the mental life of the robots more articulated and more creative. Mental images – memories, imaginations, inventions – are not only



(a)



(b)

Figure 11. Neural network of a robot that talks to itself vocally (a) or sub-vocally (b).

self-generated sensory inputs but they evoke self-generated linguistic sounds which allow the robot to analyze its mental images and its memories and to respond in more sophisticated ways to them. Predicting the consequences of one's actions becomes reasoning and planning – "if X happens, then Y happens", "if I do X, then Y happens" – and this makes the robot's behaviour "rational".

The impact of language on mental life is not restricted to its cognitive consequences. When they use language to talk with others, human beings do all sorts of things. They remind others, they discuss with others, they interpret things for others, they command others, they try to convince others to do or not to do this or that thing. When they talk to themselves, human beings do all these different things with themselves. They remind themselves, they discuss with themselves, they interpret things for themselves, they command themselves, they try to convince themselves to do or not to do this or that thing. And this makes their mental life and all their behaviour more complex and effective. Human robots must be able to talk with themselves and they must be able to do with language all the different things which they do when they use language to talk with other robots.

Talking to oneself can be called "consciousness". The English word "consciousness" has many different and ill-defined meanings and we might call all mental life "consciousness". But words are not important for science and we operationally define consciousness as talking to oneself. A robot is conscious of X if X causes the internal representation of the linguistic sound which systematically co-varies with X in the robot's experience. The robot sees a food token and it says "food" to itself. The internal representation of the food token is caused by something which is in the external world – the food token – but the robot is conscious of the external world because it lives in a linguistically "commented" world.

A robot can also be conscious of its own mental life. The robot imagines a food token and it says to itself "food". The self-generated internal representation of a food token causes the self-generated internal representation of the sound "food". And, if language changes non-linguistic experience (see the preceding chapter), language also changes self-generated non-linguistic experience.

Being (verbally) conscious of one's self-generated internal representations is crucial for higher level cognitive abilities such as reasoning and planning. Reasoning and planning are very difficult to do if one does not describe verbally to oneself events, facts, and actions and one's predictions of their consequences. As we have seen in the Chapter dedicated to language, the "space" of internal representations of linguistic sounds is smaller than the "space" of internal representations of what one actually sees or does because the linguistic module has fewer internal neurons compared to the linguistic module. And, instead of moving in the larger "space" of non-linguistic internal representations, it is easier to move in the smaller "space" of sounds and to think in words.

The consequences of having a linguistic mental life are numerous and diverse, and understanding (reproducing) them by constructing robots is one of the most important

goals of robots as science. If a robot knows how to respond to both the linguistic sounds which arrive from another robot and the linguistic sounds which are self-generated inside its brain, this reinforces the robot's sensation of living in two worlds, an external world and an internal world. A robot which self-generates linguistic sounds which describe its past experiences will more easily conceive itself as an individual with its own unique history and will more easily develop a "sense of self". A robot with a linguistic mental life will not adapt to the environment as it is but it will imagine other environments and how to change the environment to make it better – which is typical of human beings. And using language to remember things can be advantageous because delegating memory to the linguistic module allows the non-linguistic module to process other information useful to acting on the environment.

Talking to oneself may even have played a role in the evolutionary emergence of language. There is a tendency to think that language first evolved as a communication tool and, only much later, when it was already well developed and was very sophisticated and syntactically complex, human beings began to use language to talk with themselves – to think. But even a simple proto-language made up of single words may be used to talk to oneself, and, given the advantages of talking to oneself, this may have been a pressure for the evolution of language. This is what we show with the last robots of this Chapter.

The robots are those that live in the environment in which the food tokens have a roundish shape and the poisonous tokens have an angular shape but not two tokens have exactly the same shape. As we have seen in the chapter dedicated to language, having a language made up of two sounds, "food" and "poison", helps the robots to recognize if a token is food or poison. However, as we have also seen, the evolution of this language is difficult because, while the language is useful for the robots that hear the linguistic sounds produced by another robot, the advantages for the robots that produce the linguistic sounds are less clear. There is no selective pressure for producing the appropriate sound in the appropriate circumstances in that a robot's reproductive chances depend on the number of tokens that the robot eats, not on the correctness of the sounds that it produces. And, even worse, if a robot produces the appropriate sound in the appropriate circumstances in presence of another robot, the speaking robot increases the probability of reproduction of the hearing robot, thus providing a direct advantage to a competitor. The result is that a good and stable communication system does not evolve.

In the chapter on language we have already discussed this problem and we have described various ways in which the problem may have been solved by human beings. The speaker robot may have the same genes of the hearer robot and, therefore, it will act altruistically towards the hearer robot. Or the robots may have a genetically inherited tendency of learn from others – docility – which makes them learn to speak by imitating the robots that already speak. Or speaking may benefit both the speaker and the hearer. What we add here is that using language to talk with oneself may have been another cause of the origin of language.

The new robots produce the two linguistic sounds "food" and "poison" not only to help other robots to distinguish between the food tokens and the poisonous tokens but they also produce these linguistic sounds when they are alone. When they see a food token, they produce the sound "food" and, when they see a poisonous token, they produce the sound "poison". The results are that the robots which use language not only to talk to other robots but also to talk to themselves reach a higher level of fitness compared to the robots which use language only to talk to other robots. But the linguistic sounds self-produced by the robots can help them to recognize if a token is food or poison only if the robots produce the appropriate sounds in the appropriate circumstances because, otherwise, the robots can mislead themselves. This may have been another pressure for the evolution of language – a language which is not only used to talk to others but also to talk to oneself. These robots tell us that the advantages of talking to oneself do not require a complex syntactic language. Talking to oneself may have happened quite early in language evolution, before the transition from a simple protolanguage to the full-blown compositional language of modern human beings.

Mental life causes human beings to live in a natural "augmented reality". Reality is what animals or human beings interact with. Human beings live in an "augmented reality" because they interact not only with what is outside their brain – in the external environment or inside their body – but they also interact with a reality which is created by their own brain. A linguistic mental life makes this augmented reality more articulated and more complex. Human beings have what psychologist Max Wertheimer called "productive thinking" – the capacity to recombine different parts of their experiences to solve new problems. Talking to oneself makes this capacity more effective.

Mental life is not only a cognitive mental life but also an emotional mental life. The brain of an individual self-generates images and memories and the individual's brain responds to these self-generated images and memories by self-generating emotional states. Language and a linguistic mental life can also play a role in this emotional mental life, for example, they may be play a role in an important phenomenon of mental life which belongs to the motivational/emotional half of the mind rather than to its cognitive half: dreaming. Dreaming is part of mental life because dreams are images and experiences self-generated by an individual's brain when the individual is asleep. But dreams can be remembered when one is awake and they can be narrated to oneself and to others by using language – and they may reveal much about the individual that dreams to both the individual that dreams and to others. We can know if a robot dreams and, perhaps, what it dreams by examining the activity of its neural network when the robot is asleep but, if the robot has language, we can also know what a robot dreams by listening to its linguistic description of its dreams. But, notice, human robots should dream what human beings dream, not "electric sheep".

CHAPTER 6

Social robots

When ME arrives on the Earth, the first thing which it notices about human beings is they are almost never alone. They spend most of their time with other human beings interacting with them and doing things together. And, even when they are alone, they often think of other human beings, of what other human beings think of them, of how they should behave towards other human beings, of what they would like other human beings to behave towards them. What explains why human beings are so preoccupied with other human beings is that they obtain most of what they need or what they want but from other beings. They have an extreme form of social dependence.

This convinces ME that, if its robots must be really like human beings, the robots must have a rich and complex social life. But ME wants to proceed gradually. Living with others has many advantages but also disadvantages because, if one individual lives with another individual, the individual can be damaged by the other individual. So ME decides to first construct robots that reproduce the disadvantages of living together and what human communities do to contain the behaviours that damage other human beings so as to avoid that the disadvantages become greater than the advantages and human beings prefer to live alone. But ME will also address the advantages of living together which, of course, are very numerous. ME is also interested in the fact that, for a human being, another human being is a very different piece of the environment compared to all the other things that make up the environment. ME has some ideas on how to reproduce these differences but constructing these differences is a task for the future.

1. There is no social robotics today

All animals live in an environment of inanimate objects, plants, and animals of other species but some animals also live in an environment made of other members of their same species: a social environment. Of course, all animals live in the same physical environment, the Earth. But living socially has a more restricted, and more interesting, sense and in this more restricted and more interesting sense a group of individuals of the same species live socially if they live in close spatial proximity to each other so that they can perceive each other, interact with each other, and work together to produce

what no single individual would be able to produce by working alone. Living socially is a matter of degree. In all sexually reproducing animals two individuals of the opposite sex get together to mate and, in many species, parents remain in proximity to their dependent offspring to feed and protect them. But there are some species which are more social than others, and the members of these species live together beyond the necessities of mating and caring for their offspring. Human beings are a very social species and they have very complex forms of sociality. They obtain most of what they need or want from other human beings, they learn most of their behaviours from other human beings, and they know better the environment because they also know what other human beings know about the environment. And human beings often do things with the only purpose to induce other human beings to do what they want them to do and they are preoccupied of what other human beings think of them because what other human beings think of them determines how other human beings behave towards them. Human beings have a mind so impregnated with sociality that their mind can be called a social mind.

If sociality is such an important component of the adaptive pattern of human beings, to really exploit the potential of robots for making us understand human beings we must construct robots which have a rich and complex sociality. Today a social robotics is almost nonexistent. What goes under the name of social robotics is the construction of robots that interact *with us*, not *between them*. Animals that belong to different species can interact together (for example, human beings and dogs) but sociality is mainly intra-specific, not inter-specific – and, at least for now, robots and human beings are different species. The reason why current "social robots" are robots that interact with us and not between them is that robots are constructed with practical applications in mind, and many practical applications involve an interaction between robots and human beings: robots that care for old or ill people, domestic robots, therapeutic robots, robots for entertainment, robots that collaborate with human beings to produce things, etc. These are very useful applications but if robots must help us to better understand sociality we need to construct robots that interact with other robots, not with us.

There is an exception to this rule and this exception is collective robotics – or swarm robotics or ant robotics – a field of robotics dedicated to constructing robots that do things together. But, as these terms make clear, collective robotics is concerned with the sociality of simple animals such as insects, birds, and fish, and the sociality of these animals is very different from the sociality of human beings. To construct groups of robots that do things together may be an interesting alternative to single robots for accomplishing certain practical tasks because, compared to single robots, groups of robots may have a more flexible behaviour, they may be more robust in case of damage, and they may acquire a wider knowledge of the environment. But the groups of robots of collective robotics have one single goal which is shared by all the robots of the group, and this goal is the goal of the application, that is, *our* goal. Human beings

have their own goals and the goals of one individual can be different from the goals of another individual and it can even be in contrast with the goals of the other individual. Therefore, collective robotics does not address many important social phenomena which we observe in human beings, and in particular it does not address problems of self-interest and conflict and how robots might deal with them. Problems of self-interest and social conflict are at the origin of some of the most significant phenomena of the social life of human beings and this is a very important reason why collective robotics is not the social robotics that we need to understand human beings.

Collective robotics has another limitation. The teams of robots of collective robotics are almost always composed of *identical* individuals while, in real animals, no two individuals are identical and inter-individual differences have important consequences for social interaction and social phenomena. The fact that the groups of robots of collective robotics are composed of identical robots has a more specific implication. Biology tells us that behaviour towards genetically similar individuals (kin) tends to be different from behaviour towards genetically different individuals (non-kin) and, while nonhuman animals mostly live in groups of genetically related individuals, human communities, beyond a given size, are made up of genetically unrelated individuals. Therefore, if we are interested in the social life of human beings we must distinguish between groups of robots which have similar genes (robotic families) and groups of robots which have different genes.

Another limitation of collective robotics concerns how behaviour is rewarded. In collective robotics, if the robots are rewarded for their behaviour, they are rewarded collectively, not individually, while in human societies it is an open problem how to distribute the rewards of collective accomplishments. Furthermore, the groups of robots of collective robotics tend to be "horizontal" structures in which no robot tells the other robots what they must do. On the contrary, human communities are "vertical" structures that include a "chief" which tells the other members of the community what they must do, and is generally obeyed.

Given all these limitations, although collective robotics is interesting and suggestive, if we want to construct robots to better understand human sociality and human societies we must go beyond collective robotics and we must develop a real social robotics.

In this Chapter we describe robots that reproduce some very basic phenomena linked to sociality: what are the consequences of living together, what are the disadvantages of living together, what human communities do to contain behaviours that damage others to avoid to disappear because the disadvantages of living together are greater than the advantages. In this Chapter we touch on only one advantage of living together: groups of robots can function as "information centres" so that an individual can know not only what it directly knows about the environment but also what other individuals know. But in the successive chapters of the book we describe many other advantages of living together. We describe robots that live in families, robots that learn

from other robots, robots that live in communities with a "chief" which tells the other members of the community what they must do so that all the members of the community can contribute to the "common good", robots that specialize in producing only one type of goods and then exchange their goods with other robots, robots that develop money and money-based economies, and robots that organize themselves to produce goods that no single robot would be able to produce by working alone. But it is at the end of the present chapter that we discuss a crucial aspect of human sociality: how the social environment is different from the environment of inanimate objects and how behaviour towards other human beings is different from behaviour towards inanimate objects.

2. Living together

Although social behaviour changes the environment of another individual, not all behaviour which changes the environment of another individual is social behaviour. A robot may eat the food which exists in the environment of another robot and, as a consequence, there is less food in the environment for the other robot and the other robot is forced to go elsewhere to eat. This, by itself, is not social behaviour. Social behaviour is behaviour which exists *because* it affects the behaviour of other individuals. If the first robot eats the food *in order to* push the second robot to go elsewhere, then the behaviour of the first robot *is* social behaviour. And, at least for human beings, another condition that must be satisfied is that the other individual must "understand" that the behaviour of the first individual has the goal to push him or her to go elsewhere.

In this Section we don't describe robots which have social behaviours but we describe robots that satisfy one pre-condition of sociality: living together in the sense of living in physical proximity. Today, due to the development of technologies for transporting goods and information, human beings can interact with other human beings even if these other human beings are in distant places. But even today many social interactions require spatial proximity. And for millennia physical proximity has been an inevitable requirement for social interaction – and this is important to keep in mind because, as we have said in Chapter 1, if we want to understand human beings, we must understand how human beings were in the past and how they have become what they are.

Unlike plants, animals autonomously move in the environment and they can move independently from one another and live in isolation. So, why the members of some species live near to one another? One reason is the spatial distribution of resources, and this is what we show with our next robots.

If food is equally distributed in the entire environment, there is no spatial aggregation and the robots live everywhere one independently from the other (see Figure 1).

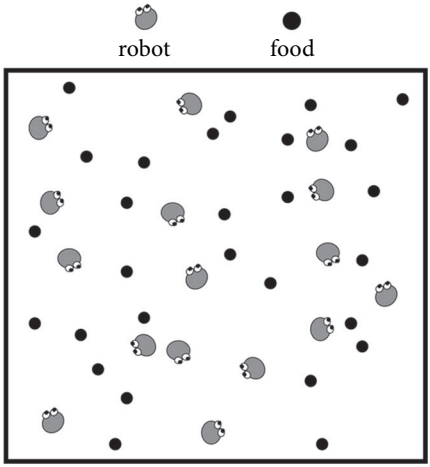


Figure 1. The robots live in the same environment but they do not live near to one another because food is everywhere in the environment.

But if food is only in certain zones of the environment, the robots tend to aggregate in these zones, and this happens even if the robots do not perceive each other and they only perceive the food tokens. The environment of these new robots has four zones which contain many food tokens and are separated by spaces with little food (see Figure 2a).

When a robot is born, it is placed in a randomly chosen location in the environment, not necessarily in one of the food zones and since, initially, the robots are not very good at reaching the food tokens, they wander aimlessly in the environment (see Figure 2b). Then, as always, the selective reproduction of the best robots and the addition of mutations to the inherited genes lead to robots that are better at reaching the food tokens and, given the nature of the environment, the robots tend to live near to other robots inside one or the other of the four food zones (see Figure 2c). This is a simple demonstration that the intrinsic nature of the environment can lead to living in spatial proximity which, as we have said, is a precondition for sociality – although, clearly, we cannot speak of sociality in this case because the robots do not perceive each other and they approach the food tokens, not the other robots.

All the robots die at the same time and they are replaced by the robots of the new generation. Therefore, there is no generational overlap and the offspring robots do not live together with their parents. We now construct another population of robots which die if they are unable to eat and therefore different robots die at different ages. Until a robot remains alive, at regular intervals the robot generates one offspring which is immediately placed in the environment so that at least for one part of their life the offspring robots live together with their parents. But, unlike the preceding robots, a newborn robot is not placed in a randomly chosen location in the environment but,

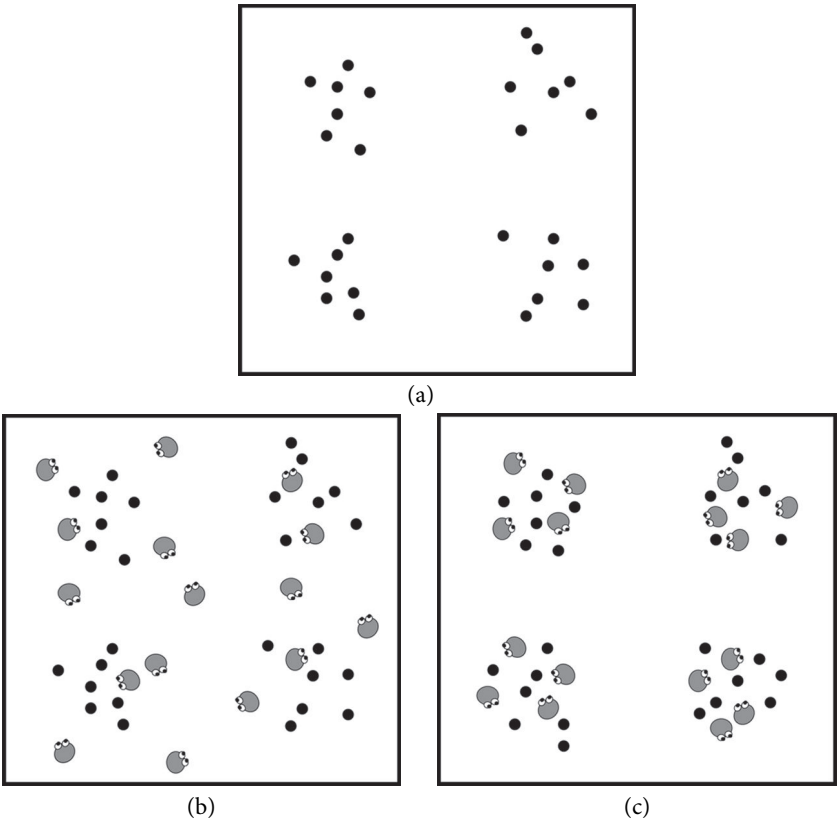


Figure 2. Food is in four zones of the environment separated by spaces without food (a). At the beginning the robots are dispersed in the entire environment (b) but, then, they concentrate in the four food zones and, therefore, they tend to live near to other robots (b).

more realistically, it is placed near to its parent. Like the preceding robots, these robots tend to aggregate in the four food zones but, since the offspring robots are born near to their parents and their parents live in one of the four food zones, social aggregation is greater and faster. But what is important is something else. In the case of the preceding robots, the robots which lived together in the same food zone were genetically unrelated robots. They were born in a randomly chosen location in the environment and only happened to end up in the same zone. On the contrary, the more realistic robots whose offspring are placed at birth near to their parents tend to form groups of robots that have the same – or very similar – genes. This is very important for sociality. The behaviour of these robots is dictated by their genes and, therefore, robots that have similar genes (parents and offspring, siblings) tend to behave similarly. Behavioural similarity is a favourable condition for sociality because it makes the behaviour of others more predictable and being able to predict the behaviour of others facilitates social

interaction. (We return to the link between social behaviour and the predictability of behaviour in the last section of this Chapter.) These robots still cannot be said to have sociality but, for them, two important conditions that lead to sociality are realized: spatial proximity and behavioural similarity.

Another interesting phenomenon linked to sociality can be observed in these robots. As we have said, the environment contains four zones where food is abundant separated by a space which is not completely empty but contains very little food. Some robots try to traverse the space between the food zones but only a small number of them succeed in reaching another food zone. This creates geographical isolation between the groups of robots which live in different zones, and geographical isolation plus the constant addition of random mutations to the inherited genotypes causes what is called genetic divergence and, therefore, behavioural divergence. Not only the robots that live in the *same* food zone tend to be genetically related and, therefore, to have similar behaviours, but they become progressively more dissimilar from the robots living in the other food zones. The robots that live in one food zone exhibit behaviours which become progressively more different from the behaviours of the robots living in the other food zones, and this behavioural divergence may act against social interaction because social interaction requires behavioural similarity and predictability. The spatial structure of the environment not only facilitates social interaction among the robots that live in the same zone because they have similar behaviours but it also makes social interaction among robots living in different zones more difficult because they have different behaviours.

The behaviour of these robots is completely innate – encoded in their genes. The robots do not learn anything during their life and their life experiences have no consequences for their behaviour. But if the robots were human robots, most of their behaviour would not be encoded in their genes but it would be learned, and learned from others. Learning from others requires spatial proximity and, therefore, this is another way in which spatial proximity creates behavioural homogeneity – although behavioural homogeneity now has a cultural rather then a genetic origin. Two robots that live near to one another learn from one another and, in this way, their behaviours become more similar. And the parallel between genetics and culture extends to behavioural divergence. The robots that learn from one another because they all live in the same food zone develop a shared culture and their culture does not only become progressively more internally homogeneous but it also becomes progressively more different from the culture of the other groups which live in the other food zones. (For robots that learn from others and have cultures, see Chapter 8.)

These robots tell us that living together may be a consequence of the intrinsic characteristics of the environment. If there is more food in one zone of the environment, the robots tend to concentrate and to live together in that zone. This applies to both nonhuman animals and human beings but, for human beings, the environment can play a more important role in determining spatial aggregation and living together

because human beings construct the environment in which they live. Cities are constructed environments and cities are places where many goods (things that one wants to have) can be found in the same way as much food can be found in the four food zones of our robots. And the environment outside cities (country, sea, mountains) has few goods in the same way as there is little food in the space that separates the four food zones of our robots. Human beings modify the environment in which they live in such a way that most of the goods they want to have are located in specific places and, as a consequence, they tend to live in these (very crowded) places. (Today, more than half human beings live in cities.)

We have not constructed robots that live in cities but we have constructed robots that modify the environment in which they live in such a way that they tend to aggregate in specific places not because of the intrinsic characteristics of the environment (as in the robots described so far) but because of their behaviour. Unlike the preceding robots, the new robots live in an environment which is homogeneous and food exists in all parts of the environment – the environment of Figure 1. Hence, the environment by itself does not produce spatial aggregation and living in groups. However, the new environment is a seasonal one. When a robot eats a food token, the food token disappears and it is not immediately replaced by a new food token. At the beginning of each season the environment contains many food tokens but then the number of the food tokens is progressively reduced because the robots eat them. Then, the next season begins and, again, there is much food in the environment. Like the preceding robots, these robots die if they do not eat, they generate one offspring at regular intervals, and at birth the offspring are placed near to their parents.

What happens to these robots from the point of view of living in spatial proximity? What happens depends on the length of seasons. If the seasons are very short and, therefore, food reappears frequently and it is always abundant, nothing special happens. Initially, the robots are distributed everywhere in the environment and they continue to be distributed everywhere, with no spatial aggregation. On the contrary, if seasons are very long, the population becomes extinct. The robots eat all the food before the next season begins and, therefore, they all die for lack of food. These results are predictable and not particularly interesting. What is interesting is what happens if the seasons have an intermediate length because, with seasons of intermediate length, we see the emergence of *periodic migratory waves* (see Figure 3).

For purely chance reasons one zone contains more food than the rest of the environment and the robots concentrate in that zone. However, the presence of many robots leads to a rapid reduction of the food of that zone and the robots are forced to leave the zone. During the period in which the robots were concentrated in the zone with more food, there were very few robots in the rest of the environment and, therefore, the food existing in the rest of the environment has not been eaten and it is still there. So the robots disperse in the environment because they are attracted by the food which is abundant in the rest of the environment. After a while the rest of the environment

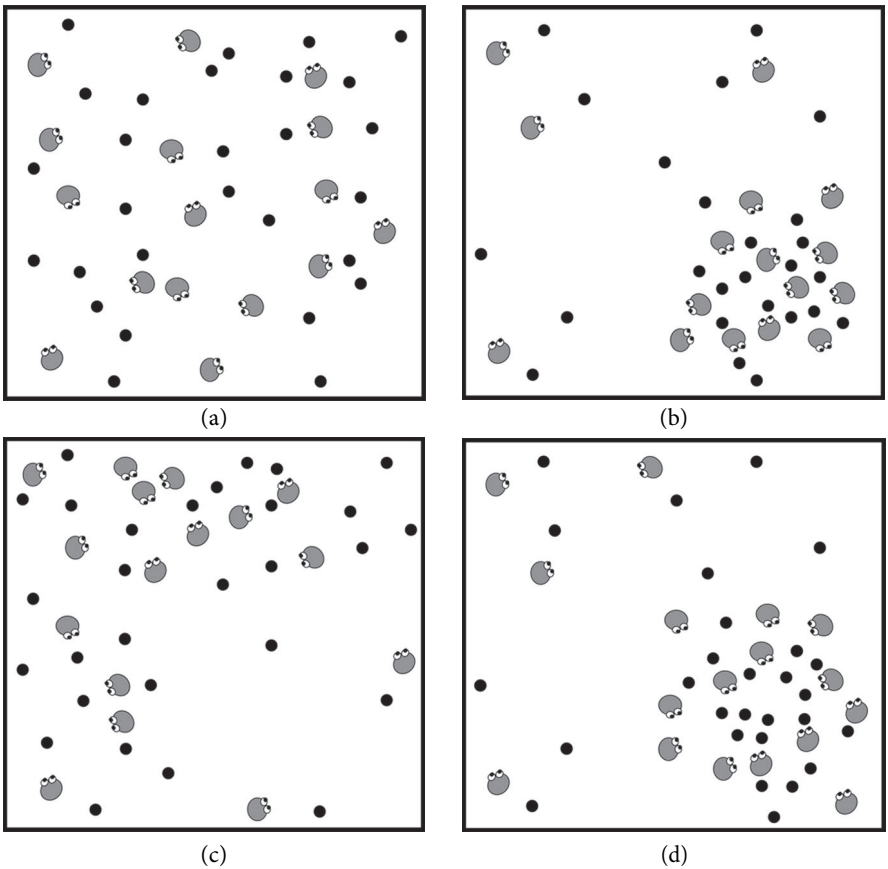


Figure 3. An environment in which food is randomly distributed and it reappears seasonally (a). The robots aggregate in a zone that happens to contain more food but, when they have eaten all the food contained in that zone, they disperse in the environment (c) to later re-aggregate in the original zone (d).

remains with little food because the robots have eaten most of the food which was there, while the original zone is again full of food. So the robots return to the original zone, and the process repeats itself cyclically producing periodic migratory waves.

These robots tell us that it may not be the environment with its intrinsic properties which causes spatial proximity and, therefore, social interaction. The environment is homogeneous and the robots might live in the environment with no spatial aggregation. It is the robots' behaviour – eating the food tokens – which modifies the environment in ways that favour spatial proximity. Modifying the environment in ways that favour spatial proximity is typical of human beings and it has probably been an important factor in the development of their complex sociality.

3. Why *not* live together

But many animals live in spatial proximity not because of the nature of the environment or because they modify the environment but because, by living in spatial proximity, they can interact together and live socially – and there are many advantages of living socially. What are these advantages? The importance of this question becomes clear if we consider that living together may have *disadvantages*, and, therefore, unless sociality is really advantageous, an individual should prefer to alive alone rather than with other individuals. In this Section we describe robots which show the disadvantages of a prerequisite of living socially: spatial proximity.

Living in spatial proximity may create problems for purely physical reasons. Like real organisms, robots are bodies and bodies are physically impenetrable. Therefore, if a robot lives with other robots, the other robots may be a physical obstacle for the robot's movements in the environment. This is what we show with our next robots.

The robots live in an environment in which food exists in a single zone and, to eat and remain alive, the robots must reach the food zone and remain inside the food zone for as much time as possible. At the beginning of their life the robots are randomly distributed in the entire environment. The robots have visual sensors that allow them to see the food zone from a distance and other sensors that tell them when they are inside the food zone. We compare two different populations of robots. In one population each robot lives alone in its environment while in the other population the robots live together in the same environment. The robots of both populations evolve the capacity to approach and reach the food zone and to remain inside the food zone but the fitness of the robots (which depends on the total time spent inside the food zone) is lower for the robots that live together than for the robots that live alone. The explanation is very simple. As bodies, the robots are physically impenetrable and, therefore, they constitute physical obstacles for a robot both when the robot is approaching the food zone and when it tries to enter the food zone. Or a robot which is trying to reach the food zone can find another robot in its route to the food zone and the other robot is an obstacle to its movements. A robot which is trying to penetrate inside the food zone can find the food zone already full of other robots so that for the robot it is difficult to enter the food zone. As a consequence, for the robots that live together it is more difficult to reach and enter the food zone compared to the robots that live alone. They spend less of their time in the food zone and, therefore, eat less (see Figure 4).

And this is true even if the robots see not only the food zone but they also see each other. When a robot sees another robot which is between itself and the food zone, the robot must avoid the other robot but this requires the acquisition of another capacity and the consumption of more time to reach the food zone. Clearly, for these robots living together is a disadvantage. And the disadvantage is greater if the environment is overcrowded. If we put more robots in the same environment, we find that the proportion of robots that penetrate the food zone is progressively reduced.

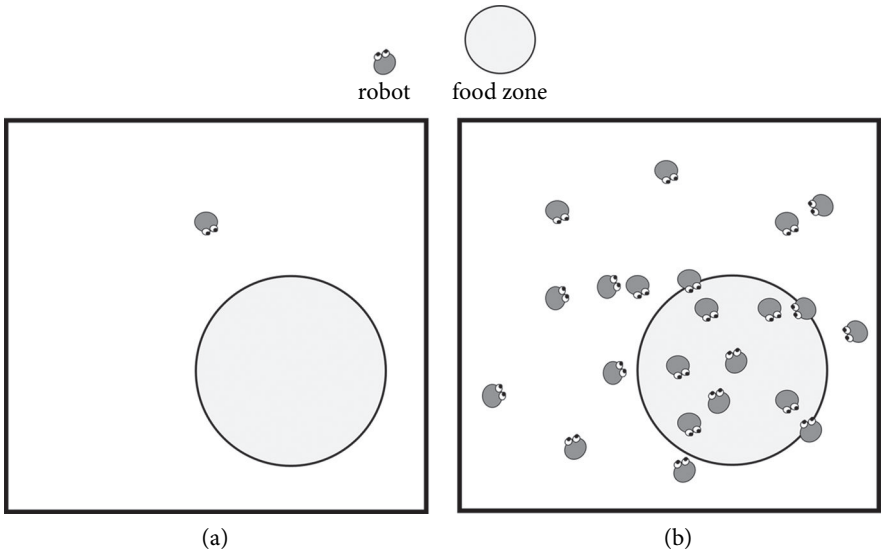


Figure 4. The environment contains a single food zone. If a robot lives alone, there are no obstacles to reaching and entering the food zone (a). If many robots live together, they may be unable to reach and enter the food zone because it is already full of robots (b).

Living together can be disadvantageous for other reasons beyond the physical impenetrability of bodies. Living together can be disadvantageous because the behaviour of an individual can change the environment in ways that damage another individual. We return to the robots of Figure 1 that live in an environment containing randomly distributed food tokens. Again, we compare a population of robots in which each robot lives alone in its environment with a population of robots that live together in the same environment although the robots do not perceive each other but they only perceive the food tokens. The capacity to eat the food tokens evolves in both populations but, again, at the end the robots that live together eat fewer food tokens than the robots that live alone. Therefore, also for these robots living together is disadvantageous. The reason is not the physical impenetrability of bodies because the bodies of these robots are physically penetrable and therefore a robot does not represent a physical obstacle for another robot. And the disadvantage is not that, by eating the food tokens, the robots reduce the quantity of food available to the other robots because, when a robot eats a food token, a new food token immediately appears in the environment. So, why living together is disadvantageous for these robots? The problem is another one and is much simpler. When a robot approaches a food token, the food token can suddenly disappear because, in the meantime, it has been eaten by another robot. Living with other robots makes the life of the robots more difficult because the robots may waste their time approaching food tokens that suddenly disappear.

These robots perceive the food tokens but they do not perceive each other, and one might think that if the robots were able to perceive each other, the problem would be solved: don't waste your time approaching a food token when another robot is nearer to the food token than you are. But the problem does not disappear. We construct another population of robots which perceive not only the food tokens but also the other robots, and, again, the robots that live with other robots eat less food than the robots that live alone. If we look at the robots on the computer screen, we see that the robots tend to avoid each other. If we measure the average distance among the robots, we find that this distance is greater for the robots that perceive each other than for the robots that do not perceive each other. The robots that perceive each other have evolved a tendency to keep a certain distance between themselves and the other robots.

We directly see the behaviour of keeping a distance between oneself and the other robots if we let the robots evolve in another environment in which the food tokens are randomly distributed in the entire environment but, at beginning of their life, all the robots are in a restricted zone of the environment. The robots that do not perceive each other tend to remain in the restricted zone of the environment in which they were born while the robots that perceive each other tend to disperse in the entire environment because they prefer not to be too near to the other robots (see Figure 5).

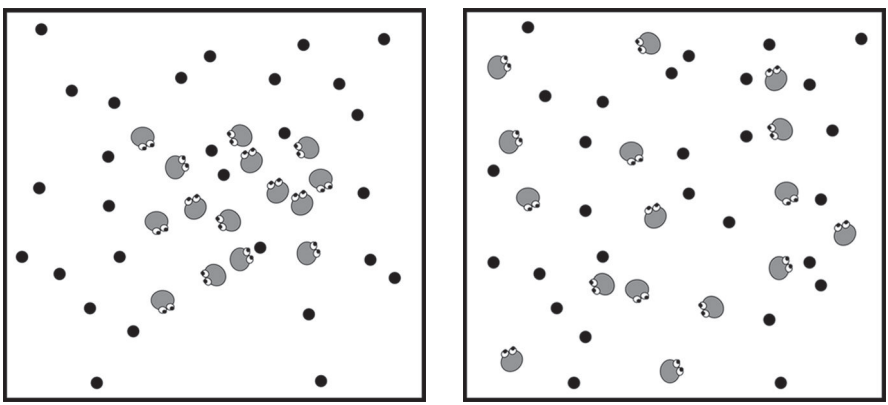


Figure 5. The robots are born in the center of the environment (a). While the robots that do not see each other tend to remain in that zone, the robots that see each other disperse in the entire environment to keep a distance between themselves and the other robots (b).

The robots "know" that the other robots can make their attempts at reaching the food tokens useless, and they adapt to the situation by evolving two behaviours rather than only one: they evolve the behaviour of approaching and eating the food tokens but they also evolve the behaviour of maintaining a certain distance between themselves and the other robots – and this is why they disperse in the environment. Living together for these robots is disadvantageous and, therefore, they tend *not* to live together.

4. Socially damaging behaviours and how to contain them

So, if one lives with others, one can be damaged by the behaviour of others. Behaviours that damage others exist among all social animals but they are so common among human beings that, if human beings don't find how to reduce the probability of occurrence of these behaviours, their societies risk dissolving because the disadvantages of living together become more numerous than the advantages. Why socially damaging behaviours are more frequent among human beings than among animals? Two hypotheses, that might both be true, could be tested by constructing robots that live together. The first hypothesis is that, while nonhuman animals live with genetically related individuals, human beings live with genetically extraneous individuals and an individual is more likely to damage another individual if the other individual does not have his or her same genes. The second hypothesis is that human beings obtain most of their goods not from nature but from other human beings. Therefore, they have more occasions to interact with other individuals and, in the course of these interactions, to damage or be damaged by other individuals.

If one individual damages another individual, the damaged individual can punish in one way or another the damaging individual, and this reduces the probability of socially damaging behaviours. Being punished reduces an individual's survival/reproductive chances and its well-being and, therefore, in the future the damaging individual will hesitate to damage other individuals. But, among human beings, the damaged individual may not know that he or she has been damaged or may not know by whom he or she has been damaged or – what is even more important – may be unable to punish the damaging individual. And not single individuals but entire organizations of individuals can damage others, and organizations of damaging individuals are more difficult to discover and punish. For all these reasons, to continue to exist, human societies must possess some mechanism for discovering and punishing socially damaging behaviours. In this Section we describe robots that reproduce, in an extremely simplified form, what human communities do to contain socially damaging behaviours.

The new robots are abstract robots and, in fact, they are "agents" rather than robots. They do not have a body, they do not live in a physical environment, and they do not behave in the sense that they respond to sensory inputs with movements of their body. But they do "things" and, since they live together, what they do may have consequences for the other robots. The robots can behave in two ways, honestly or dishonestly, where honest and dishonest behaviours are defined by their consequences. A robot behaves honestly if its behaviour increases its well-being without reducing the well-being of other robots. A robot behaves dishonestly if its behaviour increases its well-being but decreases the well-being of another robot – the robot kills the other robot or steals the goods of the other robot or damages the other robot in some other way. Each robot is characterized by an "honesty number" whose value can go from 0 to 1 and this number is the probability that, in any given cycle, the robot will behave honestly or dishonestly. For example, a robot with an "honesty number" of 0.8 behaves honestly in 80% of the

cycles of its life while a robot with an "honesty number" of 0.3 behaves honestly in only 30% of the cycles and in the remaining 70% it behaves dishonestly. For simplicity, we call the robots which have an "honesty number" greater than 0.5 honest robots and the robots with an "honesty number" smaller than 0.5 dishonest robots but, since the number can go from 0 to 1, there are robots with different degrees of honesty and dishonesty. Each robot has a level of well-being which changes during the robot's life as a consequence of the robot's behaviour and of the behaviour of the other robots. A robot can increase its well-being by behaving honestly or dishonestly but, if the robot behaves dishonestly, the robot also *decreases* the well-being of another, randomly selected, robot – the damaged robot. But the well-being of a robot can also decrease for another reason, although this only applies to the robots that behave dishonestly. If in one cycle a robot behaves dishonestly, there is some probability that the robot will be discovered and punished, where being punished means that the robot's well-being is reduced by some fixed quantity.

The "honesty number" of the robots of the initial generation is decided randomly, which means that half of the robots – more or less – behave honestly and half dishonestly (see Figure 6).

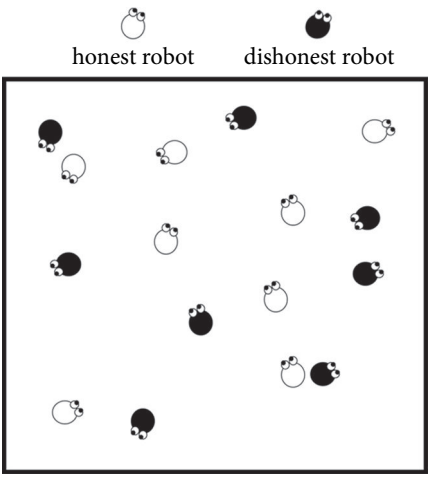


Figure 6. At the beginning of the simulation half of the robots behave honestly and half dishonestly.

Life has the same length for all robots and it is a succession of cycles. In each cycle a robot behaves honestly or dishonestly according to the probability specified by its "honesty number" and, as a consequence of its behaviour, the well-being of the robot changes. If the robot behaves honestly, its well-being increases by a fixed quantity and there are no other consequences. If the robot behaves dishonestly, the consequences are more complex. Not only the robot's well-being increases and the well-being of

another, randomly selected, robot – the damaged robot – decreases but there is some probability that the dishonest robot is discovered and punished for its behaviour, which implies a reduction of its well-being. At the end of their life, the robots are replaced by a second generation of robots. The new robots learn how to behave by imitating the robots of the first generation. Each of the robots of the second generation copies the "honesty number" of one robot of the first generation with the addition to small random changes which can make the robot more or less honest compared to its model. Each robot chooses one robot of the preceding generation as the model to imitate but its choice is not random. The robots of the first generations which have a higher level of well-being are more likely to be imitated by the robots of the second generation. The process goes on for a certain number of generations and at the end we count the number of honest and dishonest robots. At the beginning, half of the robots are honest and half dishonest. How many are the honest and dishonest robots at the end of the simulation and what explains the number of honest and dishonest robots?

The number of honest and dishonest robots at the end depends on various factors. One factor is the probability that dishonest behaviours are punished. When a robot behaves dishonestly, it *may* be punished but it is *not necessarily* punished because, to be punished, a dishonest behaviour must first be discovered, and we vary the probability of being discovered and punished in different communities of robots. A second factor is the severity of punishment. In some communities of robots punishment is more severe than in other communities, and this means that, if a dishonest robot is punished for its behaviour, its well-being is reduced by a quantity which can be greater or smaller. The third factor is how much the level of well-being of a robot increases if the robot behaves honestly and if it behaves dishonestly. In some communities behaving honestly is more convenient than behaving dishonestly while in other communities the opposite is true and dishonest behaviour increases the well-being of a robot two or three times *more* than honest behaviour.

What happens when we vary these factors? As we have said, the robots chosen as models to imitate are those which have a higher level of well-being. Which robots have a higher level of well-being, the honest robots or the dishonest robots? If dishonest behaviour increases a robot's well-being more than honest behaviour, dishonest behaviours are almost impossible to extirpate and at the end most robots behave dishonestly (see Figure 7a). Dishonest behaviours are reduced in number only if for a dishonest robot there is a 100% probability of being discovered and punished and if punishment is severe. If punishment is not severe, half of the robots continue to behave dishonestly even if the probability of being discovered and punished is 100%. And we know that, in human societies, it is very difficult to guarantee that all dishonest behaviours are discovered and punished. If the probability of being punished is not 100% or near to 100%, all robots behave dishonestly, independently of the severity with which they are punished *if* they are punished.

This is what happens if behaving dishonestly is convenient because it increases one's well-being more than behaving honestly. If behaving honestly and behaving dishonestly are equally remunerative, it is easier to contain dishonest behaviours. Dishonest robots continue to be numerous only if the probability of being punished is low and punishment is not severe but in all other cases almost all robots behave honestly. And the situation is even better if behaving honestly is *more* remunerative than behaving dishonestly. If the robots can gain more with honest behaviour than with dishonest behaviour, dishonest behaviours are (almost) entirely eliminated (see Figure 7b).

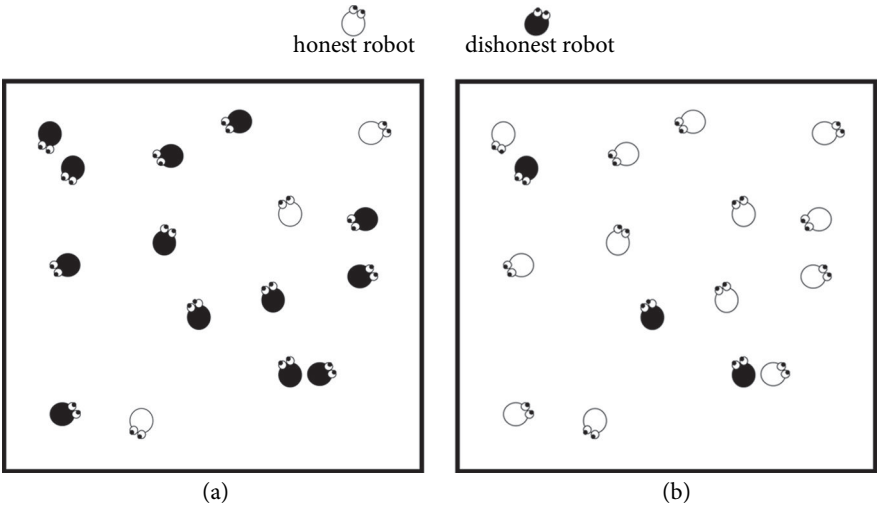


Figure 7. In some communities most robots behave dishonestly (a) and in other communities honestly (b), and this depends on how much can be gained with honest behaviour, on the probability of being punished, and on the severity of punishment.

These robots tell us that the best policy for containing dishonest behaviours is not to increase the probability that dishonest behaviours are discovered and punished or to increase the severity of punishment but to create opportunities for all the members of the community to have a good level of well-being by behaving honestly. Only if this strategy is adopted, dishonest behaviours are almost completely eliminated, independently of the probability of discovering and punishing them and of the severity of punishment.

In human societies punishing the behaviours that damage others is one of the most important task of the state – or of the precursors of the state – but our results show that for the state is not so easy to contain socially damaging behaviours. The best strategy is a preventive one, the strategy which consists in increasing the level of well-being which can be reached by behaving honestly. But this strategy requires economic policies which, for a variety of reasons, are not so easy to find and to adopt.

Behaving dishonestly can increase one's well-being to an extent which it is impossible to guarantee to all the honest members of the community, and this is particularly true in today's human societies in which even people with a good level of well-being may behave dishonestly because what is a "good level of well-being" constantly increases. The other strategy for containing socially damaging behaviours – increasing the probability that these behaviours are discovered and punished when they have already occurred (through police, investigative bodies, and courts) – is very expensive and it may not be very effective.

There is another factor which limits the capacity of the state to contain socially damaging behaviours, a factor which we have not reproduced with our robots: socially damaging behaviours can be behaviours which damage not specific individuals – as in our robots – but the entire community (corruption, collusion, evading taxes, and other crimes against the entire community). These socially damaging behaviours are difficult to discover and punish because they often involve individuals who belong to the organization of the state and therefore, the state is caught in a contradiction, the contradiction of being at the same time who punishes and who should be punished.

But there is another reason why the state may find it difficult to contain socially damaging behaviours. Our robots learn how to behave honestly or dishonestly by imitating the robots of the preceding generation, and all their learning occurs at birth – which, clearly, is not plausible. We now construct another community of robots which still learn how to behave from the robots of the preceding generation but then they continue to learn during their life by imitating the robots of *their own* generation. While the preceding robots behaved in the same way, honestly or dishonestly, during their entire life, the behaviour of the new robots can change during a robot's life and it can become more honest or more dishonest because of the interactions of the robot with the other robots of the same generation. But there are two differences between imitating the robots of the preceding generation and imitating the robots of one's own generation. In both cases imitating another robot means copying the honesty number of the other robot with the addition of some random changes to this number. But, while the robots of the preceding generation which are chosen as models to imitate are those which have a higher level of well-being, a robot imitates another robot of its own generation independently of the level of well-being of the other robot. One robot imitates another robot of its generation because it interacts with the other robot. The second difference is that imitation among the members of the same generation is reciprocal. If robot A imitates robot B, automatically robot B imitates robot A. The robots imitate each other because they interact with each other, and we represent these interactions as a network of nodes. Each node is one robot and, if two nodes are linked together, this means that the two robots represented by the two nodes interact together and imitate each other.

What determines the structure of the network? Who interacts with whom? At the beginning, the links between the nodes are random and therefore a robot can interact

with any other robot. But then the structure of the network changes because the robots that imitate the *same* robot of the preceding generation automatically interact together and imitate each other. This creates sub-communities of robots which tend to behave in the same way, with sub-communities which behave honestly and sub-communities which behave dishonestly (see Figure 8).

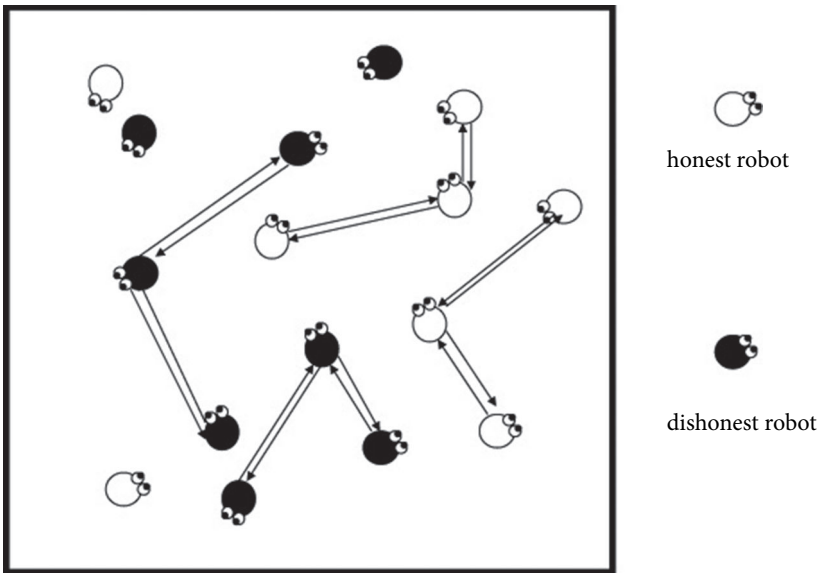


Figure 8. Robots that interact and learn from one another (links between robots) tend to form sub-communities of honest robots and sub-communities of dishonest robots.

What happens with these robots which live in homogeneous communities of honest or dishonest robots? The same factors influence the behaviour of these robots. Dishonest robots are more numerous when the probability of being discovered and punished is low and punishment is not severe but the main factor which determines the number of dishonest robots is the level of well-being that can be reached by behaving honestly or dishonestly. There are more dishonest robots if behaving dishonestly allows a robot to live better than behaving honestly.

But there is something which is new and interesting: the existence of homogeneous communities of honest or dishonest robots increases the number of *dishonest* robots. Both the honest and the dishonest robots imitate each other if they are members of the same sub-community but the consequences are not the same if the robots are part of a sub-community of honest robots or of a sub-community of dishonest robots. Imitation among honest robots has no significant consequences for the behaviour of the robots. If a robot is honest and it imitates another honest robot, it remains an honest robot.

On the contrary, dishonest robots that imitate other dishonest robots become more resistant to punishment. A robot should be discouraged from behaving dishonestly because it risks punishment but, if the robot is a member of a sub-community of dishonest robots that imitate each other, punishment is not very effective because dishonest behaviour is continually reinforced by the imitation of other dishonest robots. Therefore, the existence of sub-communities of dishonest robots increases the number of dishonest robots.

What is also interesting is that the effect is stronger if there are more links among the nodes – more interactions among the robots. We construct two different communities of robots, one in which, on average, there are fewer links among the robots and another one with more links – where "more links" means that the robots are more influenced by the other robots of their community – and we find that the increase in the number of dishonest robots due to the existence of sub-communities of dishonest robots is greater when there are more links among the robots. Punishment decreases the probability of behaving dishonestly but its effects are reduced because there is an opposing force: imitating other dishonest robots. And we find that this force is especially strong if the sub-communities of dishonest robots have more internal links.

Many other phenomena concerning socially damaging behaviours can and should be reproduced with these robots. Here is a list of these phenomena.

The existence of a link between two robots means that the two robots imitate each other. But the links might also mean that there are networks of robots that work together, and networks of dishonest robots can carry out dishonest behaviours that it would be impossible for one single dishonest robot to carry out – organized crime. Or belonging to a network of dishonest robots – criminal networks – might reduce the probability of being discovered and punished. This would be true not only for dishonest behaviours which are recognized and punished by the law but also for behaviours which damage the entire community in ways which are more difficult for the law to define: being part of a network of powerful individuals to obtain public contracts, important positions, and favourable laws – networks of power.

In our robots, when a robot damages another robot, the damaged robot is chosen randomly. We might change this by endowing the robots with knowledge on the level of well-being of the other members of their community, and dishonest robots might use this knowledge to choose which robot to damage. In our robots, a dishonest robot increases its well-being by the same amount whatever the level of well-being of the damaged robot. If damaging a robot with a higher level of well-being – a rich robot – increases the well-being of the dishonest robot more than damaging a robot with a lower level of well-being, we predict that dishonest robots will choose to damage *rich* robots rather than *poor* robots.

A robot which has damaged another robot is discovered and punished with a certain probability which is decided by us and is the same for all robots. We could make the probability that a robot is punished for its socially damaging behaviour dependent

on its level of well-being. Richer robots are less likely to be discovered and punished than poorer robots because they can use their money to avoid being discovered and to avoid punishment if they are discovered, for example, by paying better lawyers. Another thing that we might do is make the probability of being punished dependent on the probability of being denounced by other robots that know about the socially damaging behaviour of the robot. What determines the probability that a robot will denounce another robot for its socially damaging behaviour? Would honest robots be more likely than dishonest robots to denounce a dishonest robot?

Our robots do not live in physical space. They form sub-communities of honest and dishonest robots but only in the sense that the robots that belong to the same sub-community imitate each other's behaviour. But if our robots lived in physical space and they could choose where to live, would honest robots prefer to live near to honest robots and dishonest robots near to dishonest robots? For honest robots, living near to honest robots would reduce the probability to be damaged by dishonest robots and, for dishonest robots, living near to other dishonest robots would reduce the probability of being denounced for their socially damaging behaviour. This would create zones of the environment inhabited by honest robots and zones inhabited by dishonest robots.

Modern technologies tend to make the role of physical space less and less important. Due to advances in the technologies for transporting people, goods, information, and money, human beings are increasingly living in a global world and dishonest behaviours also become globalized. This poses a new problem. Globalization is economic and cultural but, at least for now, political sovereignty remains to the states. Since each state decides its laws and regulations which must be respected by its citizens but not by the citizens of other states, this creates a conflict between globalization and local laws and regulations, and this conflict is especially serious because, as we have said, socially damaging behaviours themselves become globalized. Constructing a plurality of robotic states with different legal systems and a variety of economic and cultural relations among them might be of help to understand these problems and, perhaps, to find possible solutions.

Another important phenomenon is that human societies change and these changes generate new types of socially damaging behaviours. This is especially true in today's societies because today's societies change very rapidly and they become more complex. Science and technology, the financial economy, communication and marketing in their various and multiplying forms, and the digital technologies have an increasing impact on human life and on human societies. This impact both benefits and damages human beings but human societies are better prepared to see the benefits than the damages. Constructing robotic societies which have science and technology, financial economies, modern communication and marketing techniques, and digital technologies can help human beings to identify their socially damaging impact and to find how to reduce this impact without losing their benefits.

Today's societies change also because with their behaviour human beings increasingly change the environment in ways which are socially damaging. An example is polluting the environment. Polluting the environment is the result of the behaviour of everyone and, for this type of socially damaging behaviours, there is no distinction between honest and dishonest people. Our next robots reproduce this behaviour in a very simplified form.

The robots are our basic robots that live in an environment containing randomly distributed food tokens that the robots must eat to remain alive and have offspring. The robots must move in the environment to find and eat the food tokens and, obviously, moving faster allows them to eat more. And, in fact, we find that after a certain number of generations the robots move very fast and they eat many food tokens. Now we construct a second population of robots which, when they move in the environment, they pollute the environment, and living in a polluted environment decreases the robots' survival and reproductive chances. A robot's fitness is not the number of food tokens eaten but the number of food tokens eaten minus the speed with which the robot moves in the environment. These robots have two conflicting problems to solve: moving as fast as possible to eat more and not moving too fast in order not to pollute the environment. If each robot lives alone in its environment, the robots evolve a behaviour which takes into account both problems: a robot moves sufficiently fast to eat but not too fast to avoid polluting the environment (see Figure 9a). But if the robots live together in the same environment, the situation changes. All the robots tend

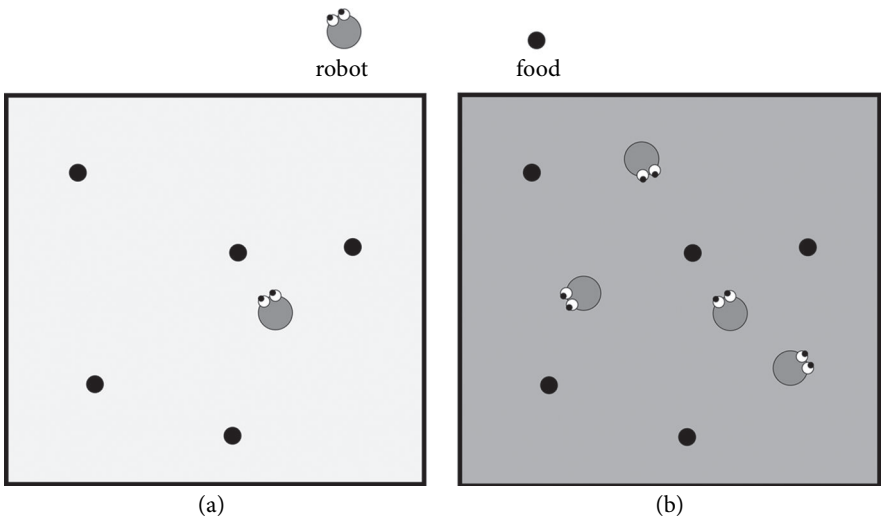


Figure 9. The environment of a robot that lives alone is not very polluted (light grey) because the robot tries not to pollute it (a). If many robots live together the environment is much more polluted (dark grey) because the robots do not care not to pollute it (b).

to move very fast and the environment becomes seriously polluted. This is due to the fact that not only many robots live together in the same environment but all the robots move very fast. The robots move very fast because each robot "knows" that, even if it moves less fast and, therefore, eats less, the other robots will move faster to eat more (see Figure 9b). The problem can be even more serious. In our robots, each successive generation of robots begins its life in a clean environment. But if, more realistically, each new generation of robots inherits the environment of the preceding generation, the robots live in an increasingly polluted environment. This is a particularly intriguing problem because future generations still do not exist and, therefore, they cannot defend their interests.

Other phenomena linked to socially damaging behaviours are of a different nature but they are equally interesting and should be reproduced with robots. Socially damaging behaviours are contained by punishing them. Punishment is a very important factor in the behaviour of all animals. If some behaviour is followed by an event which decreases the probability that the behaviour will be executed in the future, the event is a punishment and, in Chapter 3, we have described robots that learn based on punishment. But socially damaging behaviours raise new questions. Is being punished by "nature" – like touching a very hot object with one's finger – different from being punished by another individual? Is being punished by the state different from being punished by an individual? Is being punished in the presence of other individuals the same as being punished in the absence of other individuals?

Another problem is that punishment may have different effects on socially damaging behaviour as a function of whether the socially damaging behaviour was "willed" or "not willed". In Chapter 5 we have described robots which do not only do X but also *want* to do X but many problems remain unsolved. If doing X is followed by punishment, what are the consequences of punishment for robots that only did X and for robots that did not only do X but also *wanted* to do X? At which age human beings become able to want to do X? Are there pathologies that cause human beings to become unable to want to do X so that, like nonhuman animals, they can only do X?

Another important question concerns what behaviours are socially damaging. Our dishonest robots directly damage other robots but, for human beings, socially damaging behaviours go well beyond behaviours that directly damage others. Human societies can only exist if they function according to rules because, only if they function according to rules, the members of the society can coordinate their behaviours and what happens in the society can be predicted and relied on by its members. This is why, for human beings, socially damaging behaviours are not only behaviours which directly damage others but also behaviours which do not respect the rules of the society. Can we construct robots that can be said to have "rules of behaviour"? How are rules of behaviour different from culturally shared behaviours? Can only robots that predict the consequences of their planned actions and evaluate these consequences before executing the actions, have rules of behaviour? Are language and talking to oneself also

necessary? Is being punished not by a single other robot but by the entire community of robots, or by a representative of the entire community, necessary? Why are some rules of behaviours written? In what types of robotic societies are rules of behaviour written?

A final class of phenomena that should be reproduced by a robotic legal science are the differences among human societies in how they deal with socially damaging behaviours and the historical changes in legal systems. A robotic legal science should construct different robotic societies with different cultures and different histories and compare how socially damaging behaviours are dealt with in these different societies and how legal systems have changed during history. The state has the role of punishing socially damaging behaviours but human societies have, or have had, different types of states and different systems for punishing socially damaging behaviours. By definition, the state has more power than the individuals which make up the community, and this is why the state can punish the individuals which damage others. But the state can be a cast, a class, a lineage, and the rules that apply to the other members of the community may not apply to the members of the cast, class, or lineage. It is only in modern human societies that the "law" applies to all the members of the community. Therefore, we must construct robotic societies in which the law applies to all the members of the community and robotic societies in which this is not true, and see how legal systems correlate with the economic, political, and religious institutions of different robotic societies.

In modern societies it is the state which has the task of containing socially damaging behaviours by discovering and punishing them but, as we have seen, for the state it is not so easy to accomplish this task because socially damaging behaviours can be very rewarding and it may be difficult to discover and punish them. This explains why human societies have two other mechanisms for containing socially damaging behaviours. One mechanism is reciprocity and reputation. If an individual does something which damages another individual, the other individual may respond by doing, immediately or in some future occasion, something which punishes the first individual, including simply refusing to interact with the individual – which, for extremely social animals such as human beings, can be very punishing and may discourage socially damaging behaviours. This is reciprocity. And then there is reputation. If other members of the community are informed that one individual has damaged another individual, they also may punish the individual even if the individual has not damaged them. If the disadvantages produced by these social punishments are greater than the advantages produced by socially damaging behaviours, an individual will avoid behaving in ways that damage others.

The other mechanism for containing socially damaging behaviours is religious or lay morality. Morality is self-punishment. An individual who does something which damages another individual punishes himself or herself for his or her behaviour, and this self-punishment reduces the probability that the individual will do things that damage others in the future. And they can punish themselves not only for *having*

done something that damage others but also for *thinking of doing* something that damage others. Morality in human beings may have a direct genetic basis or what is genetically inherited may not be morality itself but only a predisposition to learn from others how to behave (see Chapter 8). Parents use this predisposition to teach their children and social and religious institutions to teach everyone not to behave in ways that damage others.

Another reason why human societies need these two additional mechanisms for containing socially damaging behaviours is that the state punishes certain types of socially damaging behaviours but not all types. An individual can kill or rob another individual, and these behaviours are punished by the state, but an individual can also damage the social image of another individual, or he or she may not help an individual who needs helps, or he or she may do things which cause psychological pain in another individual. These are all socially damaging behaviours which are generally not punished by the state and, therefore, they require social reputation and morality.

But the state remains the most important mechanism for containing socially damaging behaviours. Among nonhuman animals, an individual may avoid damaging another individual because the other individual is stronger or more intelligent and it would respond by punishing the first individual. Human beings are different. As we have already said, if individual A does something which damages individual B, B may not be aware of the damage, or he/she may not know that A is the author of the damage, or he/she may be unable to punish A because he/she is less strong or less intelligent than A. This is why having a state with a "chief" that discovers and punishes socially damaging behaviours is so important and why, in the large human communities of the last 5000 years, the state has had an indispensable role in containing socially damaging behaviours. In fact, discovering and punishing socially damaging behaviours is one of the most important "goods" which are produced by the state and distributed to all its citizens. (For robotic states, see Chapter 10 on political robotics.)

Today, the state has become even more important for discovering and punishing socially damaging behaviours because Western culture reduces the power of self-punishment – in both its religious and lay forms – as a mechanism for containing socially damaging behaviours. Religion as a way of living is becoming less and less important in Western culture and the reduced role of the family, economic and cultural reasons, and the fact that an increasing number of people live in cities, lead to forms of extreme individualism in which the consequences of one's behaviour for others are less taken into consideration, unless they are punished by the state. And also the way in which the state contains socially damaging behaviour has changed. In the past the state had "authority", and it could contain socially damaging behaviours without punishing them because the citizens of the state respected its authority. Today, the state has less authority and, therefore, discovering and punishing socially damaging behaviours is becoming the only way for human societies to contain these behaviours.

5. Why live together: Groups as information centres

If living together implies the possibility to be damaged by one's conspecifics, what are the benefits of living together which explain why all human beings do not live alone but they live together with other human beings? Of course, there are many advantages of living together. Living together allows an individual to help and be helped by another individual which has his/her same genes, to learn from other individuals, to specialize in the production of one type of good and obtain the other goods from other individuals through the exchange of goods, to create organizations of individuals that produce goods which no single individual would be able to produce. (For robots that do these things, see the next four Chapters.) In this Section we describe robots that reproduce one basic advantage of living together: the robots live together because the group provides information about the environment which goes beyond the information that a single robot would be able to procure if it lived alone. Since information about the environment is crucial for surviving in the environment, this may be a sufficient reason for living together.

A group of robots live in an environment in which a predator hardwired by us to do what it does appears from time to time and, if the predator is able to reach a robot, it kills the robot. The robots that reproduce are those which, when the predator appears, run away and avoid being killed by the predator. The robots have visual sensors that tell them when the predator appears but, in many cases, they see the predator when it is too late. The predator is very fast and it may reach a robot before the robot is able to see it and run away. But the robots do not see only the predator but they also see the other robots and they see when another robot runs away. This is why for these robots living together can be advantageous: the robots run away not only when they see the predator but also when they see another robot run away. (The behaviour of running away is hardwired by us.)

Since a robot can see another robot run away only if it is physically near to the other robot, for a robot it is adaptive to remain in proximity to other robots so that it can see the other robots run away when *they* see the predator. And, in fact, this is the behaviour which emerges in a succession of generations. The robots of the first generation have random connection weights in their neural network and they are not very good at responding to the visual information telling them where are the other robots by approaching and remaining near to the other robots. Therefore, at the beginning each robot tends to live alone. But since living near to other robots reduces the chances of being killed by the predator, the behaviour of the robots changes in a succession of generations and at the end the robots tend to remain in proximity to one another and to form smaller or larger groups of robots. And, if a robot runs away because it has seen the predator or it has seen another robot run away, the robot tends to come back and to remain near to the other robots (see Figure 10).

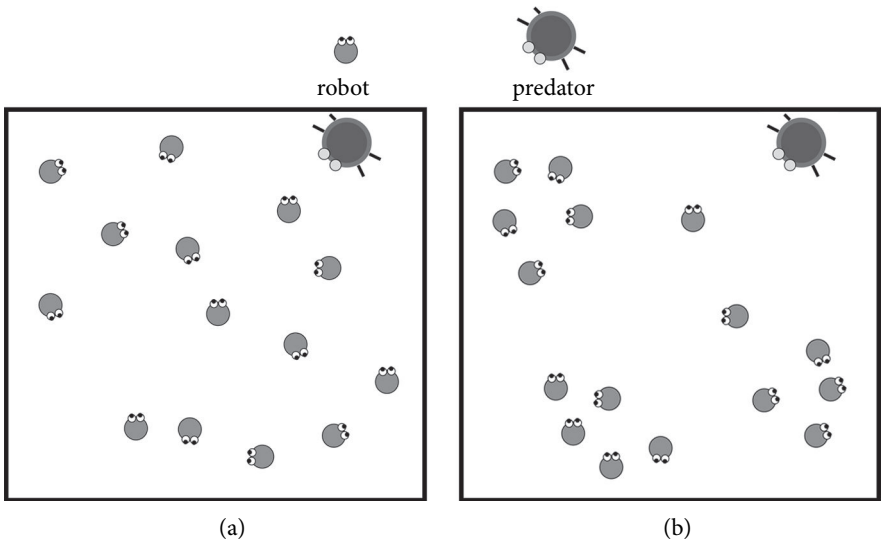


Figure 10. Initially, the robots leave alone (a). At the end, the robots live in groups because, when they see another robot of their group flee away, they know that the predator is approaching and they can also flee away and avoid being killed by the predator (b).

These robots demonstrate, in an extremely simplified form, an important advantage of living together. Groups of individuals which live together function as "information centres" for all the members of the group. All animals need information about the environment in which they live. But the sensory organs and, more generally, the experiences of an individual inevitably allow the individual to collect only a limited quantity of information about the environment. Living together solves this problem. The group functions as an information centre which collects more information about the environment and makes this information available to all the members of the group. This is why many nonhuman animals and all human beings live in groups.

A robot informs the other robots of the arrival of the predator not because this is advantageous for the sender or the receiver of the information but because the information is an automatic by-product of the behaviour of running away from the predator – which *is* a behaviour which is advantageous for the sender of the information. This type of "involuntary" communication does exist in real animals and even in human beings. What is evolved is not the behaviour of informing the other robots of the arrival of the predator but the behaviour of responding appropriately to the information communicated by the other robots. The robots do not "speak" but they "understand". (On speaking and understanding, see Chapter 4 dedicated to language.)

But the behaviour of communicating to the other members of the group the information which is available to one member of the group can be an evolved signalling behaviour which is adaptive for both the sender and the receiver of the signal. In

Chapter 9 we describe robots which are predators rather than prey. They live in an environment with a single very large prey and to attack and kill the prey it is necessary that a sufficient number of robots are near to the prey at the same time. When one robot sees the prey it communicates where is they prey to the other robots and the other robots use this information to approach the prey. In this scenario too, the group functions as an "information centre" but both the behaviour of communicating information to the other members of the group and the behaviour of responding appropriately to this information are evolved behaviours.

There is another lesson which we can derive from the robots that run away when they see another robot run away and it concerns the self-reinforcing character of collective behaviours. For our robots it is advantageous to be near other robots but the advantage increases with the size of the group. Even living in pairs is better than living alone but, if a robot lives in a large group, this makes it more probable that the robot will be informed of the arrival of the predator because it is more probable that one member of the group will see the predator when the predator arrives. To test this idea we have added to the robots' neural network visual neurons that inform them of the *size* of the different groups of robots which are around them and the robots use this information to evolve the behaviour of approaching and joining not *any* group of robots but the *largest* of these groups. This implies that there may be an intrinsic tendency of groups to become larger, and this tendency may be one of the factors which explain why, in human history, human groups have become progressively larger.

In this Section we have described robots that reproduce one advantage of living in groups: groups function as information centres for the members of the group. Our robots resemble animals rather than human beings and human groups are much more effective as information centres than animal groups because human beings have language and, with language, each member of the group can easily transmit to the other members of the group what he or she knows about the environment, But, as we have said, living together has many other advantages beyond sharing knowledge about the environment, and in the other chapters of the book we describe robots that reproduce some of these other advantages.

6. Living in small communities and living in large communities

Human beings can live in smaller or larger communities and one thing which characterizes human history is the progressive increase in the size of human communities. For millennia human beings have lived in a great number of very small communities but then their communities have become fewer in number and larger in size – and, today, globalization is leading to the emergence of a single community which includes all human beings who live on the Earth. Why? The phenomenon has many possible causes and our next robots reproduce two of these causes.

The robots live in an environment with a certain number of randomly distributed food tokens that the robots must eat to remain alive and reproduce. The environment is a seasonal one. At the beginning of each season new food appears in the environment but at the end of the season the robots have eaten most of the food and, therefore, they have problems at remaining alive until the next season begins. As always, the robots of the initial generation have random connection weights in their neural network, they are not very good at eating the food tokens and, therefore, their number rapidly decreases. Then the robots progressively acquire the capacity to approach and eat the food tokens and their number becomes larger until it reaches a value which corresponds to the carrying capacity of the environment – the maximum number of robots that can live in that particular environment. At this point we create two identical copies of the environment and we let half of the robots live in one copy of the environment and the other half in the other copy. The robots that live in one environment all belong to a single large community whose territory is the entire environment and, therefore, the robots can go everywhere searching for food. The robots that live in the other environment are divided into small communities each living in its own limited territory and the robots of one community can only search for food inside their territory but they cannot go into the territories of the other communities. What happens to these two different populations of robots? The results are shown in Figure 11. The number of robots that belong to the same large community and can go everywhere in the environment remains the same and it becomes even slightly larger. The robots of the second population, those that live in separate communities each with its small territory, become extinct.

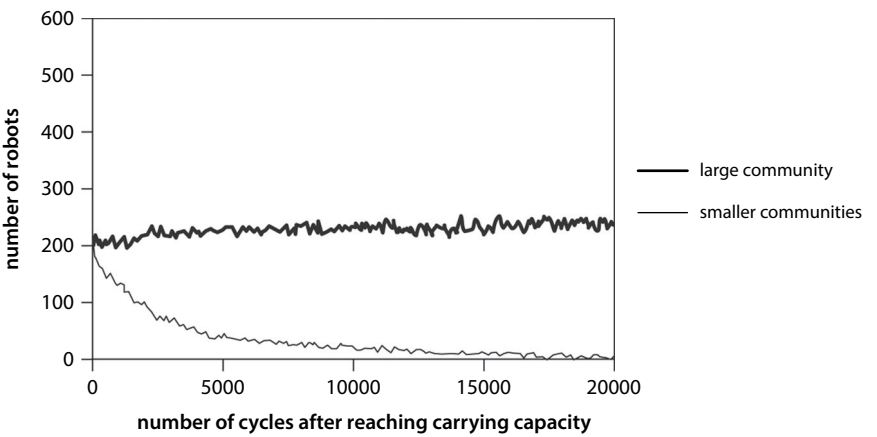


Figure 11. The number of robots living in a single large community with a large territory remains the same while the robots that are divided into many small communities each with its restricted territory become extinct.

Why this result? There are two answers to this question. Since food is randomly distributed in the environment and the environment is a seasonal one, there may be periods in which there is very little food in one particular zone of the environment and more food in another zone. The robots that form a single large community can move from one zone of the environment with little food to other zones with more food – and, in this way, they can survive. But this is not true for the robots that live in small communities with small territories. If there is little food in their small territory, there is not much that the robots can do and this explains why they become extinct.

The other answer is genetic. When a robot is born, the robot is placed in the same territory of its parent and this implies that the robots that belong to the same community and live in the same territory are genetically related. This has an important consequence. For the robots that live in the single large community, a robot which is good at finding food has a long life and many offspring while this is not true for another robot which is unable to find food. Therefore, for these robots the mechanism of selective reproduction functions very effectively because the best robots are those that have offspring and transmit their genes to the next generation. This is not true for the robots that live in a small community with a small territory. Here what is important for reproduction is not so much the robots' capacity to find food but how much food there happens to be in the small territory of the robots. Therefore, the robots that have offspring are not necessarily the robots that are better able to find food but those that happen to live in a territory which is, temporarily, rich in food. With these robots the mechanism of the selective reproduction of the best robots does not function well and the robots' capacity to find food decreases until they become extinct.

So, living in large communities with large territories is advantageous and this may explain why human communities and their territories have increased in size in the course of human history. In Chapter 8 we will see that this is also true for robots that create technological artefacts. The technological artefacts of large communities of robots are better than the technological artefacts of small communities because living in large communities increases the range of possible artefacts among which a robot can choose which artefact to reproduce.

7. The social environment is very different from the non-social environment

In this Chapter we have looked at sociality from a distance, without considering what actually happens when a robot interacts with another robot and what makes social behaviour different from nonsocial behaviour. In this last section we look at sociality more closely. We list some of the things that make social behaviour different from nonsocial behaviour and, therefore, what future social robots must be able to do.

Human beings are very social animals. They obtain most of what they need or want from other human beings and, therefore, their environment is mostly a social environment and their interactions with other human beings are very different from their interactions with inanimate objects or plants. Animals of other species, for example dogs, are also part of the environment of human beings, and they are interesting because human beings interact with them in ways which have both similarities and differences with how they interact with other human beings.

The most important difference between the social environment and the environment of inanimate objects and plants – and, to a large extent, also animals of other species – is that, while to obtain what he or she wants from a plant or from an inanimate object, a human being must simply act physically on the plant or inanimate object, to obtain what he or she wants from another human being, he or she must go through the brain of the other human being. Our robots eat the food tokens or they use the food-producing tools to produce new food but a robot cannot obtain what it wants from another robot by simply acting physically on the other robot as it acts physically on the food tokens or the food-producing tools. The other robot must *behave* in ways that allow the first robot to obtain what it wants. And, since the behaviour of the other robot is under the control of its brain, the first robot must influence the other robot's brain so that the other robot does what the first robot wants it to do. Obviously, to influence the other robot's brain, the robot must act physically because acting physically is the only thing that robots can do but what the robot must do is simply cause the arrival of specific sensory inputs to the other robot's brain and, in this way, influence the other robot's behaviour. What must these sensory inputs do? The answer is: they must change the other robot's motivations. As we have seen in Chapter 2, a robot's behaviour depends mostly on its motivations. Robots – and animals and human beings – have many different motivations and what they do is trying to satisfy the motivation which currently has the greatest strength in their brain/body. Therefore, to obtain from another robot what it wants, a robot must change the current strength of the other robot's motivations and add new motivations. This is social behaviour: changing the motivations of others so that they do what one wants them to do. A robot can send all sorts of sensory input to the brain of the other robot to modify the other robot's motivations, in some cases even deceiving and misleading the other robot. The robot can talk to the other robot. It can express emotions with its body, face, and voice. It can modify its external physical appearance by dressing and decorating its body. It can manipulate the environment of the other robot so as to induce the other robot to buy the things that it wants to sell.

An interesting question is if social behaviour requires that the robots "intend" and "understand" what is going on when they interact together. Must the first robot "intend" to do what it does? Must the other robot "understand" what the first robot is doing and why it is doing what it is doing? The answer can be Yes for human beings and No for nonhuman animals and, even for human beings, it can be Yes in some

occasions and No in other occasions. We do not try to answer this question here but we only say that robots that "intend" and "understand" are robots that predict and evaluate the consequences of their behaviours. (We have described robots that predict in Chapter 5.) But there is an important difference between making predictions on the inanimate environment and making predictions on the social environment. To be able to predict what will happen to an inanimate object, a robot must take into considerations only the sensory inputs which arrive to its sensory organs from the object and its actions on the inanimate object. To predict what another robot will do, a robot must take into consideration the sensory inputs which arrive to the *other robot's* sensory organs, what the other robot knows, what are the motivations of the other robot, what is the motivation which is currently controlling the other robot's behaviour, what is the current emotional state of the other robot. And there is another problem. All inanimate objects obey the same "laws" and these are relatively simple "laws" which, even for the ordinary robot, are not so difficult to discover. The behaviour of the other robot does not only obey more complex "laws" but, since each robot is different from all other robots, its behaviour cannot be predicted by only using "laws". (For inter-individual differences among robots, see Chapter 12.)

Living socially is living in an environment in which the robots obtain most of what they want not from nature but from other robots. But why should the other robots give a robot what it wants? The answer is that in most cases the other robots give the robot what it wants in exchange for something else. This can be an informal exchange of goods between robots that know each other or it can be an economic exchange of goods between extraneous robots in which everything is explicitly defined: what A gives B and what B gives A and when the exchange occurs. (On the economic exchange of goods, see Chapter 8 dedicated to robotic economies.) In some cases there may seem to be no exchange: robot A gives something to robot B in exchange for nothing. But in reality in this case too there is an exchange of goods: robot A gives something to robot B in exchange for not being punished by robot B. This is linked to an important social phenomenon: power. Robot B has power on robot A if it can make robot A do what robot A does not want to do in exchange for not being punished by robot B. Power can be an informal power but human societies also have socially recognized or institutional power, and informal power and formal power are at the basis of human beings' social life and of the life of their societies.

But there is something more – and deeper – which distinguishes other robots from inanimate objects as parts of a robot's environment. While inanimate objects are "strange" entities which are very different from the robot, other robots *are like* the robot. They have the same external shape of the robot (more or less), they behave in the same way (more or less), they have the same motivations and emotional states (more or less). Therefore, in so far as the robot knows itself, the robot also knows the robots with which it interacts and the robot can project what it knows about itself on the other robots. If the robot knows how it will react to some event, the robot can predict how

another robot will react to the same event – where reacting can mean both behaviours and emotional states. Of course, the other robots are not identical to the robot and they are not identical to each other. Therefore, the robot must learn how each individual robot reacts to this or that event but, at some basic level and by default, the robot can generalize from its own reactions to the reactions of the other robots. And the process is symmetrical. By knowing how other robots behave and what they think and feel, the robot can also know how to behave and what to think and feel.

Living in a social environment has still other consequences because social robots live in two worlds: a public world and a private world. Some of the sensory inputs that arrive to a robot's sensory organs also arrive to the sensory organs of other nearby robots. For example, when the robot sees an object, the robot knows that another nearby robot can also see the object and, therefore, the robot can predict which reaction the object will provoke in the other robot – provided it knows the motivation which is currently controlling the behaviour of the other robot. But other sensory inputs work differently. When a robot feels hungry or is in a particular emotional state or it talks to itself, the robot cannot assume that these sensory inputs are sensory inputs also for another robot. The world that the robot knows through the first class of sensory inputs is the public world. The world that the robot knows through the second class of sensory inputs is its private world.

Human beings are different from nonhuman animals from this point of view, and this difference should be reproduced by constructing both animal and human robots. It is a fact of physical reality that some physical events cause sensory inputs for only one robot while other physical events cause sensory inputs for all the robots which are physically near to one another, and this is true for all types of robots: animal robots and human robots, social robots and nonsocial robots. However, only human beings with their sophisticated prediction abilities actually live in both a public world and a private world. When a human robot tries to predict the behaviour of another human robot, the robot notices that it can use some of its own sensory inputs – for example, the sensory inputs from a physical object – to predict the behaviour of the other robot, while other sensory inputs – for example, the sensory inputs that tell the robot that it is hungry or in some particular emotional state or the self-generated sensory inputs that constitute the robot's mental life – are useless from this point of view.

This implies that, while a robot finds it relatively easy to predict what happens in the environment of inanimate objects – at least on a short-term and local basis – predicting what another robot will do and how it will feel is much more difficult. This is not only because robots are more complex physical entities compared to inanimate objects but because much of what the robot needs to know to predict the behaviour, feelings, and thoughts of the other robot is inaccessible to the robot. And there is still another reason. While the environment of inanimate objects is more or less always the same and the robot can learn to predict all it needs to predict about this environment in the early stages of its life, the social environment is always changing because the

robots which the robot already knows change and the robot always encounters new robots. Therefore, learning to predict what other robots will do and how they will feel and think is a task for the robot's entire life.

This is a problem because living socially requires being able to predict and to count on the behaviour of others. To solve this problem human beings have cultures, rules of behaviour, and written laws, which are all mechanisms for making all the members of a community behave in the same way so that their behaviour becomes more predictable. And this also explains why people like to live with people who behave like them and have the same culture: they can more easily predict and count on their behaviour.

This is also linked to social routinization. In Chapter 3 we have discussed routinization as a form of learning. When a robot is learning some behaviour – for example, to drive a car – the robot's behaviour changes and it becomes progressively routinized, where routinization is reducing the number of sensory inputs which the robot needs to take into consideration for executing the behaviour – the robot can talk with another robot while driving the car. This extends to social routinization. When a group of robots do something together, each robot must respond to the sensory inputs that arrive from the other robots, and routinization becomes social routinization: reducing the sensory inputs from the other robots. Social routinization may be even more important than routinization at the individual level. In socially organized behaviour the number of different inputs that arrive to the sensory organs of a robot is very large because they depend on the behaviour of the other robots and the behaviour of the other robots varies from moment to moment and each robot behaves differently. Therefore, it is not sufficient that each robot behaves in a routinized way to restrict the inputs arriving to *its* sensory organs but it is necessary that all the robots behave in the *same* routinized way so as to reduce the sensory inputs they provide to the other members of the group – as when human beings dance together, take part to social ceremonies, or construct something together.

This has an important consequence for social behaviour and social life more generally. The sensory inputs that an individual's body sends to the individual's brain are inaccessible to another individual but the brain of the other individual can self-generate the equivalent of these sensory inputs. (The brain's neurons that do this are called "mirror neurons".) By seeing or hearing another individual, an individual senses and feels like the other individual. I receive the same proprioceptive input from my immobile arm when I see your arm move. I feel the same emotional state that you feel when I see your face or voice express that emotional state. Knowing what other individuals know and feel plays a central role in all social interactions and in all socially organized behaviour, and it is an important phenomenon that future human robots should be able to reproduce.

There is one last thing that we want to say about the social environment. The social environment is not a homogeneous environment in which everyone interacts with everyone but it is an environment made up of social networks. An individual interacts

with some individuals more than with other individuals and social life can be seen as a network of nodes in which the nodes are individuals, the links between nodes are interactions between pairs of individuals, and a node has more links with some nodes than with other nodes. In Section 4 we have seen how social networks of dishonest robots can reduce the impact of punishment on dishonest behaviour but social networks exist in all forms of social life – family networks, networks of friends, networks of coworkers, networks of social activists, networks of power – and who is one particular individual can be to a large extent known by knowing the social networks of which he or she is a member. This is why human robots must not only be social robots but they must also be nodes in various types of social networks.

We conclude this Chapter by saying that the sociality of human beings is such an important component of their adaptive pattern that it must be inscribed in their genes. From their birth, human beings are particularly responsive to sensory inputs from other individuals, and their genetically inherited sociality can explain many aspects of their behaviour such as their tendency to see nature as made up of entities which are like them and want to communicate with them, to respond to work of arts as "speaking" to them and, more generally, to attribute "meaning" to reality – where "meaning" must be "meant" by someone else – including oneself. This may be at the origin of religion, art, and even philosophy and it is an important reason for constructing robots that have an evolved complex sociality. (For robots that have art and religion, see Chapter 11.)

CHAPTER 7

Robotic families

ME knows that for all organisms the most basic requirement is to leave one's genes to future generations. This requires not to die too soon and to have offspring but for many animals to remain alive and to have offspring are not sufficient. Especially for human beings, if their genes must continue to be part of the genetic pool of the population, for a long period after birth parents must help their offspring to remain alive so that they also will be able to reproduce. To help their offspring to remain alive, parents must stay near to their offspring to give them what they need, and offspring must stay near to their parents to receive from them what they need. This explains why human beings live in families and why families are so important for the life of human beings.

Another thing which ME notices about human beings is that, like many other animals, human beings have two sexes, male and female, and males and females play different roles in reproduction. ME is particularly interested in one difference. After mating with a male, for a certain period of time a female becomes nonreproductive while males can mate reproductively whenever they want – provided they find a reproductive female. This has many consequences for the life of human beings and ME decides to construct male and female robots to better understand these consequences.

ME also notices that groups of genetically related human beings live together in places called homes and these homes tend to be near to one another. Groups of homes have become increasingly larger and, especially today, this is having an increasing influence on the life of human beings.

1. Genetic families and social families

Human beings are not just human beings. They are males and females, they are infants, children, adolescents, adults, and old people, they are genetically related or genetically extraneous to other individuals – and, for both biological and cultural reasons, their behaviour reflects these characteristics. Females behave differently from males. An individual at one age behaves differently from the same individual at another age. Individuals behave differently towards their kin than towards nonkin. And kin-related individuals tend to live together in families.

Current robots do not have any of these characteristics. There are no female and male robots – although some robots do have the external appearance of a male or a female – and robots do not have one specific age – except for some robots that have the external appearance of human infants. Robots are not genetically related to other robots and they do not behave differently towards genetically related and genetically unrelated robots. They are just robots. But, if robots must be theories of the behaviour of real animals, this is another aspect of robotics that will have to change. We must construct robots that are males or females, robots that have different ages, and robots that have genetic relations and behave in ways which reflect these genetic relations. If our goal is to construct robots to understand human beings, our robots must be born from a female robot which is beyond a certain minimum age (reproductive age for females) after sexual intercourse with a male robot and after the equivalent of nine months of pregnancy. Infant robots must survive only if their mother and father provide them with food. Female and male robots that form a reproductive pair must live together and with their young offspring and they must help each other in various ways. At some age females must cease to be reproductive – menopause – and the robots must die at different ages. These demographic characteristics must influence the robots' behaviour and, if culture and medical technologies change some of these characteristics, the robots must reproduce these changes and their consequences.

This Chapter is dedicated to robotic families. Human beings live in families and, therefore, human robots must live in families. But what are families? We must distinguish between genetic families and social families. Genetic families are sets of individuals which have similar genes because they have the same ancestors but, since there are various degrees of genetic similarity, the notion of a genetic family is a relative one. For example, siblings are more genetically related than cousins and first cousins are more genetically related than second cousins. In fact, a genetic family has no intrinsic limits and the entire human species can be considered as a genetic family – the human family – because all human beings descend from the same ancestors.

Given this definition, genetic families exist in all animals. In contrast, social families exist only in some animals, those which have a genetically-based form of sociality. A social family is a group of genetically related individuals which live and interact together. While the notion of genetic family is purely biological, the notion of social family is both biological and social. The distinction is important because behaviour towards genetically related individuals tends to be different from behaviour towards genetically unrelated individuals and human robots must behave differently towards genetically related and genetically unrelated robots. And, if culture changes what is – or is considered to be – a family, human robots must reproduce these changes and their consequences.

As we have seen in the preceding chapter, social interaction implies – or has implied in the past – spatial proximity. Spatial proximity is more important for interactions among genetically related individuals than for interactions among genetically

unrelated individuals and, in fact, spatial proximity tends to co-vary with genetic relatedness: two individuals are more likely to live in close spatial proximity the more genetically related they are. For example, parents take care of their dependent offspring and taking care requires spatial proximity. Closely related individuals tend to live in the same places called "homes" and they have "domestic" life and activities. Homes serve for protection from adverse conditions and bad events but they also allow the members of a family to be near to one another by living in the same home.

Families as a social phenomenon have many aspects that must be reproduced by a social, political, economic, and historical robotics. Ancient human societies were family-level societies, that is, small societies which coincided with extended families made up of both consanguineous and non-consanguineous relatives. Then, human societies have increased in size, they have included a plurality of families, and societies have become large groups of genetically unrelated individuals. The size of human societies has become larger and larger and, today, there is a tendency towards the emergence of a single society which includes all human beings who live on the Earth. Politically, families are – or were – related to political power because the "chiefs" of a society were members of the most powerful and richest families. Economically, families are associated with economic inheritance and differences of wealth. In other chapters of the book we will describe robots that reproduce some of these political and economic phenomena but in this Chapter we limit ourselves to nuclear families made up of parents and offspring and, in one case, also grandparents, and we try to reproduce with our robots how genetic relatedness influences behaviour. Like our other robots, most of the robots described in this Chapter are unisexual robots and they reproduce nonsexually. Therefore, we talk of mother robots, daughter robots, and grandmother robots. But in the last section we describe male and female robots that reproduce by mating together.

2. Mothers and daughters

In many animals, for a more or less prolonged period after birth, a newborn individual is unable to procure the resources which are necessary for its survival and, therefore, if they want their offspring to remain alive, parents must provide them with the necessary resources. In real animals and, especially, in human beings parents provide a variety of resources to their dependent offspring to keep them alive and well (food, water, protection from all sorts of dangers and negative conditions, knowledge on how to behave, etc.) but our dependent robots only need food to survive and their parents provide them only with food.

The (female) robots live in our basic environment with randomly distributed food tokens which they must eat to remain alive. Periodically, a mother robot generates a daughter robot which, for a certain period of time, is unable to move and, therefore,

to survive she needs to be fed by someone else. Since giving some of her food to her dependent daughter reduces a mother's chances to live longer and to have other daughters, why should the mother feed her dependent daughter instead of eating all the food which she is able to find? The answer is the theory of kin selection. To guarantee the presence of her genes in future generations, what is important for a mother is not the number of her daughters but the number of her daughters that reach adulthood and therefore can themselves have daughters. Since infant robots are unable to move and therefore they would starve and die if left alone, the only chance that mothers have to leave their genes to future generations is to give some of their food to their dependent daughters. If a mother robot were to behave selfishly, without feeding her dependent daughters, the mother robot would eat much and would have a larger number of daughters but these daughters would die before reaching adulthood and, therefore, the mother's genes would disappear from the genetic pool of the population.

The genes of these robots do not only encode the connection weights of their neural network but they also encode how likely is that a mother robot will give some of her food to her dependent daughters. The robots' genes include an "offspring care" gene which has a value than can go from 0 to 1, where 0 means 0% probability of giving a newly found food token to one's infant daughters, and 1 means a probability of 100% of this behaviour. The value of the "offspring gene" of a mother robot is inherited by her daughters with the addition of some random mutations which can either increase or decrease its value and make them more or less altruistic towards their daughters.

We do not reproduce a mother robot's infant daughters and the actual behaviour of feeding one's infant daughters but we only reproduce the effects of this behaviour. When a mother robot finds a food token, she either eats the food token or she divides the food token among her infant daughters. The robots that have an "offspring care" gene of 1 give all the food they are able to find to their daughters – completely altruistic mothers – while the robots with a gene value of 0 give none of their food to their daughters – completely selfish mothers. The value of the "offspring care" gene is assigned randomly to the robots of the first generation and, therefore, the average value of the gene is 0.5 and the average mother has a 50% probability of giving her food to her infant daughters. But, after a certain number of generations, we find that the robots not only have evolved the capacity to find food in the environment but, in addition, the probability that a mother robot gives her food to her infant daughters has increased to about 75%. When a mother finds a food token, three quarters of the times she gives the food to her infant daughters. As predicted by kin selection theory, the robots have evolved a tendency to behave altruistically towards their daughters. They reduce their personal chances of living longer and having more daughters to guarantee the survival of their already existing daughters and, therefore, the continued existence of their genes in future generations.

That this behaviour is limited to genetically related robots is confirmed by other robots which also can give their food to infant robots but are not selective and they give

their food to randomly chosen infant robots, not specifically to their daughters. The genes of these robots includes the same "offspring care" gene of the preceding robots but now, when a mother robot decides to give her food according to the probability specified in her "offspring care" gene, the food goes to a randomly chosen infant robot and not necessarily to the robot's daughters. Like the preceding robots, the initial value of the "offspring care" gene in the population is 0.5, which means that on average the robots of the initial generation give half of their food to infant robots. But now the gene's value, instead of increasing, decreases and, after a certain number of generations, it is close to zero. The evolved robots are selfish: they eat all the food they are able to find and do not give any of their food to infant robots.

And there is another interesting result. If we measure the capacity to find food of the mother robots which are not selective and do not necessarily feed their daughters, we find that they are *less good* at finding food than the mother robots that selectively feed their daughters. Why? The answer is that being selective in feeding one's daughters rather than any randomly chosen dependent robot creates a stronger evolutionary pressure for evolving the capacity to find food and, as a consequence, this capacity reaches higher levels. The mothers which are better at finding food live longer and have more daughters and, since their daughters inherit their connection weights, their daughters will also be good at finding food and they also will generate many daughters. Therefore, by feeding her infant daughters, and not any randomly chosen infant robot, a mother guarantees the survival of robots that, like her, are good at finding food. In contrast, if a mother who is good at finding food but is not selective and it distributes her food among all infant robots, the infant robots that she feeds may not be very good at finding food and, therefore, the ability to find food will have more difficulty emerging in the population. This may have been one of the causes of the greater intelligence of human beings compared to nonhuman animals. The need to feed their children for a long period after birth represents a pressure on human beings to become more intelligent.

So far, infant daughters are purely virtual robots and they become real robots only when all their mothers have died and a new generation begins. This allows mothers to feed their virtual daughters with no requirement of spatial proximity to their daughters. But, in actual reality, to be fed by their mothers the daughter robots must be physically close to her, and this represents a pressure on daughters to stay near to their mothers and to interact with them – which is part of our concept of a social family. Our next robots reproduce this aspect of being a family: infant daughters follow their mothers when their mothers move around in the environment looking for food.

We construct a new population of robots in which, when a daughter is born, the daughter is immediately put in the environment together with her mother and, therefore, mothers and daughters live together in the same environment. The life of a robot is divided into two periods: infancy and adulthood. During infancy a robot does not see the food tokens but she only sees her mother. Therefore, infant robots cannot directly procure the food they need to remain alive. But when a robot becomes an

adult and she has her own daughters, the situation is inverted. The adult robot's visual neurons encode the location of the nearest food tokens but the robot cannot see her infant daughters – we might say, because she is busy finding food. When an adult robot reaches a food token and she has not generated any daughters yet, she eats the food token. But, if the adult robot has one or more infant daughters, she divides the food tokens with her infant daughters. (These robots do not have an "offspring care" gene and mothers automatically give their food to their daughters.) This allows the robot's infant daughters to survive until they reach adulthood when they become able to procure food directly and to reproduce in their turn.

However, these robots have a new problem: mothers can give their food to their infant daughters only if their daughters are *physically near* to them. Since mothers don't see their infant daughters and they move in the environment searching for food, they cannot remain near to their daughters and it is their daughters' responsibility to maintain a small distance between them and their mothers. To be fed by their mothers, infant robots must follow their mothers in their displacements in the environment in search of food. As we have said, what is evolved is, first, the capacity of adult robots to respond to sensory input from food by approaching and reaching the food and, second, the capacity of infant robots to respond to visual input from her mother by remaining near to their mother. At the beginning, since the connection weights of the robots' neural networks are random, adult robots are not very good at finding food and, therefore, they die early and generate very few daughters – and, in any case, they have little food to give to their daughters. Similarly, infant robots are unable to follow their mothers which move in the environment searching for food and, therefore, they also die prematurely both because their mother has little food to give them and because they are unable to follow their mother and get the little food that their mother could give them. However, in a succession of generations the connection weights of the robots' neural networks evolve as a result of the selective reproduction of the best robots – which must be good both at following their mother during their infancy and at finding food in order to feed their infant daughters when they become adults – and, at the end, everything has changed. Adult robots have evolved the capacity to find the food tokens. Therefore, they live longer, have more daughters, and they are able to feed their infant daughters. Infant robots have evolved the capacity to follow their mothers so that they can receive food from their mothers, and, in this manner, they can reach adulthood and reproduce in their turn (see Figure 1).

In these robots, each mother robot lives alone in the environment together with her daughters. Hence, an infant robot must follow the only adult robot existing in the environment and this capacity is evolved and encoded in its genes. But imagine that many mother robots live together with their offspring in the same environment. This requires that an infant robot learns who her mother is, and this type of learning is called "imprinting". Infant robots must learn to follow the first adult robot which they see when they are born because this adult robot is very likely to be their mother and only

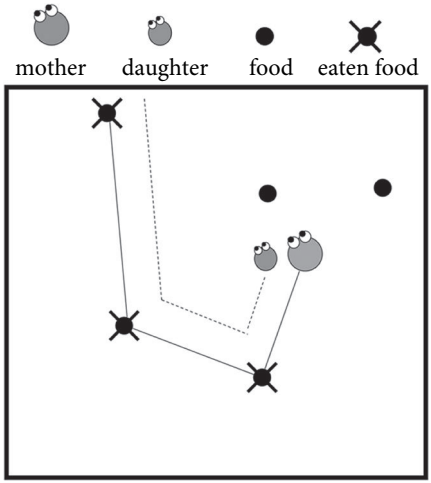


Figure 1. An infant robot follows her mother while she is looking for food in the environment so as to get food from her.

their mother is disposed to feed them. We have described robots that get "imprinted" with their mother in Chapter 3.

The robots of this Section possess two behaviours which are both due to genetic relatedness: mothers selectively feed their daughters and not other infant robots and daughters follow their mother and not other adult robots to be fed by them. These behaviours are the basis of families as groups of genetically related individuals who live and interact together. In the next Section we enlarge our perspective and we introduce grandmother robots.

3. Grandmothers

At a certain age females cease to be reproductive and this phenomenon is called menopause. In most nonhuman animals, including nonhuman primates, females tend to die not much time after they have ceased to be reproductive. In contrast, human females live after menopause for a period which may last several decades. What is the cause of this difference? Why women have a long life after menopause if they cannot have children after menopause? One hypothesis which has been advanced to explain the long life after menopause of human females is the "grandmother hypothesis". As we have seen with our preceding robots, mothers provide their daughters with the food they need to remain alive until their daughters become able to directly procure their food, and mothers tend to do so because giving food to their dependent daughters increases the probability that their genes will be transmitted to future generations.

According to the "grandmother hypothesis", women continue to live after menopause because in this way they can help their adult daughters to provide their grandchildren with the necessary food – and other goods such as protection from dangers and adverse conditions. Living after menopause has no direct adaptive value because after menopause a woman cannot have children but it has indirect adaptive value because post-menopausal grandmothers can help their grandchildren to remain alive.

Our next robots are a test of the "grandmother hypothesis". In these robots, the behaviour of finding food is not evolved but is hardwired and all the robots are equally good at finding the food tokens. A robot simply approaches and captures the nearest food token following the shortest path. However, it can happen that a robot finds itself in one part of the environment without little food and the robot may die because it cannot eat. The robots are all females and reproduction is nonsexual: at regular intervals an adult female robot generates one daughter robot. The behaviour of looking for food is hardwired and not evolved but something else does evolve in these robots. The robots have a gene, an "age of natural death" gene, which they inherit from their mother, and this gene specifies the expected time of the robot's death. All the robots necessarily die at some maximum age but a robot can die before that age not only because the robot cannot eat for lack of food but also because its "age of natural death" gene specifies an earlier time of death. A robot's daughters inherit the "age of natural death" gene of their mother with random mutations which can increase or decreases the expected length of their life.

The lifetime of a robot is divided into three stages: infancy, adulthood, and old age. During infancy, the robots are unable to move and they survive only if their mother and/or grandmother provide them with food. Adult females can move and find food and they reproduce by generating a maximum number of three daughters. When the energy of their infant daughters reaches a certain low threshold, mothers automatically give some of their food to them. Old females also move and find food but they do not reproduce because females have menopause at the moment in which they reach old age. These robots are more realistic than our usual robots because they have a "life history" – their life is a succession of different stages – and there is a greater generational overlap because infant robots live together with their mothers and grandmothers (see Figure 2).

While mothers automatically give their food to their daughters which need food, grandmothers are different. We compare two populations of robots. In one population grandmothers are altruistic and they give their food to their granddaughters. In the other population grandmothers are selfish and they keep all their food for themselves, ignoring their granddaughters. What are the consequences of the behaviour of grandmothers towards their granddaughters for the age of expected death which is specified in the robots' "age of natural death" gene? Notice that the "age of natural death" gene has an effect only on old females because only old females can die for natural death. Infant and adult females can only die of starvation while old females can die both because of starvation or because their "age of natural death" gene specifies an earlier age of death.

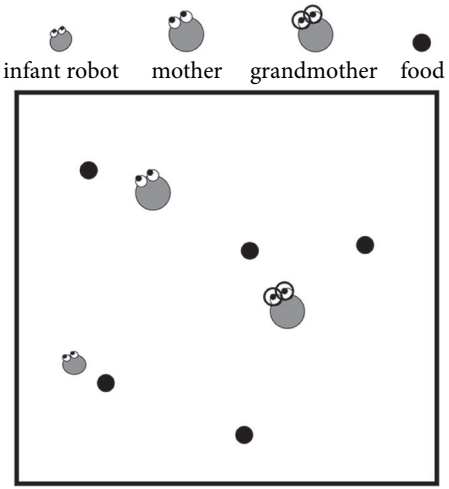


Figure 2. An infant robot, her mother, and her grandmother.

The robots of the initial generation have a value for their "age of natural death" gene which is randomly chosen between the beginning and the end of old age. The gene is inherited with random mutations that can anticipate or postpone a daughter's expected age of death compared to the mother's expected age of death and what we are interested is what is, at the end of evolution, the expected age of death specified in the genotype of the robots which have altruist or selfish grandmothers.

The results are that, in the population in which grandmothers are selfish, the robots do not live much after menopause. The age of natural death encoded in the genotype of selfish grandmothers tends to be only 11 cycles after menopause. In contrast, if grandmothers are altruistic and give their food to their granddaughters, the average female lives a considerable period of time after she has had menopause: 92 cycles. This is in agreement with the "grandmother hypothesis". Human females tend to live longer after they have ceased to be reproductive because in this manner they can help their children to keep their grandchildren alive. And populations with altruistic grandmothers have a larger size than populations with selfish grandmothers: 459 vs. 353 individuals. Altruistic grandmothers live longer to help their granddaughters to remain alive, and this leads to an increase in the size of the population.

We now ask another question. As we have said, old females can die for two different reasons: they can die for natural causes – because their inherited genotype specifies a certain age of death – or they can die because they do not eat and their energy goes to zero. Is the cause of death the same for altruistic grandmothers and for selfish grandmothers? The answer is No. In the populations in which grandmothers are altruistic, old females are more likely to die because they are unable to find food while in the population with selfish grandmothers they are more likely to die because of natural causes (their "age of natural death"). Altruistic grandmothers live longer but, since they

give their food to their granddaughters, they may die of starvation. Selfish grandmothers tend to die earlier not because they do not eat but because there is no pressure on them to live longer and their genotype specifies an earlier expected age of death. The different demography of the two populations of robots is manifested in the fact that a larger number of granddaughters have a living grandmother in the population with altruist grandmothers than in the population with selfish grandmothers.

In these robots, adult females can give birth to a maximum number of three daughters and, in addition, a mother gives birth to all her three daughters in the early part of her reproductive life – concentrated births. Now we construct another population of robots in which a mother gives birth to her three daughters during the entire length of her reproductive period – spaced births – and we compare populations with altruistic grandmothers with populations with selfish grandmothers in these new robots. What is the effect of the spacing of births? The expected age of death encoded in the robots' "age of natural death" gene turns out to be greater with spaced births than with concentrated births but what is surprising is that this is true for both the population with altruistic grandmothers and the population with selfish grandmothers. Is this in contrast with the "grandmother hypothesis"? The "grandmother hypothesis" is the hypothesis that in populations with altruistic grandmothers, but not in populations with selfish grandmothers, the genotype of the population will encode a longer life after females have ceased to be reproductive because in this way they can help their daughters to feed their granddaughters. Our results confirm the "grandmother hypothesis" for populations in which mothers generate their daughters in the early part of their reproductive life but fail to confirm the hypothesis for populations in which mothers generate their daughters during the entire length of their reproductive life. If mothers generate their daughters throughout their reproductive period, the age of death specified in the genotype increases for *both* altruistic and selfish grandmothers, and this is in contrast with the "grandmother hypothesis". Why should grandmothers live so long if they do not help their daughters to feed their granddaughters?

This result is explained by adopting a "mother hypothesis". The "grandmother hypothesis" says that women live after menopause because they help their daughters to keep their granddaughters alive. But if mothers can generate a child towards the end of their reproductive period, the same problem applies to mothers. (And, today, in Western societies mothers tend to generate their (few) children when they are older and closer to the end of their reproductive period.) When they give birth to a child, mothers can be near to menopause and, therefore, they may need to care for their young children *after* they have reached menopause. This implies a selective pressure on *mothers* to live longer after menopause. And the selective pressure on mothers to live longer after menopause applies both to populations in which grandmothers are altruistic and to populations in which grandmothers are selfish because it is a pressure on mothers, not on grandmothers. Mothers are necessarily altruistic. While grandmothers can trust their daughters to take care of their granddaughters, and therefore they can be selfish and ignore their granddaughters, mothers must necessarily be altruistic

towards their daughters because, in populations in which the young are unable to take care of themselves, this is the only possibility for mothers to leave their genes to future generations. Therefore, if mothers can give birth to a daughter towards the end of their reproductive period, there will be a selective pressure on them to live longer after menopause, whatever *their* mothers – the grandmothers – do. This, again, has consequences for the demography of the robots. Population size is larger with spaced births than with concentrated births because the pressure to live longer after menopause is greater on mothers than on grandmothers. But grandmothers remain important. In the population with spaced births, population size continues to be larger with altruistic grandmothers than with selfish grandmothers.

These robots reproduce some very basic phenomena linked to families and to life history as a succession of life stages specified in the genes. Clearly, even these basic phenomena are extremely simplified in our robots. Our mother robots can have a fixed maximum number of daughters (three) and we decide both the spacing of births and whether grandmothers are altruistic or selfish. All these parameters should be encoded in the robots' genes and should evolve autonomously so that we may better understand how they evolve and how they co-evolve and influence each other. And we have adopted a purely biological perspective while, among human beings, these parameters are also influenced by culture and they change with culture. Our robots' environment is also very simple and it contains a constant number of food tokens because any token which is eaten is immediately replaced by a new food token. What would happen in an environment with less food or in an environment in which the quantity of food varies seasonally? We predict that less food or seasonal food would mean a stronger pressure on grandmothers to be altruistic and to live longer after menopause. And a final limitation of our robots is that altruistic grandmothers feed their grandchildren but no one takes care of them when they become too old to be able to find their own food. Do human beings take care of their old parents because they hope to be helped by them and, unlike nonhuman animals, they remain "imprinted" with them for the entire length of their life? (In Chapter 9, we return to this last question.)

4. Sisters

The altruistic behaviour of the robots described so far is addressed to robots of successive generations. Mothers behave altruistically towards their daughters, grandmothers behave altruistically towards their granddaughters. Now we ask: Does a robot behave altruistically towards a kin-related robot of her *same* generation – towards her sisters? We answer this question with our next robots.

The robots live in an environment which contains both food tokens and tokens which are not food but are tools to produce new food – with a sort of simple agriculture. When a robot reaches a food token, it eats the food token which disappears and is not replaced by a new food token. When the robot reaches a food-producing tool,

the tool also disappears but an additional food token appears in a randomly selected location in the environment. (A food-producing tool can be used to produce only one food token and then it must be thrown away.) Eating the food tokens is an adaptive behaviour but also approaching and reaching the food-producing tools is adaptive because it increases the total number of food tokens that the robots can eat.

As always, at the beginning the robots are unable to reach the food tokens and eat them and they are also unable to reach the food-producing tools and produce new food tokens. The robots eat little food and they produce little new food. Then the selective reproduction of the robots that are better at reaching and eating the food tokens leads, in a succession of generations, to an increase in the robots' capacity to eat the existing food. What about the robots' capacity to reach the food-producing tools and to produce new food? If a robot is unable to produce new food, the robot is limited to eating the food which exists naturally in the environment. If the robot is able to produce new food, the robot can eat not only the food which exists naturally in the environment but it can also eat the new food which it has produced. Human beings are different from other animals because in the last 10–12 thousands years they have adopted this more adaptive strategy of producing new food through agriculture and animal breeding rather than just collecting the existing food through hunting and gathering. Therefore, if our aim is to construct human robots, our robots must possess the capacity to produce new food.

Do the robots produce new food? If each robot lives all alone in its environment, without other robots (see Figure 3), the answer is Yes. The robots evolve both the capacity to eat the existing food and the capacity to produce new food. The fitness of the robots (total number of food tokens eaten during a life of fixed length) is higher than the fitness of our basic robots that can only eat the existing food – and this is understandable since the new robots have more food available to them.

However, this is what happens if each robot lives alone in its environment and, therefore, a robot which produces a new food token is guaranteed that no one else will eat the food token it has produced. What happens if many robots live together in the same environment? The food token produced by one robot appears in a randomly chosen position in the environment and not necessarily near to the robot which has produced it. Therefore, the new food token can be eaten not by the robot which has produced it but by another robot which happens to be near to the new food token. (These robots do not own things and they do not have property rights. We describe robots that own things in Chapter 9.) The behaviour of producing new food is altruistic because it reduces the survival/reproductive chances of the food-producing robot – which could have used its time to eat the existing food tokens rather than producing new food tokens – and it increases the survival/reproductive chances of the other robots. So we ask: Does the behaviour of producing new food tokens evolve in these robots?

The answer depends on who are the robots that live together in the environment. The environments contains only four robots but what is important is who are these

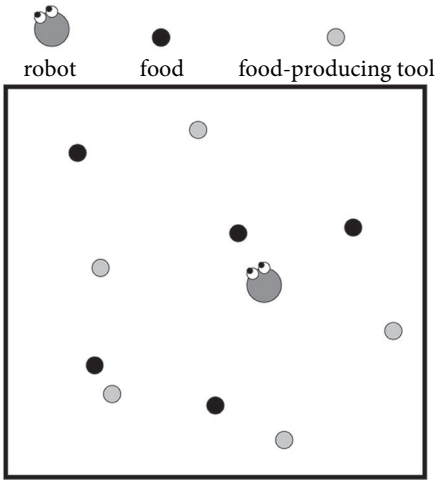


Figure 3. A robot lives alone in an environment with both food tokens and tools for producing new food tokens.

four robots. We evolve two populations of robots. In one population the four robots that live in the same environment are randomly chosen from among all the robots of the same generation. In the other population the four robots that live together are the daughters of the same mother. They are sister robots and, therefore, have the same – or very similar – genes (see Figure 4).

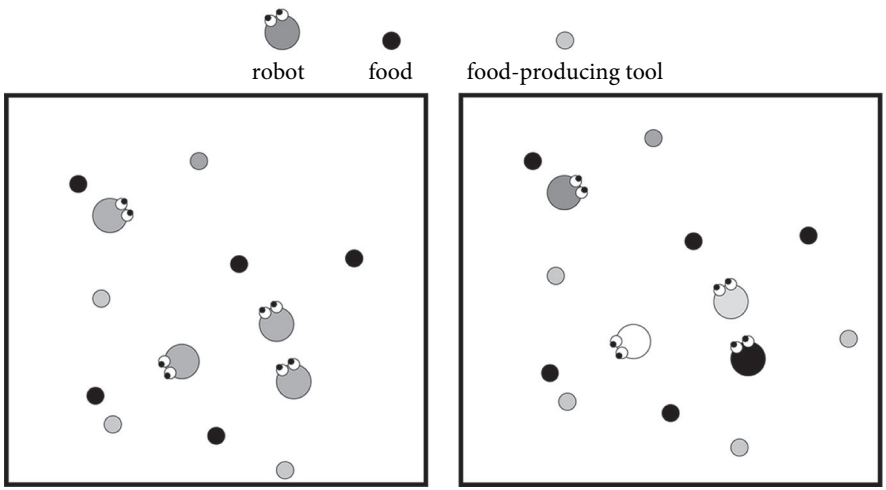


Figure 4. Four robots live together in the same environment which contains both food tokens and tools for producing new food tokens. The four robots are sisters (same colour) (a) or they are genetically extraneous robots (different colours) (b).

What are the consequences of living with genetically extraneous robots and living with genetically similar robots? Do the robots limit themselves to eating the food naturally existing in the environment or do they also spend some of their time producing new food which can be eaten by the other robots? The results are that, if the robots live with their sisters, they evolve the behaviour of producing new food while, if they live with genetically unrelated robots, they don't. The reason is clear. When a robot produces a new food token, the new food token may be eaten not by the robot that has produced it but by another robot. Hence, the food-producing robot does something which *reduces* its survival chances and *increases* the survival chances of another robot. In groups of sisters, the individual survival/reproductive chances of a robot which spends its time on producing new food are reduced but the survival/reproductive chances of the robot's sisters are increased. And here, again, kin selection is at work. Because of her altruistic behaviour, the food-producing robot may not live much and it will not have many daughters but her sisters will live more and will have more daughters. And since the genes of a robot's sisters are similar to the genes of the altruistic robot, the genes of the altruistic robot will be found in the next generation. These genes encode a tendency to produce new food and, therefore, the robots will become food-producing robots. In contrast, if a robot lives with genetically extraneous robots, the behaviour of producing new food increases the reproductive chances of robots which do not necessarily have a tendency to produce new food. Hence, the behaviour of producing new food does not evolve.

The behaviour of producing new food leads to an improvement in the life conditions of the robots because the robots can eat not only the food naturally existing in the environment but also the food that they themselves produce. And, in fact, we find that the total number of food tokens eaten by the robots which live with their sisters and, therefore, evolve the behaviour of producing new food, is greater than the corresponding number for the robots which live with genetically extraneous robots and do not evolve the behaviour of producing new food. The population size of these robots is fixed. But if population size could vary in successive generations, we would find that the robots which produce new food are more numerous than the robots that do not produce new food. This explains the emergence, in human beings, of the strategy of producing new food – agriculture and animal breeding – rather than only eating the existing food – foraging and hunting.

What happens with these robots has two implications. One implication (prediction) is that in human beings agriculture has emerged in groups of genetically related individuals – families – and this may have also happened for other behaviours such as the construction of tools and repairs for the night or the production of new goods by using the existing goods. All these new goods have first appeared in groups of genetically related individuals. The other implication is that producing goods that can be used by other individuals may have been a pressure to live with genetically related individuals – to live in families – because living in families guarantees that one's goods are used by genetically related individuals. Of course, later on human beings have found ways to

develop agriculture and to produce new goods also in groups of genetically *unrelated* individuals by introducing private property and the punishment of individuals which eat the food or use the tools produced by another individual. But in this Chapter we are not interested in the emergence of agriculture or in the emergence of property rights and other social phenomena. We are only interested in explaining why genetic relatedness is an important determinant of how individuals behave towards other individuals and that living with kin-related individuals may have been at the origin of many specifically human phenomena.

5. Males and females

All the robots described so far have only one sex and they reproduce nonsexually. (The mating partners of Chapter 2 are only virtual robots.) But human beings – and many other animals – have two sexes, male and female, and an individual reproduces by mating with an individual of the opposite sex. Hence, human robots should be either male or female and they should reproduce sexually. In this section we describe robots that are males and females and that reproduce by mating together. Sex is related to the existence of families because mothers and fathers tend to live together to take care of their offspring and to mate again.

How are male robots different from female robots? We impose only one single difference between male robots and female robots: mating has different consequences for male and female robots. After mating, males can immediately reproductively mate with another reproductive female while, after successful mating, females have a period in which they are not reproductive due to pregnancy, hormonal changes, lactation, and other factors. What are the consequences of this difference between male and female robots?

The robots live together in an environment that contains food tokens which they must eat to remain alive. Like the preceding robots, the robots have visual sensors which tell them where are the food tokens and they must respond by approaching and eating the food tokens. However, for these robots eating and having a long life are not sufficient to leave their genes to the next generation. Half of the robots are males and half females and to reproduce a male robot must mate with a female robot. A robot's visual neurons inform the robot not only about the location of the nearest food tokens but also about the location of the robots which are near to them and which can be robots of their same sex or of the other sex. Mating occurs when a robot of one sex touches a robot of the opposite sex but the consequences of mating are different for males and females. When a male has mated with a female, the male can immediately mate with another reproductive female and generate another offspring. When a female has mated with a male, the female becomes nonreproductive for a fixed number of cycles and only at the end of this nonreproductive period she returns to be reproductive and she can mate again successfully with the same male or with another male.

Male and female robots have bodies of different colours but the body of a female robot changes when she becomes nonreproductive. Male robots are black. Reproductive female robots are grey when they are reproductive but they become grey with a white part when they become nonreproductive (see Figure 5).

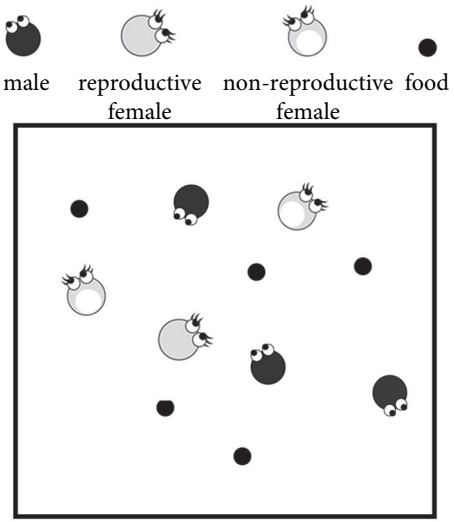


Figure 5. Three male robots, one reproductive female robot, and two non-reproductive female robots that live together in the same environment.

All the robots have visual sensors which allow them to see what is around them and to distinguish among males, reproductive females, nonreproductive females, and food tokens – which have a dark grey colour. But the brains of male and female robots are not identical because the brain of females is informed of their current state: reproductive and nonreproductive. In addition to the visual neurons which encode the location and colour of whatever is near to a female robot, the female's neural network has internal sensory neurons which encode the current reproductive state of the female. This makes the adaptive pattern of females more complex than the adaptive pattern of males because, in their behaviour, females must take into account the state of their body.

All the robots die at the same maximum age but a robot can die earlier if its bodily energy reaches the zero level. However, for these robots to remain alive is not sufficient to leave their genes to the next generation. To have an offspring, a robot must mate with a robot of the opposite sex, where mating occurs when two robots of opposite sex touch each other. Therefore, to leave their genes to the next generation the robots must be able to divide appropriately their time between finding food and finding mates. However, as we have said, females are different from males. After a female has mated with a male, the female becomes nonreproductive for a fixed period of time and, if mating occurs between a male and a nonreproductive female, no offspring are generated.

When all the robots of one generation have died, males are ranked in terms of the number of reproductive mating events they have had during their life and the highest ranking males generate a certain number of *male* offspring which inherit the genes of their father. And the same for the female robots, with the best female robots generating a certain number of *female* offspring. (This clearly is a simplification because, in real animals, offspring have a genotype which is made of parts of the father's genotype and parts of the mother's genotype.) The new male and female robots constitute the next generation of robots which is made of the same number of male and female robots.

Initially, the robots are not very good at eating and they are not very good at mating, which means that they have short lives and they rarely mate during their short life. Then the evolutionary process leads to better robots that eat more, live longer, and mate more often. Our question is: Do males and females behave differently? The only difference between male and female robots is that males can successfully mate at any time while, after successful mating, females have a nonreproductive period. What are the consequences of this single sexual difference for the behaviour of the male and female robots?

As we have said, the robots gradually evolve the capacity to approach and mate successfully with a robot of the opposite sex and, at the end, we find that, on average, a robot has 6 successful mating events – it generates six offspring – a number which, by necessity, is the same for males and females. However, if we look at the best male robot and at the best female robot, we find that the best male robot has a greater number of offspring (around 15) than the best female robot (around 10) (see Figure 6), while the worst male robot has fewer offspring than the worst female robot. This means that female robots tend to have more or less the same number of offspring while some males have a much greater number of offspring than other males. This implies a stronger selective pressure on males than on females. Females are sufficiently sure that they will have their offspring. Males must compete with other males to have offspring.

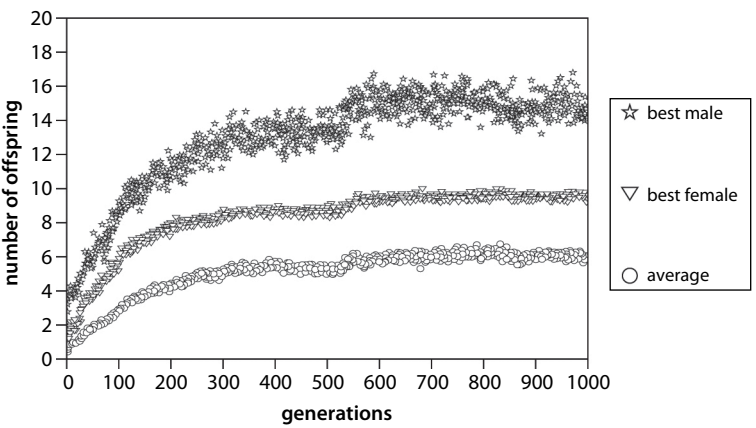


Figure 6. Average number of offspring is necessarily the same for male and female robots but, in all generations, the best male has more offspring than the best female.

But since to mate and have offspring our robots must remain alive and to remain alive they must eat, the increase in the average number of successful mating events implies that our robots do not only evolve an ability to mate successfully but also an ability to eat. These are two different abilities. Two robots can have the same number of successful mating events – same number of offspring – but one robot may be very good at eating and, therefore, it has a long life but it is not particularly good at mating successfully, while another robot may not be so good at eating and, therefore, it may die before it reaches the maximum living age, but it is very good at mating successfully during its shorter life. How males and females compare from the point of view of these two abilities?

If we look at the average male and the average female, we find that males live longer than females and this implies that males eat more than females. And, in fact, if we count the number of food tokens eaten by males and females during their life, we find that males eat more food tokens than females. Why males eat more and live longer than females? The explanation seems to be that, as we have said, females have a more complex adaptive pattern compared to males. A female's brain has to cope with a richer sensory input which includes the internal bodily input which tells the female what is its current reproductive state and the female must respond to this additional input by behaving differently when she is reproductive and when she is not reproductive. Males do not have this problem. Their brain receives sensory input only from the external environment and their behaviour must be always the same: look for reproductive females and, while you are looking for reproductive females, eat all the food tokens you can find.

If we look at the robots on the computer screen, we find that not only females behave differently from males but they behave differently when they are reproductive and when they are noneproductive. Males tend to move more and to explore more parts of the environment but this is true only with respect to reproductive females, not with respect to nonreproductive females (see Figure 7). Males tend to be always very active and they are always looking for something, be it food or a reproductive female. Reproductive females do not move much, and only when they become nonreproductive, they are as mobile and explorative as males. These differences in behaviour are associated with different mating strategies. The mating strategy of males is to move in the environment looking for reproductive females and, as soon as they find them, to approach and mate with them. The mating strategy of reproductive females is to remain where they are and to wait for males to reach and mate with them. These different behavioural strategies have consequences for eating and surviving. Although males are more interested in reproductive females than in food, they eat more than females because, overall, they move more and, therefore, they find more food than females. It is only when females become nonreproductive that they become more active and look for food.

Another difference between males and females has to do with how length of life and mating success – number of offspring – are correlated in males and females. Are

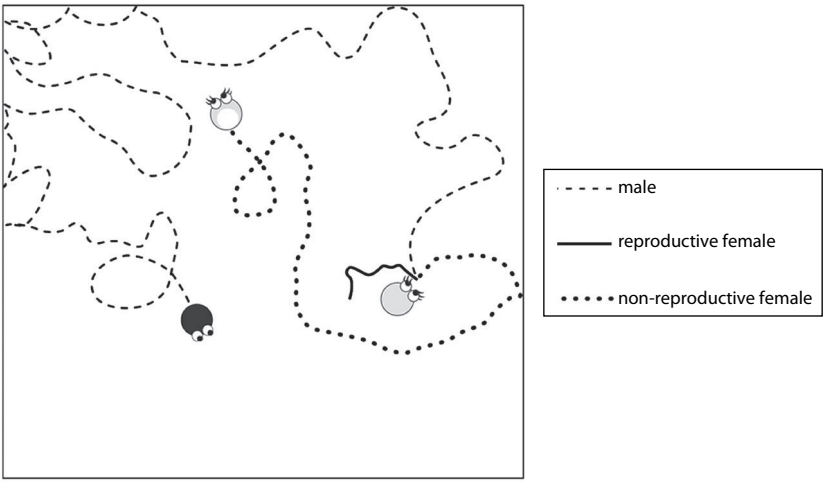


Figure 7. Parts of the environment visited by a male, a reproductive female, and a non-reproductive female during the same number of cycles. Males explore more of the environment than reproductive females. Nonreproductive females are in between.

the robots that live longer those which have more offspring? This is true for females but not for males. All females live more or less the same time and they have more or less the same number of offspring while some males have a short life but more offspring than other males which live longer (see Figure 8). Males and female have two different adaptive strategies. Because of their nonreproductive period after mating, females can only have a maximum number of offspring during their entire life. Males do not have this limit because, theoretically, they can have a successful mating event in each cycle of their life and their reproductive success is only limited by the number of available reproductive females. Males' reproductive success is measured by their ability to find reproductive females to mate with rather than by their ability to find food. In contrast, reproductive females do not have to look for males because males look for them, and their only problem is to eat and have a long life so that they can express their maximum reproductive potential. However, paradoxically, the greater mobility of males causes males to eat more food and, therefore, to live longer than females.

To confirm this interpretation of the behaviour of males and females, we measure how mobile are the different robots and how much of the environment they explore. We divide the environment into a certain number of cells and we count how many different cells are visited by males, reproductive females, and nonreproductive females in a fixed period of time identical for all robots. The results are that males visit more cells of the environment and they move faster than both reproductive and nonreproductive females (see Figure 9).

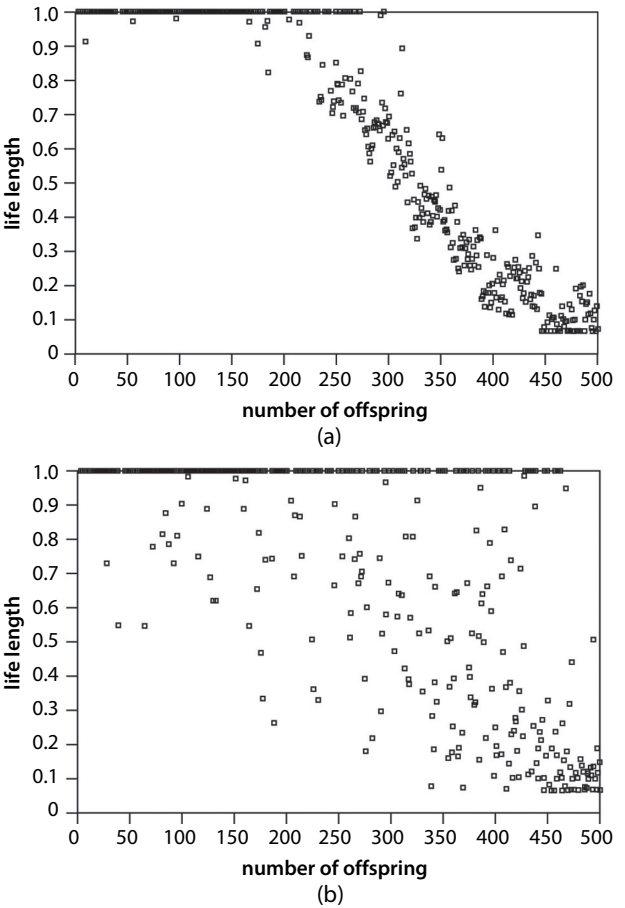


Figure 8. Females that live longer have more offspring (a) while this is not true for males (b). A male can have many offspring and a short life or it can have few offspring and a long life.

To better understand how males, reproductive females, and nonreproductive females are different in their behaviour, we examine how the robots behave in an experimental laboratory in which they are given a choice between two options. We consider all possible choices. In one experimental trial a male has to choose between a reproductive female and a food token, in another trial the same male has to choose between a reproductive female and a nonreproductive female, in a third trial between a food token and another male, and so on. And the same for reproductive and nonreproductive females. They choose between a male and a food token, between a male and a reproductive female, etc. While the robot which is the experimental subject is free to move, the robots between which the experimental subject has to choose do not move. Each

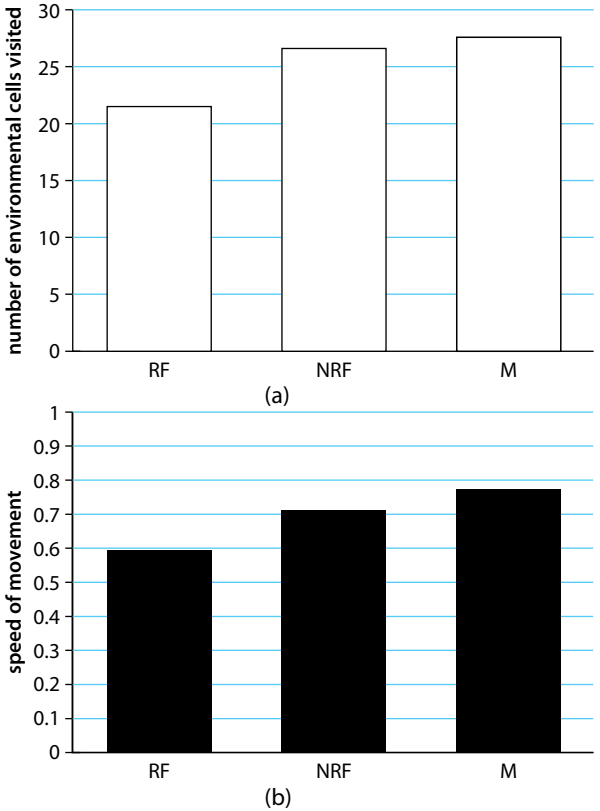


Figure 9. Number of environmental cells visited (a) and speed of movement (b) for reproductive females (RF), nonreproductive females (NRF), and males (M).

trial lasts a fixed number of cycles and, for each trial, we take three measures: (1) what is the choice of the robot; (2) how much time the robot takes to reach one of the two items – a robot or a food token; (3) the number of trials in which the robot does not appear to be much interested in the two items between which it has to choose because, when the trial ends, the robot has reached neither of the two items.

The results are the following. A male which has to choose between a reproductive female and a food token clearly prefers the reproductive female, and an even stronger preference for reproductive females is manifested by males when they have to choose between a reproductive female and a nonreproductive female or between a reproductive female and a male. The situation is reversed when a male has to choose between food and a nonreproductive female or between food and another male: the male almost always chooses food. In all cases, males are very fast in their movements and they never appear undecided. A trial rarely terminates without them having reached one of the two items (see Figure 10).

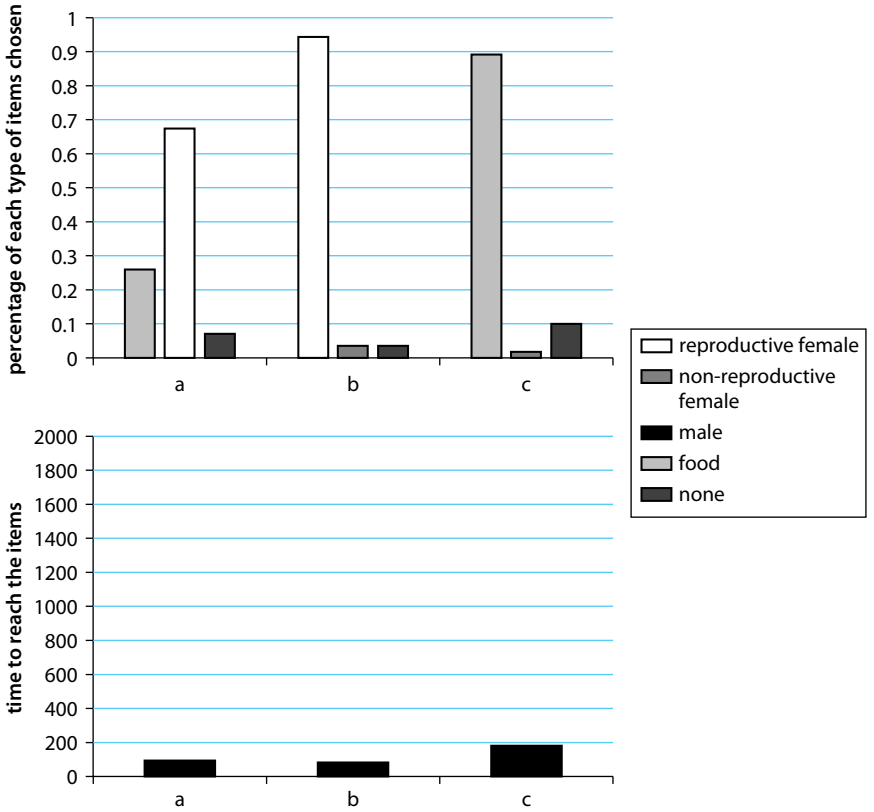


Figure 10. Males' preferences in the laboratory. Above: Males choose between food and reproductive female (a), reproductive and nonreproductive female (b), food and nonreproductive female (c). Below: Time taken by males to reach the chosen item.

In contrast to males, reproductive females prefer food to males. And even when they must choose between a male and a reproductive female, reproductive females prefer the male only a little more than the reproductive female. In both cases, and especially in the second one, reproductive females are very hesitant and slow in their behaviour (see Figure 11). Another interesting finding is that when reproductive females must decide between a reproductive and a nonreproductive female, they prefer the reproductive female to the nonreproductive female, although in this case too they manifest a great indecision and a tendency not to move much. This seems to imply a specific mating strategy on the part of reproductive females: they tend to remain near to other reproductive females in order to increase the probability to attract males.

In contrast, when they become nonreproductive, females change radically their behaviour. Nonreproductive females prefer food to everything else and they are fast

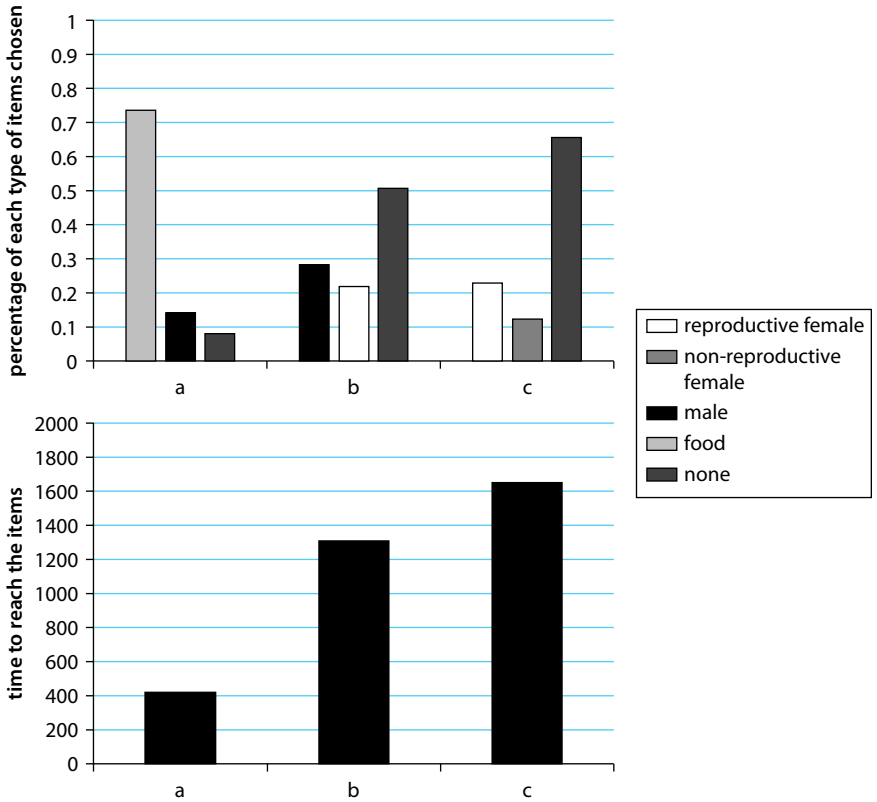


Figure 11. Reproductive females' preferences. Above: Reproductive females choose between (a) food and a male, (b) a male and a reproductive female, (c) a reproductive and a nonreproductive female. Below: Time taken to reach the chosen item.

and very decided almost like males (see Figure 12). Only when food is not one of the two choices, nonreproductive females are not interested in choosing.

The results of these experiments can be summarized in the following way. Males tend to always behave in the same manner. They are much attracted by reproductive females and they prefer reproductive females to everything else, including food. However, when reproductive females are out of sight, males are very attracted by food and, since they are always very active and explorative, they eat much food. On the contrary, females behave differently when they are reproductive and when they are nonreproductive. When they are reproductive, females prefer food to males. Furthermore, reproductive females only slightly prefer males to another reproductive female and they have a certain tendency to be attracted by other reproductive (but not by nonreproductive) females, thereby increasing the probability to attract males because two

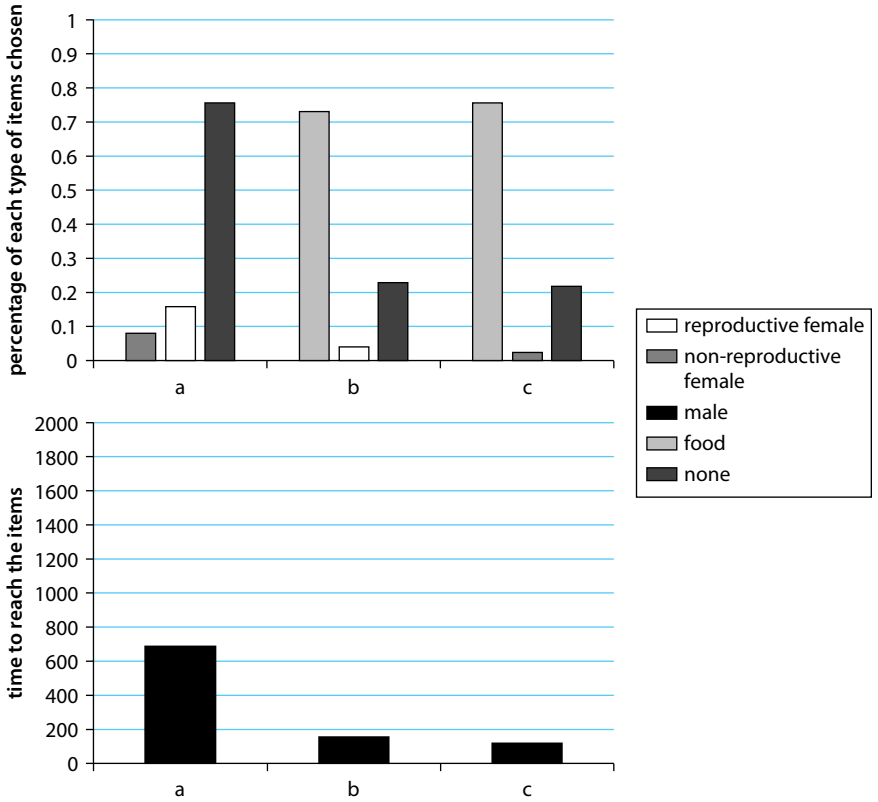


Figure 12. Nonreproductive females' preferences. Above: Nonreproductive females choose between male and reproductive female (a), food and reproductive female (b), food and male. Below: Time taken to reach the chosen item.

reproductive females together may attract a male more than a single reproductive female. But in all circumstances their behaviour tends to be slow and hesitant. However, when females become nonreproductive as a result of mating, they are very active much like males. But, contrary to males, they prefer food to everything else and are not attracted by anything else.

These robots live in an environment in which all the food tokens are identical and they all contain the same quantity of energy. Are there other differences in the behaviour of males and females if the robots live in an environment which different types of food? For example, how males and females respond to different types of food which contain different quantities of energy? How they respond to risky food?

The new robots live in an environment which contains two different types of food tokens and the two types of tokens have different colours. Half of the tokens are black and half are white and the black tokens contain twice the energy of the white tokens. In

one environment this is the only difference between the two types of food. In another environment there is still another difference, and this difference is linked to risk. The black tokens contain twice the energy of the white tokens but they have a 30% probability of being toxic and, if a robot eats toxic food, its energy is decreased. What are the preferences of males and females in this environment?

After the male and female robots have evolved in these more complex environments, we test them in an experimental laboratory like the preceding robots. A robot is placed in a fixed position in front of two food tokens, one black and the other one white, and we determine which of the two tokens the robot reaches in a fixed period of time. Male robots are tested only once while female robots are tested twice: when they are reproductive and when they are nonreproductive. The results are that in both environments males do not have any preferences for one or the other type of food. They approach and eat any type of food independently of the quantity of energy contained in the food and of the risk associated with the more energetic food. In contrast, females have clear food preferences but their preferences are different in the two environments. If they have evolved in the environment in which the more energetic food does not present any risks, females tend to prefer the more energetic food to the less energetic one. However, if the more energetic food has s probability of 30% to be toxic, they prefer the less energetic but riskless food to the more energetic but risky one.

These are the results of an experiment in which everything is controlled by us. But if we count the number of food tokens of the two types eaten by the robots in their *natural* environment, we obtain different results. In the environment in which the food tokens are different only in their energetic value, not only females but also males eat more tokens of the more energetic food than tokens of the less energetic food. In the environment in which the more energetic food is risky because it has some probability of being toxic, the situation is reversed and both males and females eat more tokens of the less energetic, but riskless, food than of the more energetic, but potentially toxic, food.

These differences between the behaviour of the robots in their natural environment and their behaviour in the experimental laboratory may surprise us but they can be explained. In the controlled conditions of the laboratory we capture the intrinsic preferences of the robots for the two types of food in the sense that what a robot does in the laboratory depends entirely on the food preferences which are encoded in its brain. In contrast, in the natural environment a robot's behaviour is a function not only of the food preferences encoded in the robot's brain but also of the particular environment in which the robot finds itself at any given time. The environment of our robots is a social environment, and a social environment is what it is because of what the other robots do. In the natural environment males eat more of the more energetic food because males are always looking for reproductive females and reproductive females have a preference for the more energetic food. Therefore, males which are looking for reproductive females are more likely to find them near to the more energetic food. This explanation is supported by two other tests. We place a male robot in an semi-natural

environment which is identical to the natural environment except that the robot is completely alone: there are no females and no other males but there is only food. In another test the environment contains other male robots but no female robots. In these semi-natural environments, the male robots show *no* preferences for the two types of food, exactly like in the laboratory. This confirms that males do not actually have food preferences inscribed in their brain but it is the presence of females that causes them to eat the food which is preferred by females. It is like human males going to a tea house instead of a pub in order to meet potential female partners: they end up sipping herbal tea, even if they would personally prefer to drink beer. (But, today, both males and females might prefer to drink beer.)

The differences in the behaviour of the robots in their natural environment and in the experimental laboratory allow us to make an important methodological point we have already discussed in the first chapter of this book. Laboratory experiments are considered as the best method for collecting empirical data not only in the natural sciences – in which they *are* the best method – but also in psychology and, increasingly, in the social sciences. But behaviour depends on the environment, and the environment of laboratory experiments is very different from the natural environment. The task of behavioural scientists is to know and to explain the behaviour of human beings in their natural environment, and laboratory experiments may not tell them all they need to know about human behaviour in everyday life. Robots are useful because we can study their behaviour both in their natural environment and in the experimental laboratory. We can compare their behaviour in the two environments and, if there are differences, we can explain these differences. And they are especially useful because, first of all, they can make students of human behaviour aware of the limits of laboratory experiments and because they can suggest hypotheses on why human beings behave in different ways in the experimental laboratory and in their everyday life.

We find other differences in the behaviour of males and females if we change another thing in our robotic scenario. For the male and female robots described so far, mating and having offspring are sufficient to leave their genes to the next generation. But, as we have already said, in many animals offspring are born without the capacity to feed themselves and, therefore, if they must survive and reach reproductive age, they must be fed by someone else. Kin selection theory predicts that parents will feed their infant offspring because by feeding their infant offspring they increase the probability that copies of their genes will be found in future generations – and in Section 2 we have described unisexual robots that renounce to some of their food to feed their infant offspring. What happens if the robots are male and female and reproduction is sexual? Will both mothers and fathers feed their infant offspring?

We add to the robots' genotype the same "offspring care" gene of the robots described in Section 2. The "offspring care" gene can vary between 0 and 1, where 0 means that the individual carrying the gene eats all the food it is able to find without giving any of the food to its infant offspring while 1 means that the individual gives all of its

food to its offspring. We construct two different populations of robots. In one population only females have the "offspring care" gene. Males don't. Hence, only mothers give their food to their offspring. In the other population both males and females have the "offspring care" gene and therefore both fathers and mothers can feed their offspring. In the population in which only mothers feed their offspring, female robots evolve an average value of 0.60 for their "offspring care" gene, which means that mothers give more than half of the food they are able to procure to their offspring – and this confirms what we have already found with the unisexual robots described in Section 2. In the population in which *both* mothers and fathers can feed their offspring, the evolved value of the "offspring care" gene is more or less the same for both mothers and fathers, although it is somewhat lower than in the population in which only mothers had the "offspring care" gene because now the offspring can be fed by two parents, not only one. If both mothers and fathers feed their offspring, the burden is less for both.

But even if both fathers and mothers have a genetically inherited tendency to feed their dependent offspring, there may be one factor that influences fathers' tendency to feed their offspring, and this factor is paternal certainty. While mothers are certain that their food actually goes to their offspring, fathers are not so certain. Males can be feeding not their offspring but the offspring of another male because the mother may have mated with another male. To reproduce this phenomenon we change slightly modify our robots. While mothers are 100% sure that they are feeding their infant offspring, when a male robot decides to give its food to its infant offspring, its offspring will actually receive the food only 50% of the times and the other 50% of the times the food goes to another randomly selected infant robot. The results confirm the importance of paternal certainty. In the population in which fathers are not so sure that they are actually feeding their offspring or the offspring of another male, males have a lower value for their "offspring care" gene (0.39) compared to females (0.52). Since males are not certain that they are feeding *their* offspring, and not the offspring of some other male, they are less disposed to give their food to the dependent offspring.

But a higher value of the "offspring care" gene for females compared to males – which means a greater investment in one's offspring on the part of mothers than on the part of fathers – may have a more general explanation. The theory of natural selection predicts that both females and males will behave in ways that increase the probability that their genes will continue to exist in future generations. However, given their different roles in reproduction, females and males may adopt two different strategies to achieve this goal. Females strongly want that their necessarily restricted number of offspring reach reproductive age and reproduce in their turn. Therefore, females tend to invest a lot in their offspring. In contrast, males tend to adopt a strategy based on "great numbers". They spend their time trying to have the largest possible number of mates and, therefore, the largest number of offspring but they are less prone to invest in their existing offspring. In order to guarantee a future for their genes, males count on the probability that, given a large number of individuals which have their same

genes (their offspring), for purely chance reasons at least some of these individuals will survive and reproduce.

Another interesting result is that, if a robot must find food not only for itself but also for its offspring, food becomes more important for the robots. We have tested the robots which have the "offspring care" gene in the same laboratory setting used for the robots which do not have the "offspring care" gene. A robot is offered a choice between two items and we determine which item it chooses, how much time it takes to choose, or if it does not choose at all. The results are that if the robots have offspring which depend on their parent(s) for food, the importance of food for the robots increases. In all the choices in which a robot must choose between two different items one of which is food, the probability that the robot will choose food is greater for the robots that must feed their offspring than for the robots whose offspring do not depend on them for food.

Let us summarize what we have found with our male and female robots. The only difference we have imposed on our robots is that after mating females have a nonreproductive period during which mating does not result in conception, and the female's body tells the female's brain what is her current reproductive state. This is not true for males which can successfully mate at any time – if they are able to find a reproductive female. We have found a number of differences in the behaviour of males and females which all derive from the fact that males and females are under different selective pressures. Reproductive females are a "scarce resource" for males because at any given time many females are nonreproductive. Therefore, males must compete with other males for this "scarce resource", and finding reproductive females, not eating and having a long life, is the main challenge for them. On the contrary, males are not a "scarce resource" for females because at any given time all males are potential mates. While females can generate a maximum of number offspring during their entire life, a male can (theoretically) generate one offspring in each cycle of its life. Therefore, for our female robots the challenge is not to find mates but to live as long as possible.

These different selective pressures result in a number of differences in the behaviour of males and females. Males tend to always explore actively the environment. They always look for reproductive females and they prefer reproductive females to anything else. How can males adopt this behaviour if food is necessary for their survival? Contrary to what one would predict on the basis of the preference of males for reproductive females rather than food, males eat more and, on average, live longer than females. The answer to our question is that males are always exploring actively the environment searching for the "scarce resource" represented by reproductive females and, while they are exploring the environment, they are able to find a lot of food. However, as soon as they perceive a reproductive female, males tend to ignore food and to approach and mate with the female.

Reproductive females behave very differently. They do not search for males although, being reproductive, they might successfully mate with them. And, even when they search for food, they are not very active and move rather slowly. This difference

in the behaviour of males and females is due to the fact that males and females have a different (symmetrical) reproductive strategy: males actively look for reproductive females while reproductive females wait for males to come and mate with them. Not only reproductive females do not search for males but they do not go actively after food because they prefer to remain where they are so that males may more easily find them and mate with them. In our laboratory experiments we have even found a tendency of reproductive females to stay near to other reproductive females, which appears to be another strategy for attracting males.

Our robots must eat to remain alive and, although remaining alive is not sufficient for transmitting their genes to the next generation, it is a necessary pre-condition. Therefore, one may ask how females can remain alive although they do not look actively for food. The answer is that a reproductive female will soon become nonreproductive because some male will find her and mate with her. And a female behaves very differently when she becomes nonreproductive. The same female which, when she is reproductive, is not very active and does not look very much for food but only waits for males to mate the her, when she has mated and has become nonreproductive, becomes very active and looks actively for food ignoring everything else.

Males can be said to have a simpler life in the sense that they are expected to do always the same thing. They must actively look for reproductive females and eat all food that you can find while looking for reproductive females. In contrast, the life of a female is more complex because it includes two different states or periods. When a female is reproductive, she is not very active and she waits for males to mate with her. When she becomes nonreproductive, her behaviour changes. She actively looks for food and is not interested in anything else. Although they live in the same physical environment, males and females live in different behavioural environments. While the physical environment in which an animal lives is independent of the animal, its behavioural environment depends on the animal's characteristics, including its reproductive strategy. Males and females have different roles in reproduction and, therefore, their ecological niches are different. This causes a number of differences in their behaviour.

The differences in the adaptive pattern of males and females also emerge in more complex environments in which there are different types of food. Males do not have food preferences. Their feeding strategy is to eat whatever food they happen to find while they are actively exploring the environment in search of reproductive females, independently of the energetic value of the food and of the risk associated with the more energetic food. Females are more discriminative. Since their main adaptive challenge is to live as long as possible, they have a preference for the more energetic food over the less energetic food. But when the more energetic food is risky because it can be toxic, they show an opposite preference for the less energetic food over the more energetic one.

Other differences in the adaptive pattern of males and females emerge if to leave one's genes to future generations it is not sufficient to have offspring but it is also

necessary to feed one's offspring to keep them alive. While both fathers and mothers have the same tendency to give their food to their dependent offspring when parental certainty is as great for males as for females, when males are less certain than females that they are really feeding their offspring, males are less disposed to feed their (supposed) offspring than mothers are disposed to feed their (sure) offspring.

Our robots are not aimed at reproducing any specific animal. They are very simplified, "general", animals. For example, in our robots the reproductive state of females is externally visible because they have different colours when they are reproductive and when they are nonreproductive. This is true for some nonhuman primates but in human beings the reproductive state of women is hidden, and this, clearly, has a series of important consequences. But what is interesting is that even the single difference that we have imposed on our robots (and that generally exists in nature), which causes females, but not males, to have a nonreproductive period after mating, results in a number of differences in the behaviour of our male and female robots which tend to be found in many real animals and in human beings. All females have more or less the same number of offspring while a few males have many more offspring than other males. Males actively explore the environment and, consequently, if they were able to learn, they would learn more about the environment – the natural environment, not necessarily the social environment. Males exhibit risky behaviours. Males give a greater importance to mating compared to eating and males invest less in their offspring compared to females.

In real animals the greater investment in one's offspring of females compared to males is attributed to a variety of factors. One is the different size of the female egg cell compared to the male sperm cell: egg cells are much larger and they need more energy to survive than sperm cells. Hence, for females the egg cell is more important than the sperm cell is for males, and females are more careful about mating (reproductively) compared to males. Another factor is parenting effort, which is greater in females than in males in that females spend more of their energy in assisting their offspring after conception, both in uterus and after birth, compared to males. Our robots show that even in the absence of these factors females invest more than males in their offspring due to the smaller number of offspring that they can possibly have and to their greater parental certainty compared to males.

There is one aspect, however, in which our robots differ from human beings: our male robots live longer than female robots while in human beings females tend to live longer than males. To reproduce this phenomenon we should add other factors which we have ignored in constructing our robots. For example, movement has no costs for our robots. If, more realistically, we add movement cost, males might have a shorter life than females because they move more than females. Or, males may physically, or in other ways, compete with other males for access to females and this may result in a shorter life length of males compared to females because of the resources that males must consume to compete with other males. And if mothers are more involved than

fathers in caring for their offspring to keep them alive, there might be a greater pressure on females than on males to live longer in order to take care of both their children and their grandchildren (see Section 4 on grandmothers).

Other phenomena could be reproduced by adding other features to our robots. Our robots are males or females and the two categories of robots can be distinguished on the basis of their colour. But there are no inter-individual differences among the robots of the same sex that can be perceived by the other robots. What if the robots that belong to the same sexual category were not all identical? Will a robot behave differently towards two individuals that belong to the same sexual category but have different characteristics? Male robots could be larger or smaller, and the size of a male robot might be perceived by another robot and have consequences for the other robot's behaviour. For example, if a male robot must physically compete with another male robot for a reproductive female, the size of the two robots might be a factor in deciding how the two robots behave: the smaller male might simply go away.

Inter-individual differences among robots that belong to the same sexual category are linked to a very important phenomenon: sexual selection. Since different individuals have different genes, different individuals will generate offspring with different chances of survival and reproduction. If the quality of the genes of an individual is reflected in the body and behaviour of the individual, both females and males should develop a capacity to identify the physical and behavioural characteristics associated with good and bad genes in potential mates and select mates based on these characteristics. This is sexual selection and the characteristics on which sexual selection is based constitute the sexual attractiveness (beauty) of an individual. Beauty is different in males and females. Males tend to be chosen as mates by females if they have physical and behavioural characteristics which are correlated with their ability to find food (in modern human beings, money) or to defend their mate and offspring from all sorts of dangers. Females tend to be chosen as mates by males if they have characteristics correlated with their ability to give birth more easily, to give birth to healthier offspring, to feed their dependent offspring with their milk, and to interact in a caring way with their offspring. In many animals, sexual selection tends to be more important for females than males in the sense that female choose their mating partners while males are less selective. This should also be true for our robots since females are a "scarce resource" for males but males are *not* a "scarce resource" for females. And if there are physical or behavioural traits of males which tend to increase the probability that a male will be chosen as sexual partner by females, these traits of males should also emerge in the population. This may involve what is called costly signalling: the possession of a trait which is not practically useful and it consumes resources but increases the chances to be chosen as mate. (An example is the beautiful tail of male peacocks). In human beings females spend more of their resources than males to attract males than vice versa and what makes an individual attractive for an individual of the other sex tends to be different from males and females.

Another phenomenon linked to sexual attractiveness that needs to be reproduced with robots is that men like always the same characteristics in women while women like two different types of physical and behavioural characteristics in men. Some characteristics of men are more masculine. They reveal good genes, the ability to protect females from other males that want to mate with them, intelligence and the ability to procure resources, but not necessarily a tendency to invest in their offspring. The other type of male characteristics which attract females are more feminine, and they reveal a character which is more sensitive to the needs of others and more prone to help others. This difference between women and men may explain while all women (with exceptions) tend to be "feminine" while some men are more "masculine" and other men more "feminine". Another phenomenon that we should reproduce with robots is that women seem to like the first type of male characteristics (masculine) when they are reproductive and they are looking for a male which has good genes, can protect them, and is able to procure resources, while they like the second type of male characteristics (feminine) when they are nonreproductive and they need a male who takes care of them and their offspring.

Mating and caring for one's offspring are motivations which have an intrinsic great strength and they often win the competition with other motivations in controlling an animal's behaviour. In Chapter 2 we have shown that adding a special "emotional circuit" to the neural network of a robot tends to make the robot's behaviour more effective and to increase the robot's survival/reproductive chances because the states of this circuit (emotional states) help the robot to make more appropriate motivational decisions. Emotional states are based on the interaction of the brain with the rest of the body and they cause changes not only in the internal organs and systems of the body but also in the posture, movements, and other external characteristics of the body which can be perceived by other robots (expression of emotions). This allows a robot to let other robots know its emotional states, and this knowledge might influence the motivational decisions of the other robots. Mating and caring for one's offspring are an area in which the external expression of emotions plays a very important role, inviting or rejecting mating attempts, establishing attachment between parent and offspring, and in other ways.

The differences in the adaptive patterns of males and females are also reflected in the motivations and emotional states of males and females. For example, males and females tend to have two different forms of jealousy. Jealousy in females tends to be "emotional" whereas jealousy in males tends to be "sexual". Females resent more the behaviour of their (stable) partner if this behaviour signals an emotional involvement of their partner with another female. Males are more hit by their partner's behaviour if this behaviour signals sexual intercourse with another male. The emotional involvement of a female's male partner with another female may imply a reduction in the partner's tendency to invest in his family (wife and offspring) and. therefore, females are more sensitive to the emotional "treason" of their partners, while being relatively

less sensitive to their sexual "treason" (if it is only a one time affair). On the other hand, males are more sensitive to the sexual involvement of their female partner with another male because this may reduce their certainty that the female's offspring are their offspring, while being less sensitive to their emotional "treason" because females cannot but invest in their offspring. These differences between males and females are interesting and we should reproduce them in our robots.

6. Homes

At the beginning of this Chapter we have defined a social family as a group of genetically related individuals which live close to one another and interact together. A group of individuals can live near to one another and interact together even if they move in space because they can move together in space – and this is what happens in many animals. But moving together poses many problems to human beings. It reduces an individual's freedom of movement because the individual must move together with the other individuals and, therefore, he or she can explore much less the environment. Very young individuals may not be able to move and therefore they must be carried around by adults. If an individual loses contact with the group, the individual may not be able to re-join the group. If the individuals store their food in external stores, the external stores cannot be too large and contain too much food because they must be transported by the individuals. (For robots that have external stores, see Chapter 9.)

The solution to all these problems is homes. A home is a place in the environment to which all the members of a family return periodically. Homes make it possible for the members of the family to move freely in the environment and then rejoin the other members of the family by returning home. Young members of the family which are unable to move can stay at home. Homes can be places in which to store all sorts of resources which are needed by all the members of the family. And they can provide protection from dangers and adverse conditions such as cold and rain. Some nonhuman animals have "homes" – nests, repairs, places in which they store food – and some of these "homes" are family homes. Human beings have had family homes for most of their long history although, today, an increasing number of individuals live in individual homes.

The male and female robots described in the preceding section come together only for mating and, therefore, they do not have families because families are groups of genetically related individuals which live, more or less permanently, near to one another. But we can change our robots and reproduce both robotic homes and robotic families. Our female robots can move as easily when they are reproductive and when they are nonreproductive, and we have even found that, when they are noreproductive, they move and actively search for food much more than when they are reproductive. But this is not what happens in real animals and, especially, in human beings. When they

are pregnant and when they must breast-feed and take care of their infant children, female human beings cannot move so easily and they can only look for food near to where they are. We might reproduce this phenomenon by constructing robots that possess the following behaviour – but we have not actually constructed these robots. When a male robot and a female robot mate together, they construct a home where the female robot and her offspring that still cannot move and must be fed by their mother live and where their father returns periodically. Homes also have another function: they increase paternal certainty. To be sure that they are feeding their offspring and not the offspring of other males, male robots tend to remain in proximity to their "wife" and, therefore, to live in the home together with their "wife" and their infant offspring. (Homes can also be places in which to store one's food. We describe robots that have family stores in Chapter 9.)

Homes have another implication which has to do with spatial abilities. While female robots search for food near to their home – woman the forager – male robots look for food in larger spaces far from their homes – man the hunter. Female robots do not need spatial landmarks because they tend to remain near to their home and their only landmark is the home. But males need spatial landmarks to return to their home from distant places – and this may cause males to have more developed spatial abilities compared to females.

To demonstrate the importance of spatial landmarks, we evolve three different populations of robots which must return to their home after they have collected food in the environment and, for these robots, fitness is the time taken to return home. The environment of the first population does not contain landmarks and, therefore, a robot can only explore more or less randomly the environment until it finds its home (see Figure 13a). The environment of the second population contains a non-directional landmark – a white object. The landmark is sufficiently near to the robot's home so that, when the robot sees the landmark, it reaches the landmark and, by moving around the landmark, it finds its home (see Figure 13b). The landmark of the third populations is a directional landmark – an object which is half white and half black. When a robot sees this landmark, the robot does not need to reach the landmark because its home is always on the black side of the landmark and, therefore, the robot can directly go towards its home (see Figure 13c).

The results are that the robots that live in the environment without the landmark take much time to reach their home because they cannot but move around in the environment until they happen to find their home and, therefore, they have a low level of fitness. The robots that live in the environment with a non-directional landmark have a higher level of fitness because they use the landmark to reach their home. And the robots that live in the environment with a directional landmark have an even higher level of fitness because when they see the landmark they follow the direction indicated by the landmark and they reach their home in less time (see Figure 14).

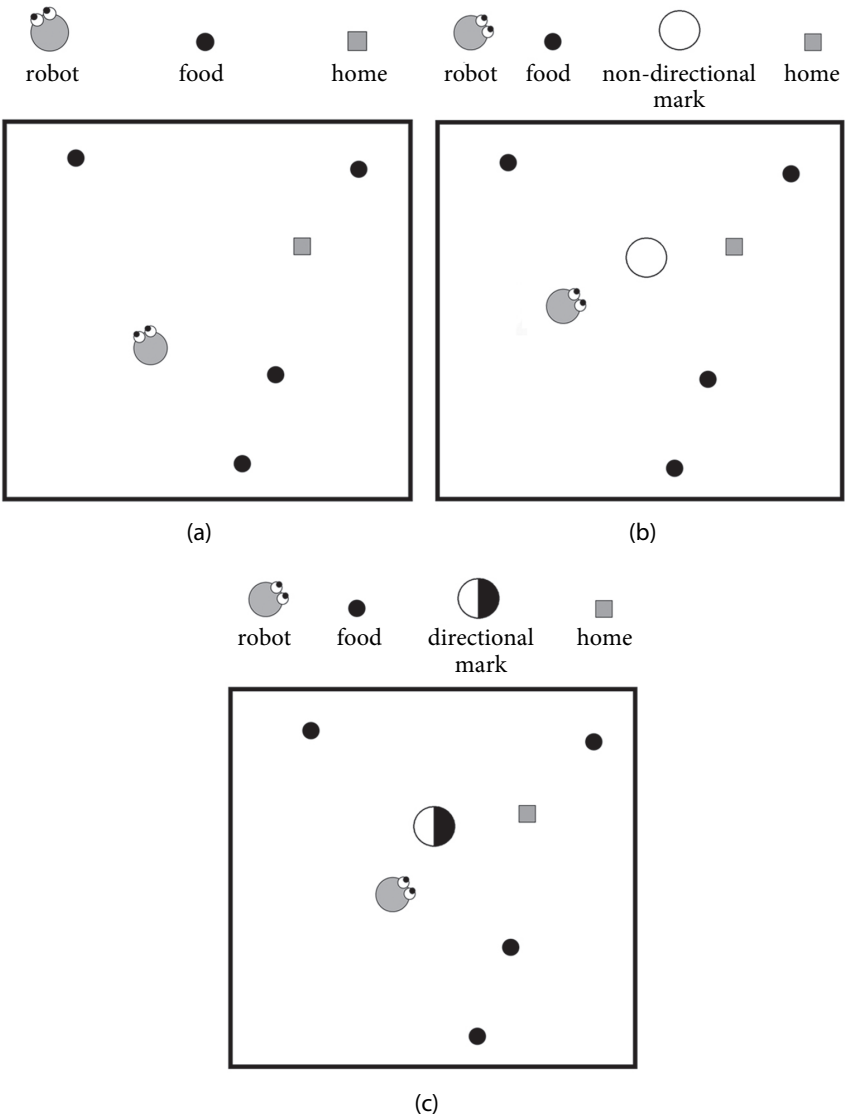


Figure 13. The robots must return to their home after collecting food in the environment and their environment can contain no landmark (a), a non-directional landmark (b), or a directional landmark (c).

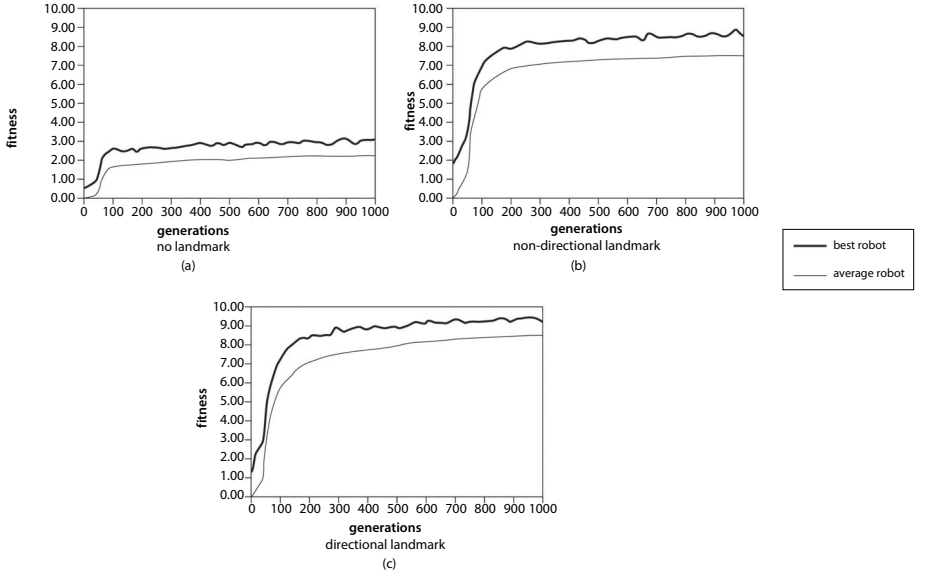


Figure 14. Fitness of robots living in an environment without landmarks (a), with a non-directional landmark (b), and with a directional landmark (c).

To better see how the robots respond to the directional and to the non-directional landmarks, we test the robots in an experimental laboratory. We place the robots in front of a landmark, either directional or non-directional, we let the robots move for a certain time and we measure the average distance of the robots from the landmark during this time. The robots that live in the environment with a directional landmark keep a greater distance from the landmark than the robots that live in the environment with a non-directional landmark. This demonstrates that the directional landmark tells the robots where to go to find their home without reaching the landmark.

These robots are unisexual robots. But if male robots explore more widely the environment in search for food compared to female robots, this may represent a greater evolutionary pressure to develop a capacity to do landmark navigation and to develop "spatial maps" in the male brain compared to the female brain.

7. Conclusions

This Chapter has been dedicated to how robots behave towards other robots which have their same genes, to male and female robots that must mate together to have offspring, and to robots that live in families. Our results can be explained by the theory of kin selection according to which an individual can guarantee the future presence of its genes in the genetic pool of the population not only by having offspring but also by helping genetically related individuals to remain alive and have their own offspring. We will return to families in Chapter 9 dedicated to robots that store their food in external stores instead of immediately eating all the food they find in the environment, and the reason is that the external stores of human beings are not individual stores but family stores to which all the members of a family have access. In that Chapter we will describe robots that replicate other phenomena linked to families such as the inheritance of goods, reciprocal help among the members of a family for the entire length of their life, and the need for genetically related families to live in homes which are near to one another – which is the beginning of villages and cities.

Kin selection theory also explains why one male and one female human beings stay together and help each other although they do not have the same genes: they stay together and help each other because this allows them to have other offspring and to keep their already existing offspring alive. This poses many interesting questions to which future robots will provide an answer. Is love towards one's offspring different from love towards one's mate? Why mates can cease to love each other while this is almost never true for parents and offspring? What is "love" for human beings? Is love behaviour or emotions and mental life? Other questions are less clearly linked to kin selection theory. Our robots are inter-individually different, so we can ask: How individuals choose their mate? Can we predict which individual of one sex will be attracted by which individual of the other sex and when the attraction will be reciprocal? Can

we predict how long a specific male and a specific female will remain together? How can kin selection theory explain homosexual couples?

In this Chapter we have looked at robotic families from a purely biological perspective but human families also depend on social, economic, and cultural factors which influence how human families function and how they have changed in the course of human history and continue to change (very rapidly) today. Reproducing the effects of these social, economic, and cultural factors on robotic families is another important task for the future.

CHAPTER 8

Robots that learn from other robots and develop cultures and technologies

ME is convinced that one of the most important reasons why human beings live together is that, by living together, they can learn from one another. Unlike nonhuman animals, which are born with most of the behaviours encoded in their genes, human beings learn almost all their behaviours – and also what they believe and what they value – in the course of their life and they learn all these things not by interacting with nature but by interacting with other human beings. The result is that a group of interacting human beings tend to have the same behaviours, the same beliefs, and the same values – what is called their culture – and that the culture of one group is different from the culture of another group if the two groups do not interact together and do not learn from one another. But ME also notes that human cultures change, and they change because only some of the behaviours, beliefs, and values of one generation are transmitted to the next generation and novelties are constantly invented or imported from other cultures.

ME is also struck by the great number of artefacts that human beings construct to satisfy their needs – and this is another thing which distinguishes human beings from nonhuman animals. The artefacts of a group are part of the culture of the group because how to construct an artefact and how to use the artefact are both behaviours learned from others, and the artefacts change because they are transmitted from one generation to the next with variations and inventions. Recently, human beings have invented artefacts that allow them to interact together not only if they live near to one another but to interact with any other human being, wherever he or she lives on the Earth. ME thinks that this is leading to the emergence of a single culture, and this single culture is the Western culture because it is the Western culture which appears to the eyes of all human beings as the culture which allows them to have the best life.

1. Learning from others

All animals learn but what makes human beings different from nonhuman animals is that, while nonhuman animals learn by interacting with the inanimate environment, human beings mostly learn from other human beings. They learn by imitating other human beings, by being told by other human beings, by using the same artefacts used

by other human beings. And, since two individuals that learn from one another tend to become more similar in their behaviour and in what they believe and value, groups of individuals who interact together develop a set of shared behaviours, beliefs, and values which is called the culture of the group and which is transmitted from one generation to the next because the members of one generation learn from the members of the preceding generation.

Culture and the cultural transmission of behaviours are so important for human beings that human robotics must necessarily include a cultural robotics. We must construct robots that learn from other robots and we must reproduce the emergence of a shared culture in groups of interacting robots and how this culture changes in a succession of generations. No cultural robotics exists today and this, again, is due to the applied orientation of current robotics. If robots are constructed to help us achieve practical goals, the robots must do what we want them to do. Even if constructing robots that learn from other robots could be practically useful, practical robots cannot choose the robots from which they learn and they cannot have a culture which autonomously evolves and changes. Most current robots are programmed by us to do what they do but, even when the robots learn their behaviours, we decide which behaviours they must learn. Some robots are constructed to interact with us and these robots might learn from us. But, in this case too, it is we who decide what they must learn. Like most human robotics, cultural robotics is a future robotics.

Cultural robotics is one part of social robotics. If the robots must learn their behaviours from other robots, the robots must live and interact together and the possibility of learning from other robots must be one of the most important advantages of living together. We have dedicated the two preceding chapters to living together. In this Chapter we are specifically interested in robots that learn from other robots and in the emergence of robotic cultures.

We begin with robots that learn by imitating other robots. The robots are our basic robots that live in an environment with randomly distributed food tokens. The robots see the food tokens and they must respond by approaching and eating the food tokens. There are two types of robots, adult robots and child robots, and each adult robot carries around with itself a child robot while it moves in the environment searching for food (see Figure 1).

The adult robot already knows how to reach the food tokens. Its neural network has connection weights that allow the robot to respond to the sight of a food token by approaching and reaching the food token. In contrast, the child robot does not know how to reach the food tokens. The connection weights of its neural network are random, which means that, if left alone, a child robot would respond to the sight of food with random movements and would not eat much. The adult robot and the child robot move together in the environment and, therefore, they receive the same sensory input from the environment and they see the nearest food tokens in the same way. When the two robots see a food token, the motor neurons of the neural network of both robots

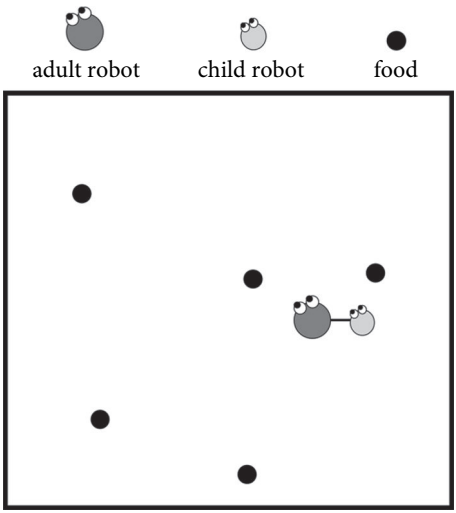


Figure 1. An adult robot carries around a child robot while it is looking for food. The child robot learns how to approach the food tokens by imitating the behaviour of the adult robot.

encode one specific movement but only the movement encoded in the motor neurons of the adult robot is physically executed and the robot actually moves in space. The movement encoded in the motor neurons of the child robot is not physically executed and the child robot moves in space only because it is carried around by the adult robot.

But the child robot is not passive. The child robot looks at how the adult robot responds to the sight of a food token and it learns by imitating the adult's behaviour. The two robots receive the same sensory input from the food tokens but their neural networks have different connection weights and, therefore, their motor neurons encode different movements. The child robot compares the pattern of activation of its motor neurons with the pattern of activation of the adult robot's motor neurons and, on the basis of this comparison, its connection weights change and the changes are such that how the child robot's responds to sensory input from food becomes progressively similar to how the adult robot responds to the same input (see Figure 2). (To make the robots learn, we use the backpropagation algorithm, with the motor output of the adult robot functioning as teaching input for the child robot.) Since the adult robot already knows how to reach the food tokens, the child robot also learns to reach the food tokens. The capacity to reach the food tokens is socially or culturally transmitted from the adult robot to the child robot.

The child robots learn by imitating the adult robots but learning from others goes well beyond that. A robot can learn by imitating another robot whatever its age and whatever the age of the other robot. One robot functions as the model to be imitated and the other robot learns by imitating the model (see Figure 3). The other robot can be a genetically related robot, say, the robot's parent, or a friend, or any other robot.

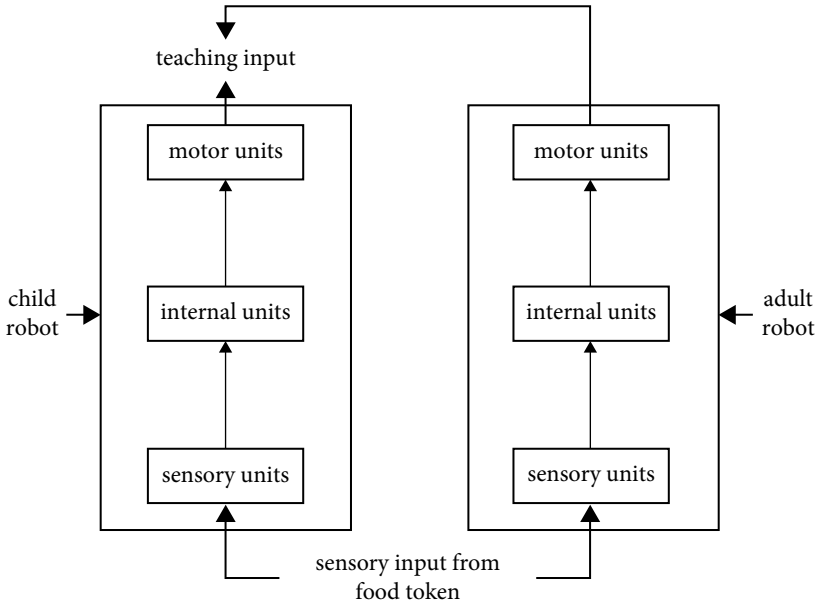


Figure 2. A child robot learns by imitating an adult robot. The child robot compares its behaviour with the adult's behaviour (teaching input) and, as a result of this comparison, its behaviour changes and becomes more similar to the adult's behaviour.

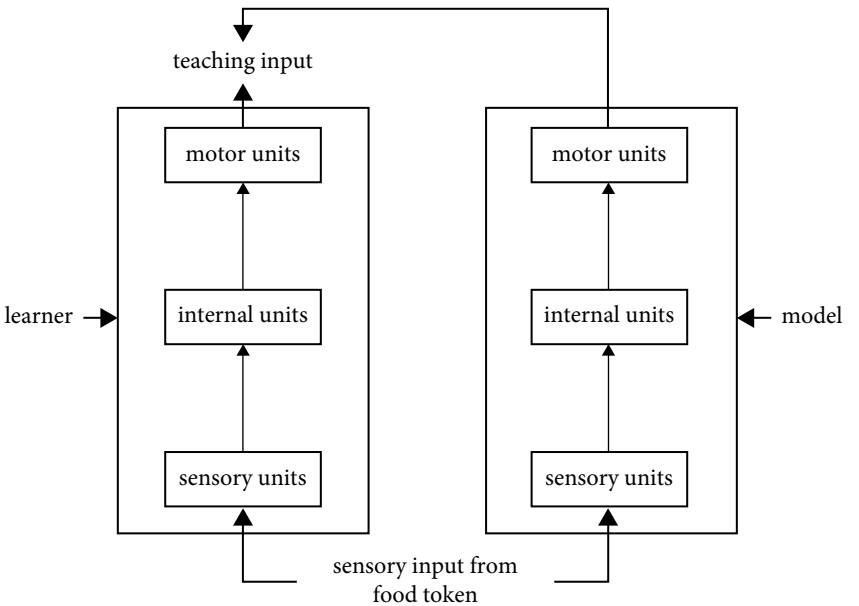


Figure 3. A robot can take any other robot as a model to imitate.

And learning from others can be reciprocal: robot A learns from robot B, and robot B learns from robot A. And what can be learned from others may be not only behaviours, but also beliefs, values, and how to construct and use technological artefacts. Anthropologists use the expression "cultural trait" to refer to anything which can be learned from others, and we will use this term in Section 7 when we will be describe more abstract robots that learn from other robots without specifying what they learn.

If a robot learns by imitating another robot, the robot which is being imitated may have a passive role. It is only a *model* from which to learn. But a robot can also be a *teacher* which actively makes another robot learn by showing or telling the other robot how to behave or by pointing to aspects of the environment which the learner must attend to in order to learn. Imagine that the child robot is learning language. The child robot and the adult robot live in an environment with different types of tokens. The adult robot knows the names of the different tokens but the child robot doesn't. The child robot learns the names of the tokens by imitating the naming behaviour of the adult robot. The adult robot and the child robot see the same token and the adult robot responds by producing one particular sound – the name of the token. The child robot also sees the token and at the same time it hears the sound produced by the adult robot and it responds to these two sensory inputs by producing one sound with its phono-articulatory organs. The neural network of the child robot compares the two sounds, the sound produced by the adult robot and its own sound, and the connection weights of the child's neural network change so that, progressively, the sound produced by the child robot and the sound produced by the adult robot in response to the token become the same. The child robot has learned the name of the token.

The adult robot may name the tokens for its own reasons but it can also name the tokens to make the child robot learn the name of the tokens, and in this case the adult robot is a teacher, not just as a model to be imitated. And the role of the adult robot as a teacher can be even more active. If two different tokens are in front of the robots and the two tokens have different names, the adult robot points with its finger to one of the two tokens and it produces the appropriate sound. In this way the adult robot actively draws the attention of the child robot on the token which is being named and this facilitate the learning of the names of the tokens by the child robot.

2. The cultural emergence of behaviours

So far, the adult robots already know how to reach the food tokens and they already know the names of the tokens, and it is because they already possess these two capacities that these capacities can be transferred to the child robots. But how have the *adult* robots acquired the capacity to reach the food tokens or the capacity to name the food tokens? What we have reproduced so far is the cultural *transmission* of behaviours from one robot to another robot. One robot possesses some behaviour and this behaviour is learned by another robot by imitating the first robot. But human beings do not only

have cultural transmission. They also have cultural creation. The cultural transmission of behaviours leads to the emergence of *new* behaviours. One behaviour does not exist and then it progressively emerges through cultural transmission, leading to a change in the repertory of behaviours possessed by the members of the community. This is what happens with our next robots.

We begin with a group of robots which have random connection weights in their neural network, and therefore no robot is able to reach the food tokens. The robots are born, they live for some length of time, and then they die. When a new robot is born, there is no *genetic* transmission from parents to offspring and the newborn robot's neural network has random connection weights – which means that the newborn robot does not know how to reach the food tokens. The newborn robots learn from the robots of the preceding generation by imitating their behaviour but the problem is that the robots of the preceding generation also have random connection weights in their neural network and therefore they also do not know how to reach the food tokens. So, what can they teach to the robots of the second generation?

Since the connection weights of the robots' neural network are random, each robot responds to the sight of a food token in a different way – and this is the crucial variable. Although no robot possesses the capacity to reach the food tokens, for purely chance reasons some robots have better connections weights than other robots and these robots are slightly better than the other robots at reaching and eating the food tokens. Now we make an assumption. We assume that our robots that carry around the child robots when they move in the environment are not chosen at random but they are chosen *on the basis of their capacity to reach the food tokens*. The robots that, for purely chance reasons, have better connection weights and therefore are better at finding the food tokens are those which are imitated by the robots of the second generation. What these robots have to teach to the robots of the second generation is not much but it is not zero. And their capacity to find food, although very limited, gets transmitted to the robots of the second generation.

And now we make another assumption. When a robot of the second generation learns from a robot of the first generation, the learner does not learn to behave *exactly* like its model. The behaviour of the learner is similar to the behaviour of the model but it is not an exact copy of the behaviour of the model. It has something which is new. This may be because learning is never perfect, because the learner can add something new to what it has learned, because a robot can put together in novel ways what it has learned from different robots, and for other reasons. We reproduce all these factors by adding some random noise to how the movements of the model are perceived by the learner and, in this way, the behaviour of the learner at the end of learning is somewhat different from the behaviour of the model. In many cases the behaviour of the learner is less good than the behaviour of the model but in some rare cases the learner performs better than its model. What is important is that if the learner turns out to be *less* good than its model, it will not be chosen as a model by the robots of the next

generation, whereas if it is *better* than its model it will likely be among the models for the robots of the next generation. In this way, any improvement in behaviour which is due to the random noise that always accompanies cultural transmission is maintained and it spreads throughout the community of robots.

The process of cultural transmission from one generation to the next is repeated for a certain number of generations and at the end we find that all the robots possess the capacity to find the food tokens – with of course individual differences among the robots. This capacity was initially nonexistent but the cultural transmission of behaviour has created the capacity. The two mechanisms of choosing the best individuals of each generation as models for the robots of the next generation and adding random noise to the transmission process, determine a progressive increase in the average number of food tokens eaten by the robots of the successive generations. This is the cultural emergence of behaviours.

The *cultural* emergence of behaviours has many similarities with the *biological* emergence of behaviours. What in biological evolution is copying the genes of the parents into the genes of the offspring, in cultural evolution is imitating the behaviour of the members of one generation by the members of the next generation. In both cases the transmission is selective. Genetically, some individuals have offspring and some don't. Culturally, some members of one generation have imitators in the next generation and some don't. The robots that have offspring are the most successful robots of their generation and the same for the robots that are imitated by the robots of the next generations. In both cases there is "noise" in the transmission process. Random genetic mutations result in offspring which are different from their parents and which, in some cases, are better than their parents. Various factors in cultural transmission guarantee that learners do not behave exactly as their models and some learners turn out to be better than their models. This is what makes cultural evolution similar to biological evolution. But there also are differences. One difference is that, while genetic mutations are random, human beings can predict the consequences of their actions and, therefore, a learner can add purposefully, not randomly, something new to the behaviour learned by imitating a model. Another difference is that an individual has only *two* biological parents, mother and father, while the same individual has *many* cultural models. Having two parents rather than only one introduces novelty in biological transmission because the genes of one individual are a new combination of pieces of the genotype of its mother and pieces of the genotype of its father. Having many cultural models introduces even more novelty in cultural transmission. An individual learns from many other individuals, both of the preceding generation and of his or her own generation – and human beings also learn from dead individuals by interacting with what dead individuals have left in the form of artefacts and written documents. And they recombine together what they learn in novel ways and invent new things. A final difference concerns the size of the pool from which an individual chooses his or her spouse for biological transmission and his or her models to imitate

for cultural transmission. In biological transmission pool size is more restricted, although one can choose one's spouse from one's own group or from another group. In cultural evolution the size of the community to which one belongs and from which one chooses one's models to imitate has a very important role. Later in this Chapter we will show how community size can be an important factor in cultural evolution (Sections 6) and how modern technologies tend to make spatial proximity irrelevant for cultural learning (Section 8). All these factors make cultural change faster than biological change.

3. Staying near to others in order to learn from them

To imitate the behaviour of another robot, a robot must see the other robot and it must also see what the other robot is seeing and is doing – and this require spatial proximity. Today various technologies make it possible for an individual to learn from another individual even if the two individuals are in different places. But spatial proximity remains a favourable condition for learning from others. And what is interesting is that spatial proximity is not only a precondition for learning from others but it can also be a consequence of learning from others because the advantages of learning from others may push individuals to live near to other individuals. So far, in describing our robots that learn from others we have taken spatial proximity as given. Our child robots are carried around in the environment by the adult robots from which they learn. We can imagine that the adult robots that carry around the child robots in the environment are their parents, but only very young children are carried around by their parents. At one year or so, children move in the environment autonomously but they need to remain in proximity to their parents because their parents give them food, protection from dangers, and other goods. (On this, see the preceding chapter on robotic families.) But in human beings offspring stay near to their parents also to learn from them. And the phenomenon is more general. All individuals want to stay near to other individuals in order to learn from them and learning from others is an important reason for living together. This is what we show with our next robots.

The robots live in the same environment and they survive and reproduce to the extent that they possess a capacity which is not genetically inherited but which they learn from other robots. To underscore the generality of the phenomenon, we have chosen an arbitrary and very abstract capacity – the capacity to recognize if two things are the same or different. The neural network of the robots has two input neurons and, in each cycle, each of the two input neurons has an activation level of either 0 or 1. If the pattern of activation of the input neurons is 00 or 11, the response must be 1 – the two things are the same. If they are 01 or 10, the response must be 0 – the two things are different. The robots that have offspring are those which are better at giving the correct responses. At birth, the neural networks of the robots have random connections

weights, which means that no robot is able to recognize if two things are the same or different. To speed up the process, a small number of robots, randomly chosen, are told by us what the correct response is. Using the backpropagation learning algorithm, these robots compare their response with the correct response provided by us and, like the child robots of Section 2, they learn by changing their connection weights so that their responses become progressively similar to the correct responses. The other robots are not taught by us but they learn by imitating other robots. A robot compares its response with the response of another robot, randomly chosen, and the connection weights of its neural network change in such a way that its responses become more similar to the response of the other robot. What we expect is that the capacity to recognize if two things are the same or different that we have taught to only some of the robots will progressively diffuse in the entire community of robots.

But there is a problem. A robot can learn from another robot only if it is physically near to the other robot. The learner must perceive the same two things which are perceived by its model and it must perceive how the model responds to the two things – and this is possible only if the two robots are near to one another. Therefore, to learn, the robots must approach and remain in proximity to the other robots. The robots that remain isolated cannot learn from other robots and they are unlikely to survive and reproduce. The robots' neural network is made up of two modules (see Figure 4).

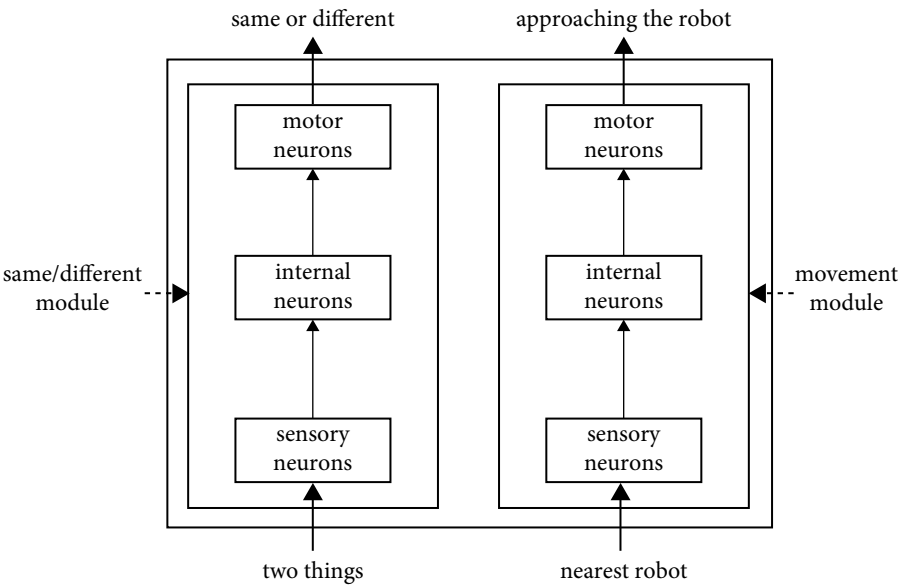


Figure 4. The neural network of a robot is made of one module with which the robot learns from other robots if two things are the same or different and another module with which the robot approaches other robots in order to learn from them.

One module is the module which allows the robot to learn to recognize if two things are the same or different. The connection weights of this neural module are assigned randomly at birth and they change during the robot's life because the robot learns by imitating other robots. The second neural module controls the movements of the robot in the environment. The visual neurons of this module encode the location of another robot which is near to the robot and the motor neurons encode the movements that allow the robot to displace itself in the environment. The connection weights of this second neural module are genetically inherited from parents to offspring and they do not change during the robots' life but they evolve in a succession of generations in which the robots that are selected for reproduction are those that are better at recognizing if two things are the same or different. Therefore, the robots are not selected for reproduction on the basis of their capacity to stay near to other robots but they are selected on the basis of their capacity to recognize if two things are the same or different – and this is a capacity which they learn during their life.

At the beginning the robots do not possess neither the capacity to stay near to other robots nor the capacity to recognize if two things are the same or different. They move independently from one another and their capacity to recognize if two things are the same or different does not increase very much because they are not near to other robots and, therefore, they cannot learn from other robots. Then, in a succession of generations the scenario progressively changes. The robots biologically evolve a tendency to remain in proximity to other robots because in this way they can learn from other robots to recognize if two things are the same or different (see Figure 5).

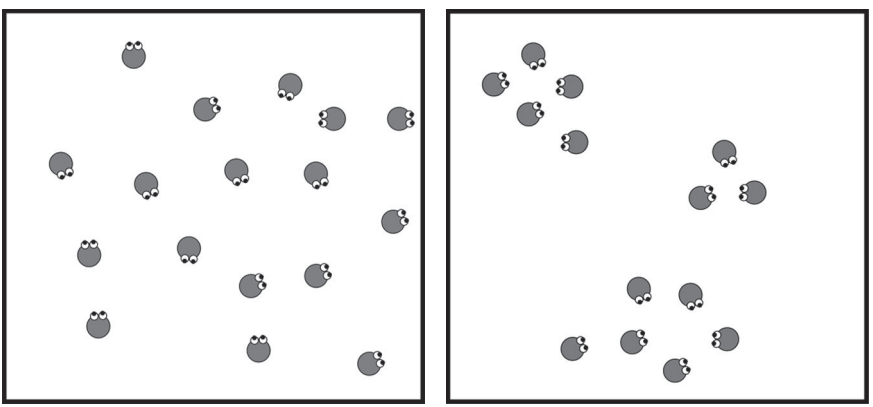


Figure 5. At the beginning the robots are dispersed in the entire environment (a) but then, since they can learn from another robot only if they are near to the other robot, they evolve the behaviour of staying near to other robots in order to learn from them (b).

These robots tell us two things. First, living in groups makes it possible to learn from others and, since human beings learn from others most of what they do and they know, human beings tend to live in groups. Second, they tell us that biological evolution and cultural learning interact together. In our robots there is no *direct* selective pressure for the *biological* emergence of the behaviour of staying near to other robots. The robots that have more offspring are not the robots that stay near to other robots but the robots which are good at recognizing if two things are the same or different. But the capacity to recognize if two things are the same or different is learned from others and learning from others requires spatial proximity. Therefore, there is an *indirect* pressure for the biological emergence of the behaviour of staying near to other robots. (See Chapter 2 dedicated to robots that both evolve and learn.) Learning from others causes the biological emergence of the behaviour of staying near to other robots because only if two robots are near to one another they can learn from one another.

4. Should adolescents learn from adults or from other adolescents?

Children do not choose their parents as models for learning because they necessarily learn from their parents but, when they grow up and become adolescent, their freedom of choice increases. And, as we have seen, the cultural emergence of behaviours is based on choosing the best robots of the *preceding* generation as models for learning. But if the generations overlap and adolescent robots live with the robots of the preceding generation, an adolescent robot can learn from the robots of the preceding generation but it can also learn from the robots of its own generation – its peers. What are the consequences of these two different forms of cultural learning? Should a robot learn from older robots or from its peers?

The answer depends on the environment in which the robots live and, in particular, it depends on whether the environment is always the same or it changes from one generation to the next. The environment contains both food tokens and poisonous tokens and, to survive and have offspring, the robots must eat the food tokens and avoid the poisonous tokens. We evolve two different populations of robots that live in two different environments. In one environment the food tokens are always black and the poisonous tokens always white. In the other environment, the colour of the food and poisonous tokens changes unpredictably from one generation to the next (see Figure 6).

The capacities to eat the food tokens and avoid the poisonous tokens are not biologically evolved and genetically inherited but they are culturally evolved and learned from other robots. At birth, all robots have random connection weights in their neural network and therefore they do not know which tokens are food and which are poison. But, like the robots described in Section 3, they learn the two capacities by imitating

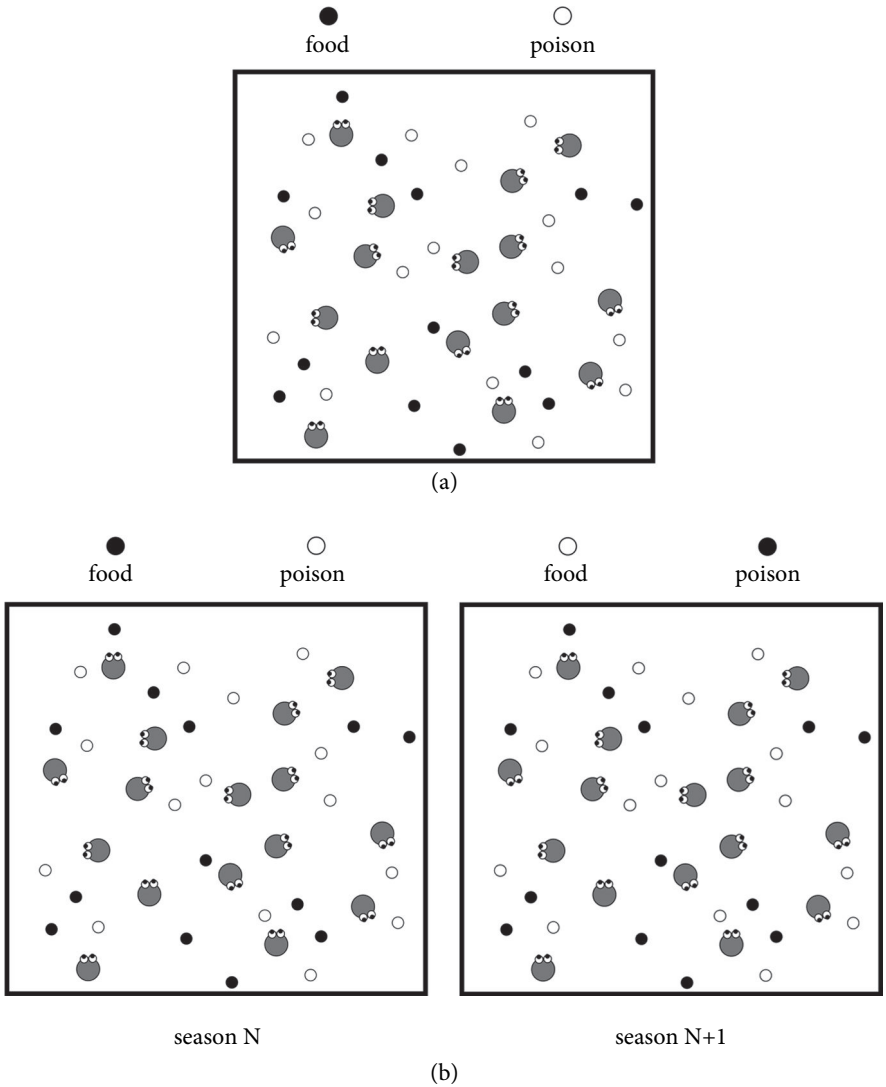


Figure 6. The robots live in an environment in which the food tokens are always black and the poisonous tokens always white (a). The robots live in an environment in which the colour of the tokens changes unpredictably from one generation to the next (b).

other robots. The robots of each generation choose the best robots of the preceding generation as models from which to learn and they add some random noise to what they learn which in some cases makes a learner better than its model. This leads to the cultural emergence in a succession of generation of the capacity to approach and eat the food tokens and to avoid the poisonous tokens.

But from whom the robots learn? We compare two populations of robots. In one population the robots which are chosen as models to imitate are the best robots of the *preceding* generation (adults). In the other population the robots that play the role of models are the best robots of the *same* generation (peers). What are the results? The results are that, if the environment does not change and the food and poisonous tokens have always the same colour, learning from the robots of the preceding generation gives much better results than learning from the robots of one's own generation (see Figure 7).

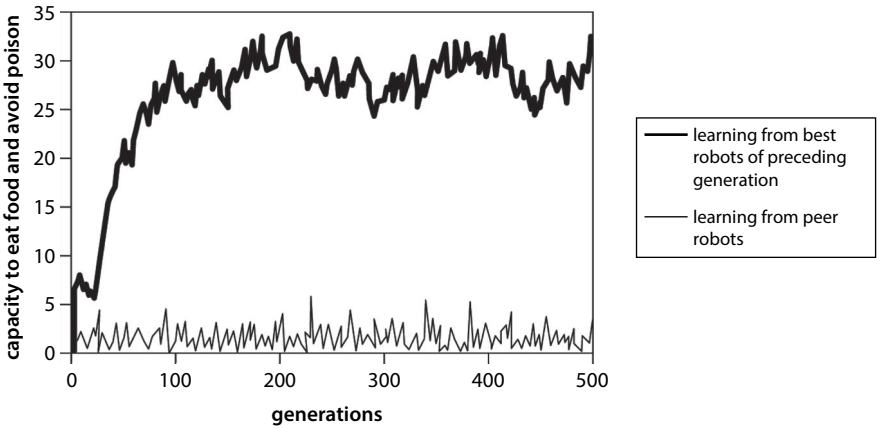


Figure 7. Cultural evolution of the capacity to eat the food tokens and avoid the poisonous tokens for robots that learn from the best robots of the preceding generation and for robots that learn from their peers.

This is not surprising since the robots which are chosen as models are adult robots that already possess the two capacities so that the robots of the new generation can learn something useful from them. On the contrary, if the robots learn from their peers, their peers do not have much to teach because their neural network also has random connection weights. And even when some useful behaviour emerges when the robots learn from their peers, this behaviour is not transmitted to the next generation because for the robots of one generation the robots of the preceding generation are adults, and these robots do not learn from adults but from their peers. When the robots learn from their peers, each new generation starts from zero and there is no culturally cumulative process.

But, even in the environment in which the colour of the food and poisonous tokens does not change from one generation to the next, learning from one's peers can have some usefulness. Like biological evolution, cultural evolution needs variation in what is culturally inherited. In the case of biological evolution this variation is the addition of random mutations to the inherited genes which in some cases result in better genes that help evolution to go forward. In the case of cultural evolution, the learners do

not imitate exactly their models but they introduce some variations and these variations, even if they are random, may result in better behaviours. The importance of noise is clearly demonstrated if we eliminate noise. If we eliminate noise in biological transmission, the genes of the robots of the second generation are exact copies of the genes of the best robots of the preceding generation but evolution stops there. There are no further changes and no cumulative improvements. The same is true for cultural evolution. If the robots of the new generation copy exactly the behaviour of the best robots of the preceding generation without introducing any change in this behaviour, there is no further improvement and no further cultural evolution.

This is what we find with our robots. If we eliminate the random noise which accompanies the imitation process, even if the robots of the new generation copy the behaviour of the best robots of the preceding generation, the capacity to eat the food tokens and avoid the poisonous tokens does not evolve. And this is why learning from one's peers can be useful. If we allow the robots of each generation to sometimes learn from their peers rather than from the robots of the preceding generation, learning from peers has a positive role. It adds the random noise which is needed by all evolutionary processes to produce new and better behaviours.

But the importance of learning from one's peers rather than from adults becomes much greater if after a certain number of generations the environment changes. Imagine that the robots live in an environment in which the food tokens are black and the poisonous tokens are white. The young robots of each generation take the best members of the preceding generation as their models for learning and this leads to the emergence of the behaviour of eating the black tokens and avoiding the white tokens. But, after a certain number of generations, the environment suddenly changes: the black tokens become poisonous and the white tokens edible. What happens is easy to predict. The behaviour of the robots of the generation in which the change occurs is disastrous. The robots learn from the robots of the preceding generation but the behaviour that they learn from the robots of the preceding generation is inappropriate to the new environment. The robots approach and eat the black tokens, which are now poisonous, and they avoid the white ones, which have become edible. In the successive generations the situation gradually improves because the robots which, as a result of the random noise that is constantly added to the process of cultural transmission, exhibit a behaviour which is more appropriate to the new environment become the models for the robots of the next generation. However, the improvement is very slow and it takes many generations.

This is where the importance of learning from peers rather than from adults becomes clear. If when the environment changes the young robots continue to learn from the adult robots, the evolutionary emergence of the newly appropriate behaviour is very slow. In contrast, if when the environment changes some of the young robots learn not from the robots of the preceding generation but from the robots of their own generation, their peers, the behaviour which is necessary to survive in the *new* environment emerges more quickly, in fewer generations (see Figure 8).

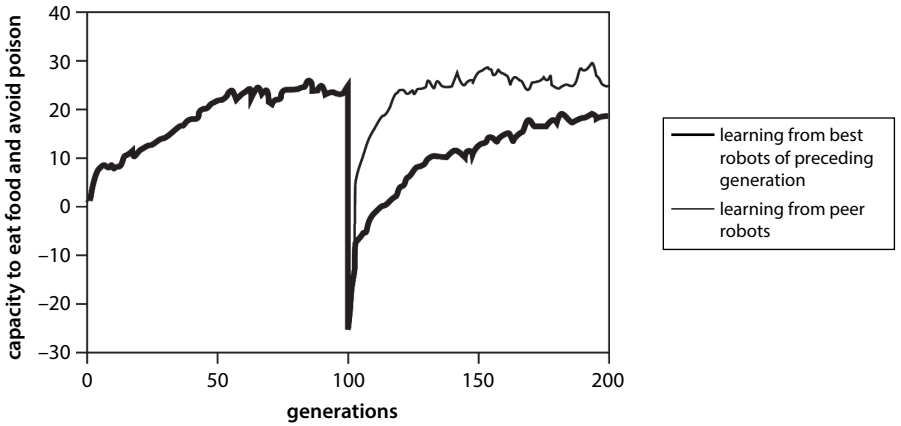


Figure 8. When, after 100 generations, the environment changes, if the robots continue to learn from the adult robots, their capacity to eat the new food tokens and to avoid the new poisonous tokens takes more time to emerge than if they learn from their peers.

These robots tell us something interesting about today's societies. In traditional societies, the young learn from adults. In today's societies the young tend to learn from their peers rather than from adults. Our robots can be of some comfort for parents, teachers, and adults generally, who are preoccupied with this phenomenon. The tendency of young people to learn from their peers rather than from adults may have adaptive value in rapidly changing environments such as today's societies because adults may not have much to teach to young people and therefore it does not make much sense to learn from them.

But our robots tell us that the component of intra-generational transmission – learning from peers – and the component of inter-generational transmission – learning from adults – must be properly dosed. If the young learn only from their peers, there is no evolution because there is no manner to preserve any suitable behaviour which has emerged in the past. If, on the contrary, the component of intra-generational transmission is completely absent and young people learn exclusively from adults, adaptation to new environments is slow and creates many problems: our robots continue to eat the black tokens although the black tokens have become poisonous and they avoid the white tokens which now are those that contain the energy they need to remain alive. The solution seems to be an appropriate proportion of intra-generational and inter-generational transmission, with more intra-generational transmission if the changes in the environment are fast – as they are today. Inter-generational transmission ensures that what is still useful is maintained and that the evolutionary process is cumulative. But intra-generational transmission makes it possible to remove existing but no longer suitable behaviours and to facilitate the emergence of new and more appropriate behaviours.

5. The evolution of artefacts

Human beings are different from nonhuman animals because most of their behaviour is learned from other human beings and is not genetically inherited or learned by interacting with the natural environment. But human beings are different from nonhuman animals for another behaviour which derives from their capacity to learn from others. Human beings create all sorts of technological artefacts, and their survival and well-being depend on their ability to construct useful artefacts. How are learning from others and making artefacts linked together?

Making artefacts and the cultural transmission of behaviours are linked together in many ways. *Using* an artefact requires to be able to use the artefact. A hammer requires that one is able to use the hammer to push a nail on the wall. A vase requires that one is able to store something in the vase and to transport the vase from one place to another without breaking the vase. A computer or a mobile phone requires that one is able to use the computer or the mobile phone. Therefore, artefacts are vehicles for the cultural transmission of behaviours. If an individual constructs an artefact that to be used requires certain behaviours or capacities, another individual who wants to use the artefact needs to acquire these behaviours and capacities. In this way, new behaviours and capacities are transmitted from one individual to another individual and they diffuse in the society together with the diffusion of the artefacts.

Another way in which making artefacts and the cultural transmission of behaviours are linked together concerns not the *use* of artefacts but the *construction* of artefacts. Constructing an artefact requires specific capacities, and these capacities are culturally learned by observing another individual when he or she constructs the artefact and imitating the behaviour of the other individual. But it is not only behaviours and capacities which are culturally transmitted because they are needed to use and to construct artefacts. The artefacts themselves are culturally transmitted and, in this case, *cultural* transmission becomes *technological* transmission. Human beings do not only copy the behaviours of others but they also copy the artefacts made by others. They see an artefact and they make a copy of the artefact.

The similarities between cultural transmission and technological transmission do not stop here. As there are *cultural* evolution and the cultural *emergence* of new behaviours (Section 3), so there are *technological* evolution and the *emergence* of new artefacts. Artefacts exist to make human beings have a better life. But this requires that the artefacts are *good* artefacts. Initially, a community of individuals may construct artefacts which are not particularly good. But, then, if the artefacts of one generation which are copied by the next generation are the best artefacts of the preceding generation and the copies include some variations that at least in some cases result in artefacts which are better than the original artefacts, the quality of the artefacts tends to improve generation after generation. This is *technological evolution*. Our next robots reproduce technological transmission and technological evolution and they show how living in larger communities may accelerate technological evolution.

The new robots have a biologically evolved capacity to find the food tokens which exist in the environment. But these robots also construct containers of food (vases) which they use to store, cook, and transport the food tokens. These vases are very useful because by using them a robot can increase the quantity of energy extracted from the food it finds in the environment and, as a consequence, its survival and reproductive chances. However, the vases are not all identical. They have different characteristics and these characteristics determine their quality. The vases vary in their capacity to conserve or cook food, in how easily they can be transported from one place to another, in how much they resist breaking. The best vases allow the extraction of more overall energy from food while low quality vases add little to the natural energy contained in the food or they may even lead to a loss of this energy – when a vase with food breaks. In this scenario the survival and reproductive chances of a robot depend on two capacities: the capacity to find food and the capacity to construct and use vases of good quality. (We assume that each robot uses the vases that it constructs and there are no robots which specialize in the construction of vases which they then sell to other robots. For specialization and exchange among robots, see Chapter 10). The capacity to find food in the environment is biologically evolved but the capacity to construct vases is learned from others and is culturally evolved. A robot constructs its vases not by starting from zero but by copying the vases constructed by the members of the preceding generation of robots. This is the cultural transmission not of behaviours but of artefacts.

The robots have a neural network made up of two modules (see Figure 9).

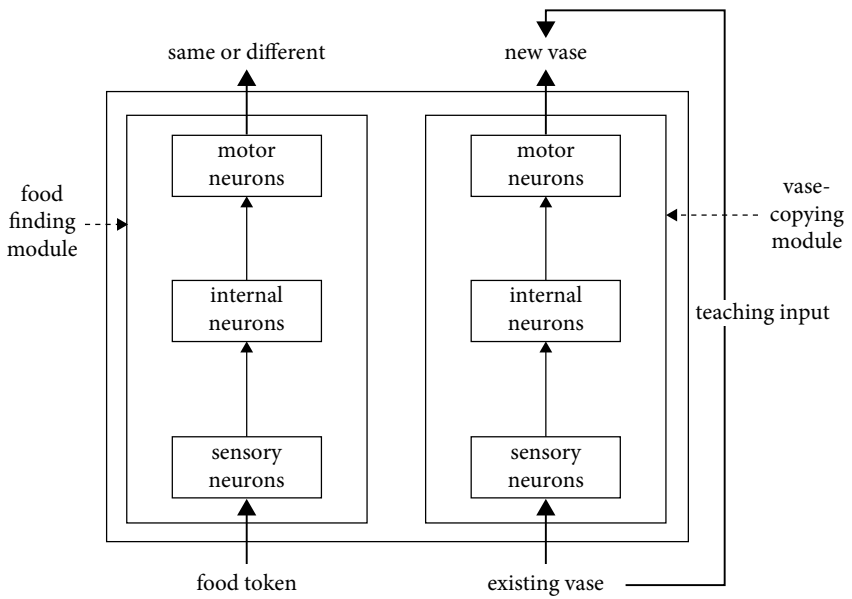


Figure 9. Neural network of robots that make copies of existing vases. The left module is used to find food. The right module to make copies of existing vases.

One module controls the behaviour of the robots when they look for food. The sensory neurons of this module encode the location of the nearest food tokens and the motor neurons encode the movements with which the robots approach and collect the food tokens. The connection weights of the module are genetically inherited and they do not change during the robots' life. The robots of the initial population have random connection weights in their food-finding neural module but the capacity to find food evolves biologically in a succession of generations of robots which inherit the connection weights of their parents with random genetic mutations.

The second module of the robots' neural network controls the behaviour of the robots when they construct a vase by copying an already existing vase, and the connection weights this second module are not genetically inherited but they are randomly assigned when the robot is born. The robot sees an existing vase and it responds by constructing another vase which is like the vase chosen as model. Its neural module for constructing vases has sensory neurons which encode the characteristics of the vase chosen as the model to imitate and motor neurons which encode the movements which allow the robot to construct a new vase which is a copy of the model. The robot learns how to construct a copy of an existing vase because the robot compares the vase it has constructed with the original and the connection weights of its vase-constructing module change gradually so that the copies become progressively more similar to the original. (We use the backpropagation algorithm to make the robots learn to copy the vases chosen as models.)

At the beginning, we generate a certain number of different vases and we arbitrarily decide which vases are good vases, those which increase the survival/reproductive chances of the robot which uses them. (These properties might be the materials with which the vase is made or its shape.) We contrast two communities of robots. In one community, the vases which are chosen by a robot as models to be copied are those that have been constructed by the robot's parent. A robot does not search for the best vases existing in its community but it reproduces the vases used in its family. In the other community, the vases that a robot chooses as models to imitate are the best vases existing in the entire community of robots.

What are the results? The results are that technological transmission which is restricted to the family is less effective than technological transmission which takes place at the level of the entire community. If the robots make copies of the vases constructed by their parents, the quality of the vases improves slightly in the very first generations but then it stabilizes at a rather low level. In contrast, if the vases used as models for reproduction are the best vases existing in the community, their quality continues to improve across a succession of generations and at the end it is much higher (see Figure 10).

What explains these results? A robot's parent obviously is a robot which has been able to have offspring, and this means that its vases are not too bad since the probability to survive and have offspring depends on *both* the robot's capacity to find food and the quality of its vases. But since having offspring depends on two factors, it may be that the robot's parent has been able to have offspring *not* for the quality of its vases but

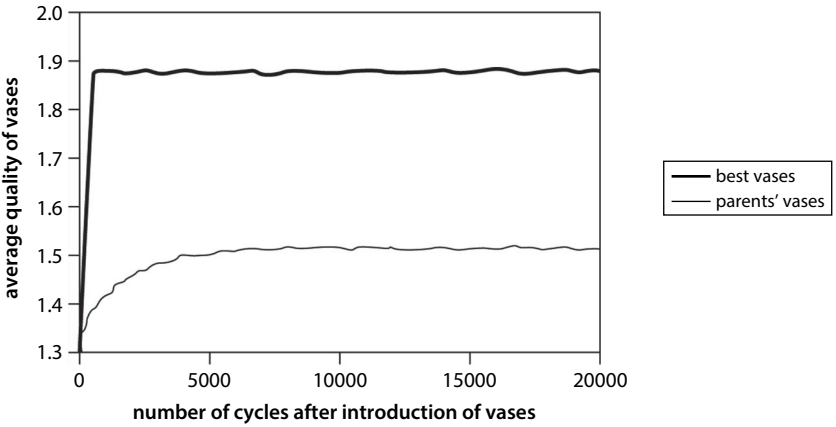


Figure 10. The vases of the robots that copy the best vases existing in their entire community are better than the vases of the robots that copy the vases made by their parents.

for its ability to find food. And, in any case, even if the vases of one's parent are good vases, there may be other vases in the community of robots which are better. Like learning from one's parents, technological transmission inside the family does not make it possible to choose the models from which to learn. A robot cannot but reproduce the vases which exist within its family. In contrast, technological transmission at the level of the entire community allows a robot to choose the best vases existing in its community as models to imitate and these vases may be better than the vases used in its family. This predicts that technological and, more generally, cultural evolution will be quicker and better if the best vases which exist in the entire community are taken as models to reproduce rather than those existing in one's family – and this prediction is confirmed by our results.

If a robot copies the vases of its parent, the robot has no choice to make. But if technological – and, more generally, cultural – transmission takes place outside the family, this poses the question of how to choose the vases to imitate. Our robots choose the best vases existing in their community, where "best" means the vases that increase the most a robot's survival and reproductive chances. But it is we, the researchers, who decide which vases are the best vases. What if the robots themselves choose the vases to be reproduced? How can a robot know which are the best vases?

A robot might try all the existing vases and select the best vases as models to be copied. But this is a long and impossible task. There must be shortcuts. One shortcut is to look at the energy possessed by the different robots. The energy – the well-being or wealth – possessed by a robot is something which is not too difficult to know, and a robot which has more energy is probably a robot which uses good vases. The robots use the observed energy of a robot as a proxy for the quality of the robot's vases and, in choosing the vases to be copied, they adopt the criterion of copying the vases of the robots which have more energy.

This criterion for choosing the vases to reproduce may be easy to apply but it is less informative than the ultimate criterion of trying all the vases. A robot which has a lot of energy can be a robot which uses good vases but it can also be a robot which is very good at finding food while its vases are not necessarily particularly good. This is what we find with our robots. We contrast two populations of robots. In both populations a robot chooses the vases to copy from among all the vases used in its community but in one population the robot copies the vases which actually increase the most its survival and reproductive chances while in the other population the robot copies the vases which are used by the robots that possess the most energy (wealth). As predicted, in both cases the quality of the vases improves across the successive generations of robots – and of vases – compared to the robots that copy their family's vases but it improves much more and more quickly in the first than in the second population. If a robot has a high level of energy, it probably uses good vases, but adopting this criterion has its limits. The problem is that for our robots – and for human beings – the more direct and effective criterion for choosing what to copy by trying all of the possible models, is not accessible. This inevitably makes technological evolution, and cultural evolution more generally, less than perfect.

It is interesting to compare these robots with human beings. A human being copies – or, better, buys – the artefacts used by the richest people or by the people who have more prestige or appear to live a better life even if he or she does not really know if the artefacts used by these people are the best ones. Human beings are very social animals and, for them, using the same artefacts used by the richest and most prestigious people is by itself advantageous. Human beings obtain most of what they want from other people, and other people are more disposed to give them what they want if they use the same artefacts of the richest and most prestigious people.

But in all human societies and, especially, in today's human societies, this criterion for choosing one's artefacts tends to be replaced by another criterion which consists in copying and using the artefacts used by the *majority* of people in one's community. This is the conformist criterion, and we will discuss this criterion of cultural and technological evolution in the next section.

As we have seen, if technological transmission is restricted to the family, technological evolution is slower and it results in vases of lower quality compared to technological transmission which goes beyond the family and extends to the entire community. But the problem is not family vs. community but the size of the community within which the robots chooses the vases to copy. Even if the robots go outside their family to choose what to copy, the size of the community in which the robots live influences technological evolution.

This is what we find if we compare robots that live in smaller communities and robots that live in larger communities. In both cases the robots choose the vases to copy from the entire pool of vases which exist in their community and in both cases the quality of the vases improves across a succession of generations. But the quality of the vases improves more rapidly and reaches a slightly higher value if the robots live in larger communities (see Figure 11).

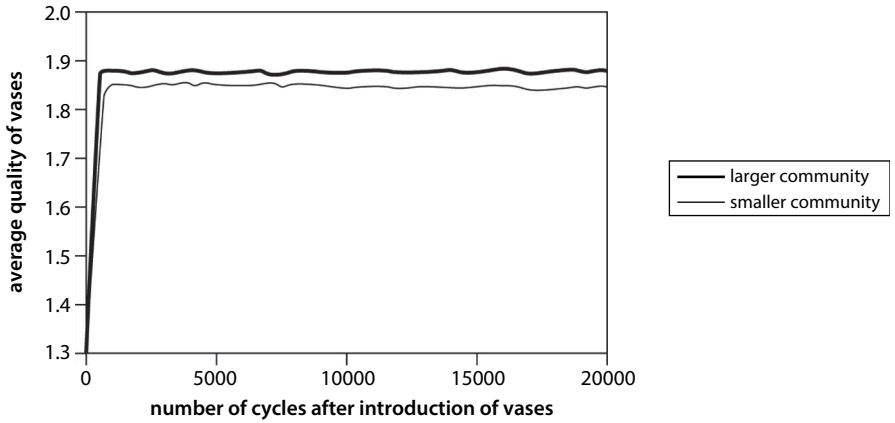


Figure 11. Evolution of the quality of vases as a function of the size of the community. Larger communities have better vases than smaller communities.

This is confirmed when we compare the number of robots living in a larger community with the total number of robots living in a series of smaller communities. At the beginning the robots of the larger community are as numerous as the total number of robots of the small communities but very soon the robots that live in the large community and choose to copy the best vases existing in their large community are more numerous than the total number of robots that live in the separate small communities and copy the best vases of their small communities (see Figure 12).

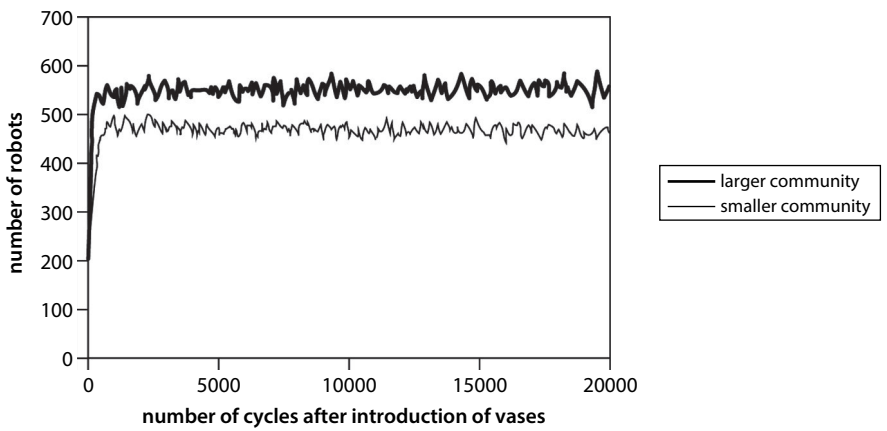


Figure 12. The number of robots that live together in a large community is greater than the total number of robots that live in many separated small communities. The robots copy the best vases existing in their community.

This may explain why, historically, human communities have increased in size. Larger communities offer a larger number of different behaviours and different artefacts from which to choose one's models to imitate and this leads to the emergence of more effective behaviours and better artefacts. The increase in the size of human communities began with agriculture and animal domestication – which require specific behaviours and specific artefacts – but it has continued with the increase in the number of different technological artefacts.

But what is interesting is that not only population size leads to better technological artefacts but also technological artefacts lead to an increase in population size. If we compare a population of robots that live in an environment with both food tokens and food-producing tools with a population of robots that live in an environment with only food tokens, we find that the first population has a larger size than the second population because the food-producing tools increase the number of food tokens available to the robots and the robots can eat more and live longer. This is what has happened in the long history of human beings and it is happening at an increasing speed today.

6. Cultures

For animals which are biologically predisposed to learn from others such as human beings, social interaction necessarily leads to cultural homogeneity. A group of individuals start with different traits (behaviours, artefacts and, if these terms can be translated into robotic terms, beliefs and values) and then learning from others causes the individuals to progressively have the same, or very similar, traits. A collection of traits which are shared by a group of individuals because the individuals learn these traits from other individuals, is a culture. Cultures emerge and are maintained because, although learning from others requires interacting with others, interactions can be indirect. If A interacts with B and B interacts with C, A also interacts with C. This means that a shared culture can emerge in a set of individuals even if an individual does not interact with all the other individuals.

Traditionally, social interaction depends on space. Two individuals are more likely to interact together if they live near to one another – although for today's human beings this is less and less true. In this Section we describe robots which interact together based on their spatial proximity and we see what are the consequences for their cultures.

The robots that learn from their parents and those that copy the vases made by their parents do not choose their models for learning. They simply copy the behaviour of their parents and the vases that exist in their family. But we have also described robots that *choose* their models of learning. They look for the best robots of the preceding generation – inside or outside their family – and learn from these robots or they look for the best vases existing in their community and copy these vases. But there

is another criterion for choosing one's models for learning which does not require "look for the best and copy the best", and this is the conformist criterion: "copy the behaviour of the majority of individuals around you". If one uses this criterion, there is no problem of evaluating what is good and what is bad, what to copy and what not to copy. You just do what the majority of people do. The conformist criterion is easy to apply because you can easily see what the majority of people around you do and this criterion may also be a substitute for the goodness of a behaviour because, if the majority of individuals behave in a certain way, it is probable that this way of behaving is a good one – or at least not a bad one. And the conformist criterion is self-reinforcing because, if an individual chooses to imitate what the majority of individuals do, this majority is automatically enlarged.

But the conformist criterion for deciding what to learn from others has another, more specific, advantage which explains why it is so often adopted by human beings. For the robots described so far, the behaviours and the vases which they copy from other robots directly increase their probability to survive and live well. Being able to find food and constructing/using vases to store or transport food is advantageous even for robots which live alone and evolve their capacity to find food and to make vases without external help. But human beings do not live alone. They live with others, and much of their chances to survive and live well depend on others. Why living with others makes the adoption of the conformist criterion in cultural transmission intrinsically advantageous? The answer is that the conformist criterion guarantees that all the members of the community behave in the same way, use the same artefacts, and have the same beliefs and values. The uniformity of behaviours, beliefs, values, and artefacts is a good thing for at least two reasons. First, it makes it easier for an individual to *predict* the behaviour of other individuals, and being able to predict the behaviour of other individuals makes it easier to interact with them and to coordinate one's behaviour with their behaviour. Second, it creates a sense of *identity* in all the members of the community, and a sense of identity makes it possible for the community to better compete with other communities, for example, in making wars with other communities.

We now describe robots that learn from others by adopting the conformist criterion. Since we are not interested in specific behaviours or in specific artefacts but we want to reproduce the emergence of cultures as sets of traits which are shared by a community of robots, the robots are more abstract than our other robots. In fact, the new robots are "agents" rather than robots because they do not have a body and their behaviour does not consist in movements of their body. Each robot has a set of cultural traits which can be behaviours, artefacts, beliefs or values, and this set of traits is represented as a string of bits, where a bit is either 0 or 1. If a robot possesses one particular trait, the bit of its string which represents the trait has a value of 1 – presence of the trait. If the robot does not have the trait, the value is 0 – absence of the trait. The robots change their traits because they interact with other robots, and the rule for changing one's traits is the conformist criterion: at each time step, randomly

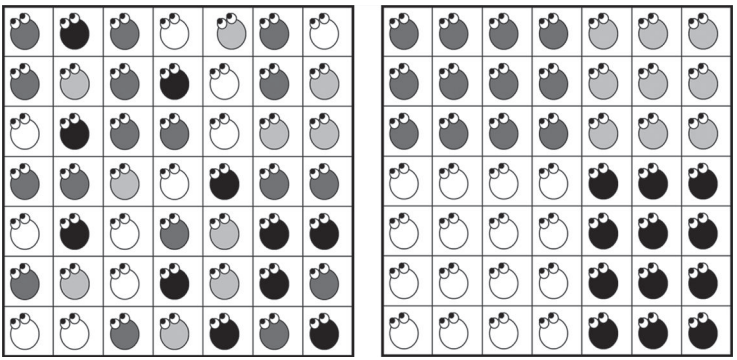
choose one bit of your string of bits, look at what is the value of the bit among the majority of robots around you, and adopt that value. If the robots around you have one trait (1) and you don't (0), adopt the trait (change the value of your bit from 0 to 1). If they don't have the trait (0) and you do (1), abandon the trait (change the value of your bit from 1 to 0). If you have the same trait as the majority of robots around you, do not change.

Although these robots are not really robots but they are "agents" without a body, they live in physical space, and we want to explore the role of physical space in the emergence of cultures. Our question is: With whom a robot interacts? Do the robots interact with the robots which are physically near to them, or are the interactions among the robots independent of physical distance? We have defined culture as a set of traits which are learned from others and which are shared by a group of robots. What are the consequences of learning from physically near robots instead than from robots which can live anywhere?

The robots live in an environment made up of square cells and each cell contains one robot. As we have said, these robots are completely abstract. They are strings of bits which represent their behaviours, their artefacts, their beliefs, and their values. A robot interacts only with the eight robots which live in the eight cells around its cell. At the beginning the value of each bit – 0 or 1 – of the robots' string of bits is assigned randomly and, therefore, the robots tend to have different traits. Then, the robots change their traits by using the conformist principle: "For each of your traits, change the value of the trait if it is different from the value of the majority of robots which are around you, and keep the original value otherwise". The traits of the robots begin to change but, after a certain number of cycles, there are no more changes and what we find are robotic cultures. There are groups of robots which live near to one another and have the same traits – the same string of bits – because they have copied the traits of the robots which are near to them.

But if our expectation is that the conformist criterion leads to the emergence of a single homogeneous culture in *all* robots, this expectation is not fulfilled. Although a certain number of cultures emerge in groups of robots that live in cells which are near to one another, there is not one single culture which includes all robots (see Figure 13). Why? The conformist criterion leads to the emergence of local robotic cultures and, after these local cultures have emerged, it is too late for the creation of a single unified culture. Imagine robot A which belongs to a certain culture and it lives near to robot B which belongs to a different culture. Robot B represents a pressure to change for robot A but this pressure is overwhelmed by the conservative pressure from the other, more numerous, robots with which robot A interacts and which share its same culture. The local cultures are separated by stable and impenetrable boundaries which make the emergence of a single homogeneous culture impossible.

Cultural traits change not only because of the interactions among individuals but also for other factors such as "errors" in cultural transmission – one individual makes



(a) (b)

Figure 13. At the beginning robots with sets of different traits (white, black, light grey, dark grey robots) are randomly distributed in the environment (a). The robots learn from the robots which are near to them and, at the end, we see the appearance of four different cultures (b).

an inexact copy of one of the cultural traits of another individual – and the invention of new traits – new behaviours, new artefacts, new beliefs, and new values. We capture all these factors of change by adding some random noise to the cultural imitation process which takes place among the robots. In each cycle, there is some probability that the value of any trait (bit) of any robot will autonomously change from 1 to 0 or from 0 to 1. What are the consequences of these intrinsic changes for the contacts among local cultures and for the emergence of a single global culture? Intrinsic changes in culturally transmitted traits should make cultures less internally homogeneous and cultural boundaries between different neighbouring cultures more penetrable. And, in fact, we find more cultural homogeneity among the robots. But, although the number of local cultures is reduced, cultural boundaries still exist and we don't see the emergence of a single unified culture.

Other factors can influence cultures and cultural change. For example, the environment can contain mountains, deserts, seas, and rivers which are an obstacle to social interaction and, therefore, to cultural influence. We reproduce these geographical obstacles to social interaction as lines that separate neighbouring cells and make social interactions between the two sides of a line impossible, and what we find is that the number of local cultures increases (see Figure 14). (These robots replicate historical phenomena such as the existence of many different languages in territories such as the island of Papua New Guinea which has many valleys separated by high mountains.)

Another factor which influences cultural phenomena is territorial expansion: populations expand to new territories. We reproduce territorial expansion by placing one robot in one of the cells of the environment and leaving all the other cells empty.

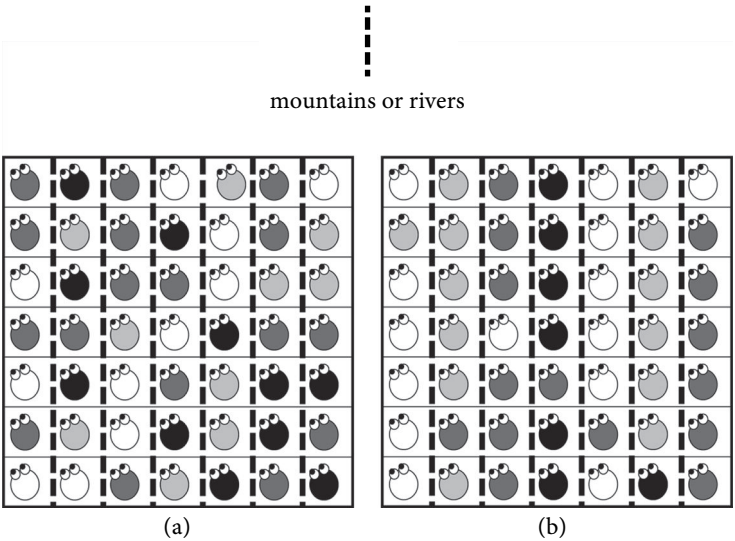


Figure 14. The robots live near to other robots but they are separated by mountains or rivers which make interactions between them difficult or impossible (a). This leads to the emergence of local cultures which reflect the nature of the environment (b).

The robot generates an offspring which is placed in one of the eight cells around the parent's cell and the offspring robot inherits the cultural traits – string of bits – of the parent robot. Then the new robot in its turn generates an offspring which is placed in one of the (empty) cells near to its parent's cell. This is repeated for a certain number of cycles and the process of expansion to new territories goes on until all cells are filled with robots. (Notice that, when a parent robot dies, it leaves one of its offspring in its cell.) What are the consequences of territorial expansion for culture and cultural change? If all the interactions are among culturally identical robots and there is no cultural innovation, cultural change is impossible and at the end all the robots have the same culture, which is the culture of the original robot (see Figure 15).

But if, more credibly, we add intrinsic cultural changes under the form of random noise, we find that in this case too a number of different cultures emerge, with stable boundaries between neighbouring cultures – and this process of cultural diversification is facilitated by the existence of geographical obstacles such as mountains, rivers, and seas. These robots replicate another important cultural phenomenon: in many cases the different cultures which exist in different parts of a given territory are historically derived from a single local culture which has expanded to the entire territory and has given birth to different cultures. (We return to this phenomenon in the last section of this chapter.)

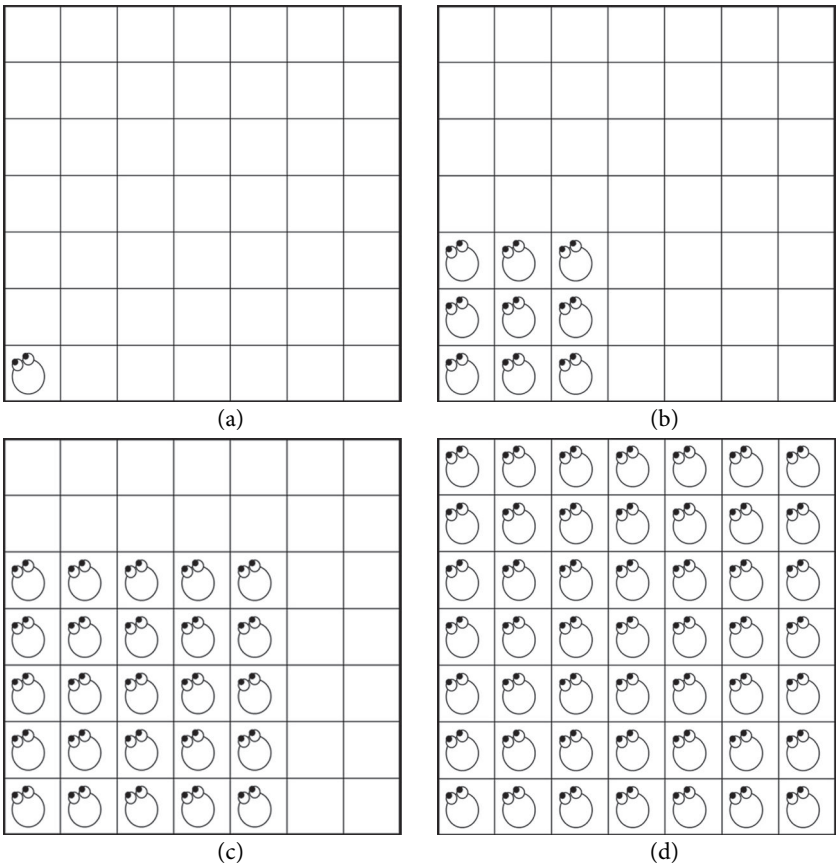


Figure 15. Four phases – (a), (b), (c), and (d) – of the progressive territorial expansion of robots with the same culture.

But the consequences of intrinsic cultural changes – random noise – are different in this new scenario of territorial expansion. In the preceding scenario in which all the cells were already occupied by the robots, adding random noise led to a *decrease* in the number of different local cultures. In the new scenario of geographical expansion, random noise leads to an i*ncrease* in cultural diversity. When the initial condition is one of extreme cultural heterogeneity, intrinsic cultural changes erode the rigidity of cultural boundaries and lead to greater cultural uniformity and fewer local cultures. On the contrary, in a situation in which all local cultures are derived from a single initial culture, intrinsic cultural changes lead to cultural heterogeneity and more cultural regions (see Figure 16).

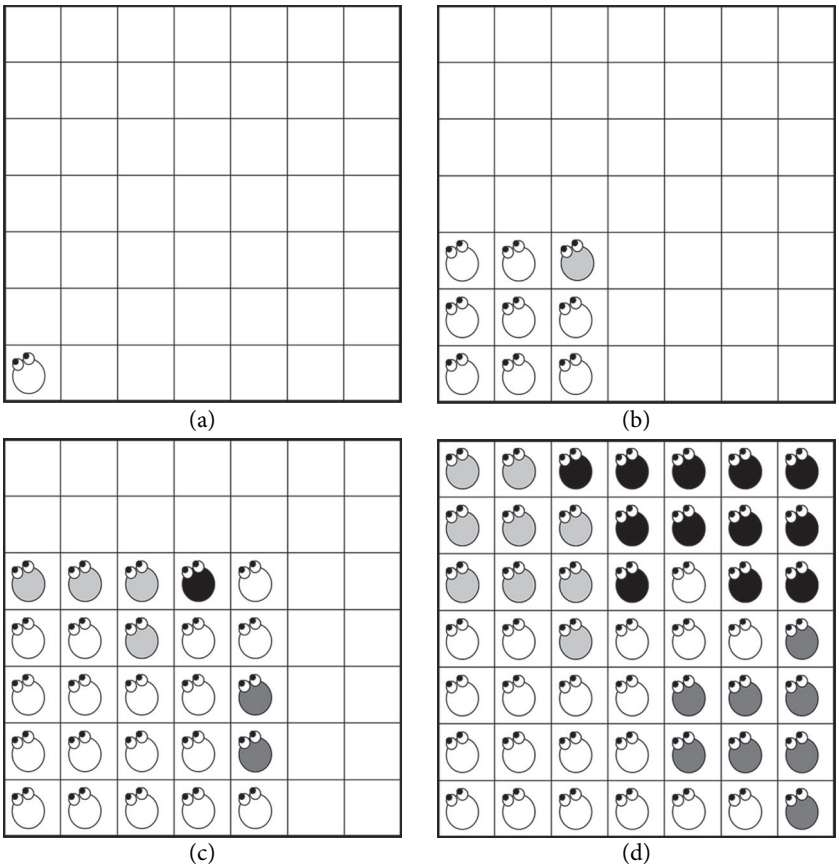


Figure 16. Intrinsic cultural changes lead to the emergence of different cultures which all originate from the same culture.

7. The diminishing role of space in cultural processes

The robots described in the preceding section interact only with the robots that live in neighbouring cells and, therefore, space has a decisive role in their learning from others and in the emergence of cultures. In the past, spatial proximity was a crucial condition for interacting with others and for learning from others. Except for territorial expansions and economic exchanges, only people who lived near to one another interacted together and their behaviours, artefacts, beliefs, and values became similar to the behaviours, artefacts, values, and beliefs of the people which lived near to them. Today, the creation of new technologies for transporting people, goods, and information has made space progressively less important as a precondition for social

interaction. Two individuals can interact together and can influence each other even if they are not near to one another in physical space. In fact, today's technologies make it possible to interact with people living anywhere on the Earth and, therefore, what is important is with whom an individual interacts, not where the two individuals who interact together are located in physical space. What are the consequences for robotic cultures if the robots can interact with, and learn from, robots that can live anywhere?

Our next robots are even more abstract because they are only nodes in an network of nodes. The existence of a link between two nodes – two robots – indicates that the two robots interact together but it does not tell us where the two robots are in physical space and what is the physical distance between them. The network has a topology but not a topography. The topology specifies with which other robots one robot interacts. The topography would specify where the robots are in physical space but for these robots spatial location and spatial distance are irrelevant for social interaction and therefore our networks of robots do not have a topography but only a topology.

An important consequence of networks of robots which interact together independently of physical space has to do with *indirect* cultural influence. For the robots that interact only if they are near to one another in physical space, physical space determines how many intermediate robots a robot must pass through in order to influence and be influenced by another robot. If there are, say, 50 cells between two robots, the two robots can influence each other only if they pass through 50 intermediate robots, which means that they are rather unlikely to influence each other. For the robots for which space is irrelevant in determining if they interact or do not interact, the situation is different. Any robot can interact with any other robot by passing through a limited number of intermediate robots.

We construct a network of robots connected by bidirectional links which represent interactions between pairs of robots. Like our preceding robots, each robot (agent) is a string of bits which specifies what are the cultural traits possessed by the robot. At the outset, the robots have random strings of bits and, therefore, they have all sorts of combinations of traits. In each cycle, each robot interacts with the robots with which it is connected by a link and these interactions can modify the robot's string of bits – the robot's cultural traits. As in our preceding robots, the rule for change is the conformist rule: one, randomly chosen, bit of the string of bits of a robot is changed from 0 to 1 or from 1 to 0 if the majority of the robots which are connected to the robot by a link have a different value for that bit. Otherwise, the bit maintains its current value.

As in the preceding robots which lived in physical space, at the beginning there are no cultures. Each robot has its own random bit of strings which tends to be different from the string of bits of all the other robots and, since culture is the possession of the same string of bits by a group of interacting robots, there are no cultures. But then the conformist rule for cultural change progressively leads to the emergence of groups of interacting robots which have the same string of bits – the same behaviours, the

same artefacts, the same beliefs, and the same values. And here we find an interesting difference with respect to the robots that interact together only if they are near to one another in physical space. The robots which interact only with physically near robots at the end have many *different* cultures. The new robots which interact independently of physical space at the end have a *single* homogeneous culture (see Figure 17).

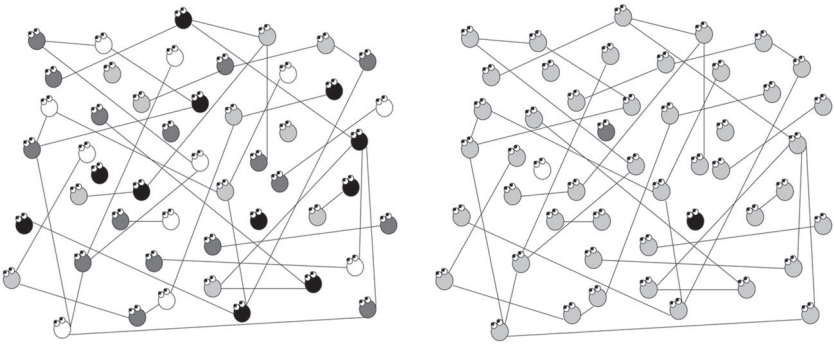


Figure 17. Robots that learn from other robots independently of where they are in physical space (a) at the end have the same culture (b).

Another interesting result concerns the time needed for the emergence of this single homogeneous culture. We construct different networks of robots, with some networks having, on average, fewer links and other networks having more links – which means that in the first type of networks the robots interact with few other robots while in the second type of networks the number of robots with which a robot interacts is larger. What we find is that a single homogeneous culture emerges in all cases but it emerges *more rapidly* in the networks in which the robots interact with *more* other robots.

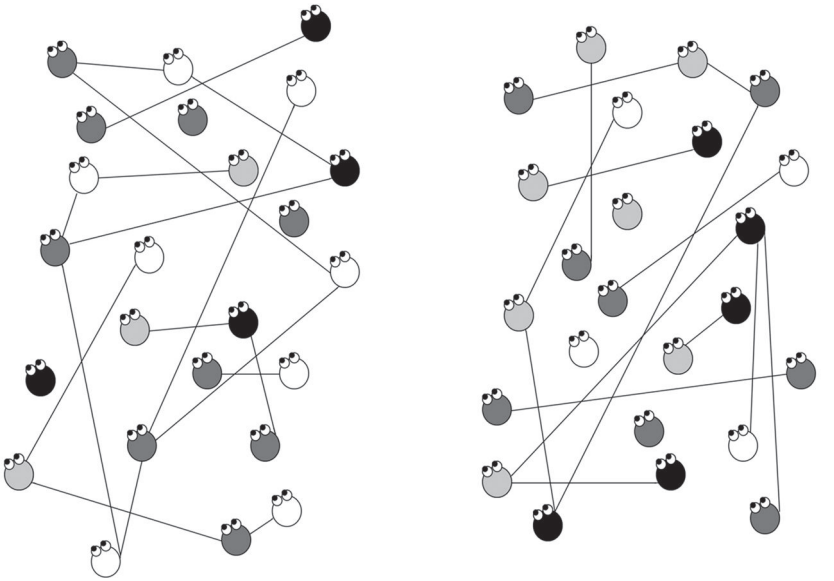
These are very interesting results if we want to understand today's human beings. Today's human beings, compared to the human beings of the past, can communicate and interact with other human beings largely independently of physical space and the number of people with whom a human being can communicate and interact with is constantly increasing. Communication and interaction take place in a variety of different ways. Communication can be direct one-to-one communication, media-based one-to-many communication, and the new Internet-based many-to-many communication. Interaction can be two-way direct interaction with other individuals or it can be the use of the same artefacts used by other individuals because using the same artefacts requires the same behaviours. Advances in the technologies for transporting people, goods, and information lead to a growth of communication and interaction of all types. What our robots tell us is that the growth of communication and interaction leads to the emergence of a single homogeneous culture.

Obviously, the process of cultural homogenization which is taking place today is a gradual process and it is not completed yet. Physical space and physical distance still play a role in communication and interaction among human beings and, therefore, they still act as obstacles for the emergence of a single homogeneous culture. We can reproduce this aspect of the phenomenon with our robots. In the networks described so far we vary the number of links among the robots but in all networks the links are *entirely* independent of physical space. We now construct other networks of robots in which, if two robots are near to one another in physical space, the two robots are more likely to be linked together and, therefore, to interact together, and in some networks physical space plays a more important role than in other networks. What we find is that in the networks in which space plays a more important role in deciding if two robots interact or do not interact together, a single homogeneous culture takes more time to emerge. But a single culture emerges in all cases if space ceases to be the only determinant of social interaction.

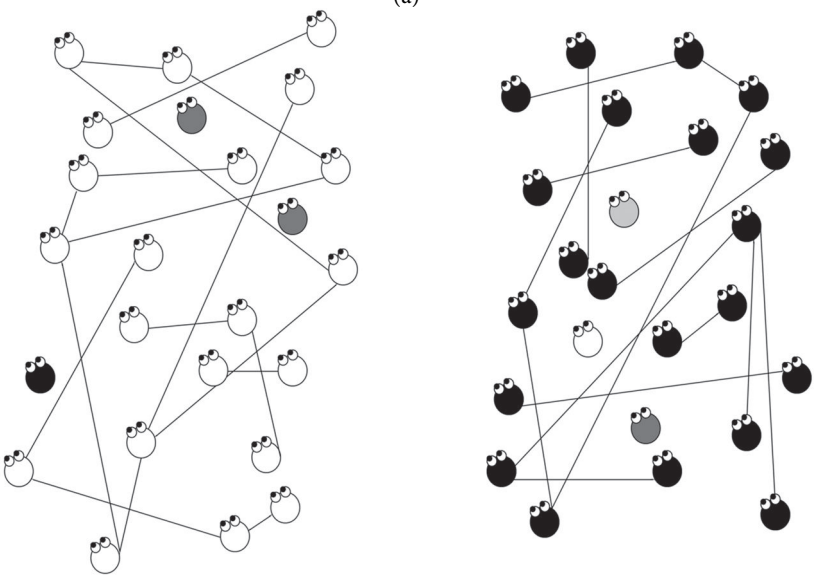
8. Cultural globalization

If globalization is defined as the progressive emergence of a single homogeneous culture for all human beings who live on the Earth, what we are talking about here is globalization. But there is a problem with our networks of robots. The networks of robots start with zero cultures – at the beginning each robot has a randomly assigned string of bits – and, then, a single homogeneous culture emerges for all the robots that make up the network. On the contrary, globalization is taking place today in a world which already contains different cultures. Therefore, to reproduce globalization we must start with a plurality of networks each of which has a culture which is different from the culture of other networks and then we must reproduce the emergence of a single unified culture which includes all the robots.

We construct two separate networks of robots which have internal links but no external links. The robots of one network interact with the other robots of their network but they do not interact with the robots of the other network. We assign a random string of bits to all the robots of both networks and we apply the conformist criterion for learning from others: in each cycle the value of one bit – cultural trait – of a robot is changed if it is different from the value which the bit has in the majority of robots with which the robot interacts. As expected, after a certain number of cycles the robots that belong to the same network have a homogeneous culture and the same is true for the robots that belong to the other network. But, as we also expected, the two cultures are different. All the robots of one network have the same string of bits but this string of bits is different from the string of bits of the robots of the other network (see Figure 18).



(a)



(b)

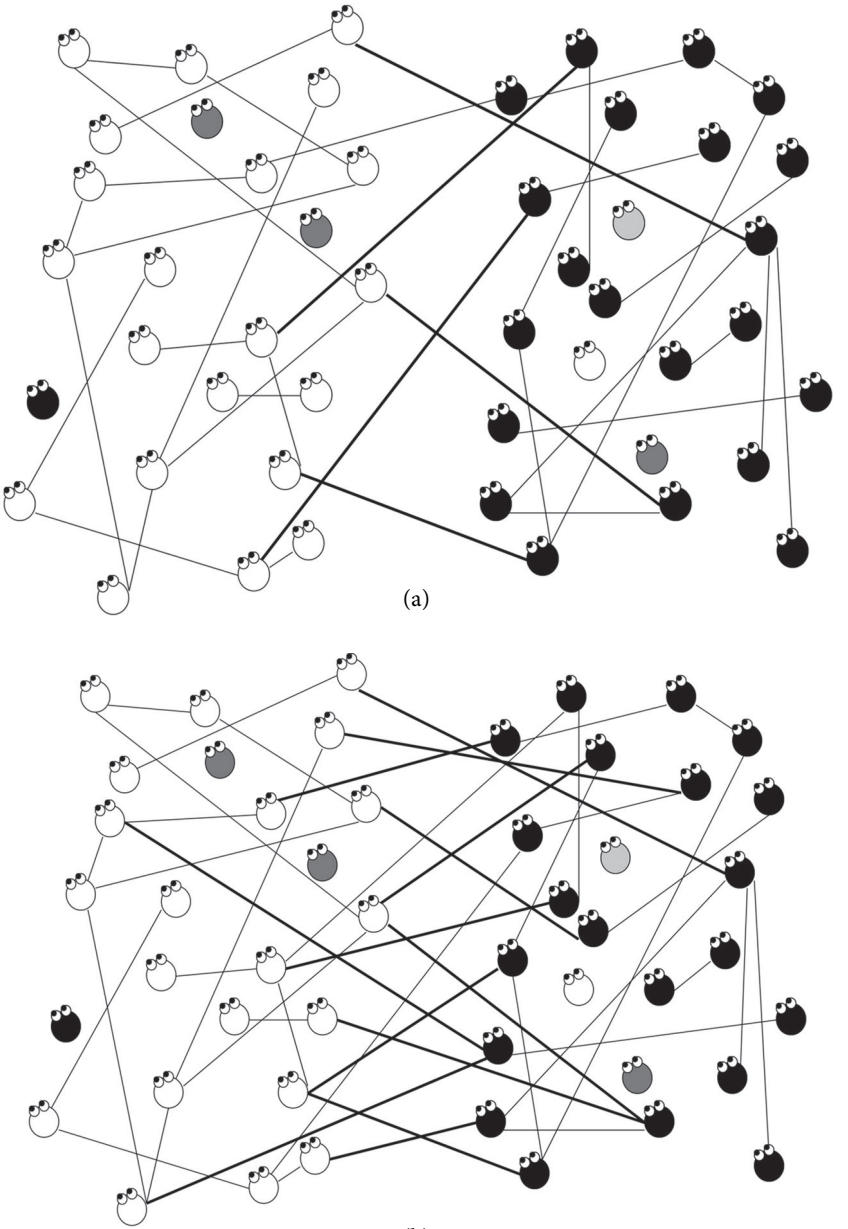
Figure 18. Two communities of robots which do not interact together (a) develop different cultures: a white culture and a black culture (b).

Our next question is: What happens if, after two different cultures have emerged in two separate networks of robots, we allow the robots of one network to interact with the robots of the other network?

The inter-network interactions introduce extraneous cultural traits in the two cultures because the traits of a robot can be influenced by the traits of a robot belonging to the other culture. But if the criterion for learning from others is the conformist criterion, this does not have significant consequences because the two robots' original cultures are strong enough to block the penetration of extraneous cultural traits. However, this is only true if the inter-network links – inter-cultural interactions – are limited in number (see Figure 19a). If the inter-network links increase in number – if the interactions of the robots of one culture with the robots of the other culture become more numerous (see Figure 19b) – we find that all robots end up with a single, unified, culture.

There is a threshold for the appearance of this unified culture. If the number of inter-cultural interactions is below this threshold, the two cultures remain separate and different. But if the number of inter-cultural interactions increases and it goes beyond the threshold, we see the appearance of a single, global, culture. (This is what is happening today.) And there is another factor which plays a role in the emergence of a single culture, and this other factor is the *internal* connectivity – average number of links – inside the two networks. If the two networks are more internally connected – more internally compact – we find that cultural globalization is slowed down. (This may be true today for Chinese culture.)

Other factors can also slow down or, on the contrary, speed up the process of cultural globalization – for example, demographic factors. During infancy human beings interact with fewer individuals compared to adults, and these individuals – parents, siblings – tend to live in spatial proximity to them, while the situation changes when they become older. To reproduce these factors we add something new to our robots. The robots described so far are completely abstract. Not only they do not live in physical space and they do not have a body and a brain, but they do not have a life and a life history. We now construct other robots which have a life history. They are born, they live for a certain number of cycles which is equal for all robots, and then they die and are replaced by other robots which have a random string of bits. A robot's life is divided into two periods, infancy and adulthood, and what changes from infancy to adulthood is with whom the robot interacts and, therefore, from whom the robot learns. During infancy, the child robot interacts with a *small* number of *adult* robots and these adult robots all belong to the *same* network (culture) of the child robot. They are the robot's family. When the child robot becomes an adult, the number of the robot's links is *increased* – the robot interacts with many other robots – and the links can be links to both robots of the same culture and robots of the *other* culture. What we find is that infant robots – and the existence of families – slow down the process of cultural globalization, and cultural globalization requires a greater number of inter-cultural links to be completed.



(a)

(b)

Figure 19. The number of interactions between two communities of robots that have different cultures can be smaller (a) or larger (b).

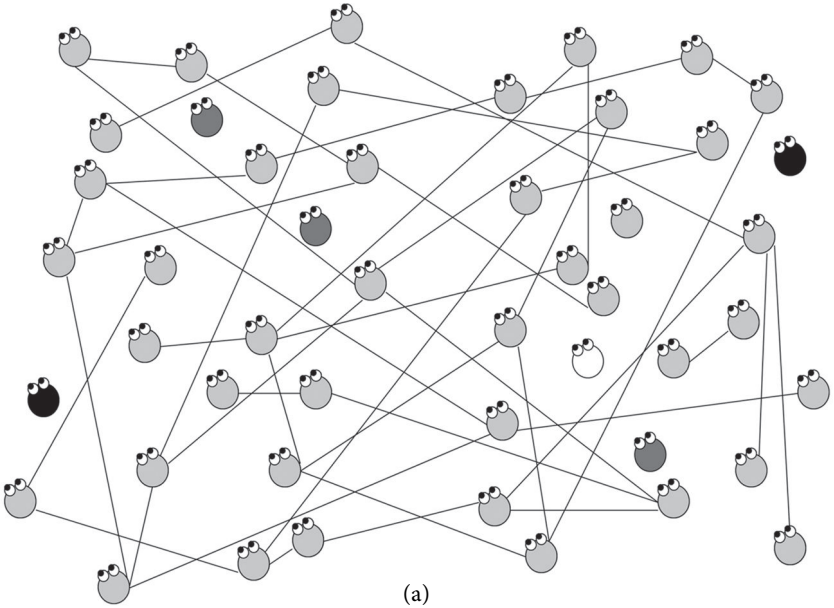
Another period of the life history of human beings which, unlike infancy, has a positive impact on globalization is adolescence. We construct two new networks of robots which have a life divided not into childhood and adulthood but into adolescence and adulthood, and what characterizes adolescence is that during adolescence a robot has links to both other *adolescent* (but not adult) robots of its own culture and to any robot of the *other* culture. This reflects what is happening in today's societies in which adolescents tend to interact with other adolescents rather than with adults – like the adolescent robots described in Section 5 – and they tend to interact with individuals who may live in very distant places by using all kinds of communication technologies,. What we find is that, in contrast to childhood, adolescence speeds up the process of cultural homogenization and a global culture appears even if the inter-cultural links are fewer in number.

But there is another factor which slows down globalization and this other factor is the existence of institutions such as the church or the state which have an influence on the cultural traits of a community of individuals. In our model, these institutions are represented as a single super-node in the network of nodes that make up a community of robots. These super-nodes have *one-way* links to all the robots of the community, which means that the string of bits of the super-node influences the string of bit of all the robots of the community but is *not* influenced by the string of bits of the robots. As expected, the effect of these super-nodes is to slow down the process of cultural globalization because churches and states tend to impose how to behave on believers and on citizens and different churches and states impose different behaviours.

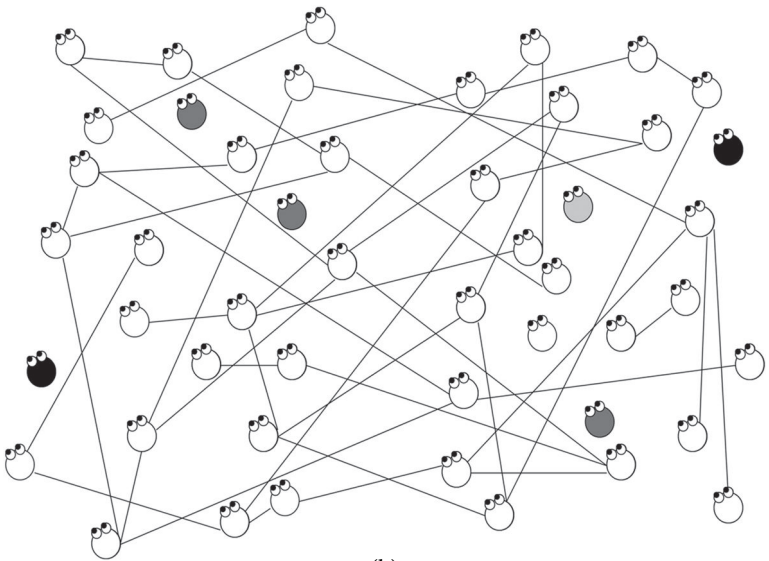
When two cultures interact together to such an extent that these interactions lead to the emergence of a single homogeneous culture, two different things can happen. We can see the emergence of a third culture which is different from the original cultures of both groups or, more probably, one of the two cultures colonizes the other culture and it becomes the culture of both groups (see Figure 20).

If the two networks have the same size and the same internal connectivity, all results are possible because our robots use the conformist criterion for learning from others. This criterion is not based on the *value* of a trait but it is only based on the *number of robots* who have the trait. If the majority of robots with whom a robot interacts have one particular trait, the robot adopts the trait without considering the value of the trait. But there may be other criteria for adopting the traits of the people with whom one interacts. The robots described in Section 3 and in Section 6 use other criteria. They base themselves on the energy (fitness) possessed by the different robots and they take as models to imitate the robots that have more energy or they copy the vases which are constructed/used by the robots that have more energy.

The adoption of this criterion for cultural learning may explain why, today, what we call globalization is, in fact, Westernization – although both Western people and non-Western people may have difficulties in recognizing this fact because Western



(a)



(b)

Figure 20. When the interactions between two communities of robots with different cultures (black and white) increase in number, a single culture tends to emerge which can be a new culture (grey) (a) or the culture of one of the two communities (white) (b).

people do not want to be seen as imposing their culture on other people and non-Western people do not want to be seen as abandoning their culture. Cultural globalization is the emergence of a single, unified, culture for all human beings who live on Earth. But what is this global culture? This global culture is Western culture. Western culture is associated with greater "energy" – more wealth, more freedom, longer life, better medicine, etc. – and this pushes non-Western people to adopt Western culture. (But the globalization of Western culture does mean that Western countries will continue to dominate the world economically and politically.)

Robots which have cultures should help us to confirm or disconfirm both the hypothesis (prediction) that economic and technological globalization will lead to cultural globalization and the hypothesis (prediction) that Western culture will be *the* global culture.

But, until Western culture becomes the single culture of all human beings who live on the Earth, the Earth will be inhabited by people who have different cultures and we must reproduce with robots these different human cultures – Western culture, Chinese culture, Indian culture, African culture, etc. Human cultures are very complex phenomena and, clearly, reproducing with robots the different human cultures is a very difficult task even if we are ready to accept all sorts of simplifications. But reproducing the different human cultures with robots has an important advantage. Scientists – in this case, anthropologists – belong to specific cultures which, until recently, have mostly been Western cultures and, notwithstanding their efforts, they cannot but look at other cultures with Western eyes. If we construct robotic cultures which reproduce the different human cultures, we can look at these cultures – including Western culture – from outside, with neutral eyes, and this may help us to really understand the different human cultures.

Cultural globalization is a very interesting phenomenon. Our most distant human ancestors had one single culture or very few very similar cultures. Then, human cultures became much more numerous and much more different. What may happen in the (near?) future is that all human beings will return to have one single culture – and robots should help us to predict if this single culture will be Western culture (see Figure 21).

What we have said with respect to culture also applies to language – and this not surprising since language is one of the most important components of culture. The thousands of different languages which are spoken today all descend from the vocal or gestural language which was spoken by the common ancestors of all human beings. And in the next section we describe a simulation of the multiplication of Western languages beginning from a single language which existed many millennia ago. Today's globalization can imply that, in the future, all the people living on the Earth may be speaking one single language: English – or a form of English which has been called Globish.

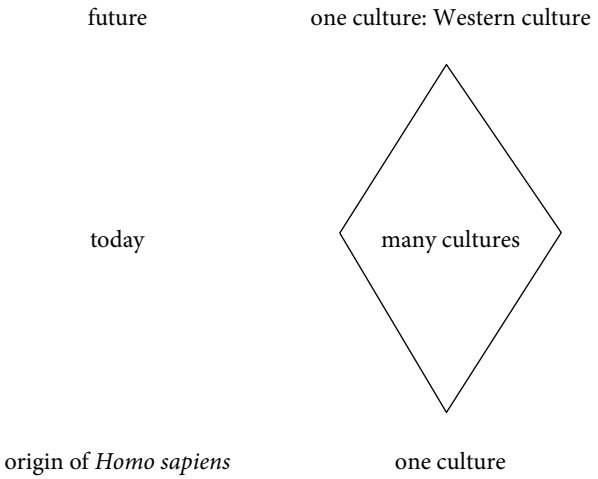


Figure 21. When Homo sapiens came into existence, there was a single human culture and this culture gave birth to the many different cultures which exist today. Globalization may lead back to a single culture, and this single culture will be Western culture.

But the globalization of the English language will take more time than the globalization of Western culture. Human beings learn language in the first years of their life when they interact only with a small number of people who live near to them – mostly their parents and siblings – and who speak their native language. As we have shown in Section 9, this may imply a delay in cultural and, even more, linguistic homogenization. But human beings may invent some technology that causes infants to be exposed to English and to learn English from their birth. And when this will happen, in a few generations English will become the native language of all human beings.

9. The expansion of Indo-European languages

We end this Chapter by describing a simulation of a real historical process: the expansion of Indo-European languages into Europe. (We ignore the expansion of Indo-European languages in the Asian continent.) According to one interpretation of the historical evidence, around ten thousand years ago a group of human beings living in ancient Anatolia (today's Turkey) and possessing one particular culture and one particular language began to expand into Europe and in the space of 3–4 millennia they, or perhaps their language, colonized almost the entire continent. We reproduce this event in the computer but there are no robots in this simulation, not even robots as nodes of a network of nodes. The robots are purely virtual and we only count how many of them live in one particular territory. This is a high price to be paid for our

robotic approach to the study of human beings but it has the advantage of showing us that the approach can be extended to human phenomena that take place in large spaces and last long periods of time and it can reproduce real historical phenomena. And there is another thing which is interesting in these "Indo-European robots". They show us how the nature of the environment can influence – and has influenced in the past – the process of cultural expansion and the manner in which one single culture can give birth to different cultures.

We divide the entire territory of Europe and of ancient Anatolia into square cells and, by using real historical/geographical data, we assign to each cell a "liveableness index" which measures how good is the cell for human living. This index depends of four factors: (a) presence of mountains; (b) presence of rivers; (c) annual rainfall; (d) suitability of terrain for agriculture and animal breeding. The "Indo-Europeans" originally live in one single cell located in Anatolia. The general rule for the expansion of Indo-Europeans is the following. The number of people living in a cell depends on the liveableness index of the neighboring cells. More people live in a cell if the neighboring cells are empty and they have a high level of liveableness. When the number of people living in a cell goes beyond a given threshold, some of the people living in the cell colonize one of the neighboring cell provided that the new cell does not already contain Indo-Europeans (we ignore other people) and its liveableness index is not too low. Cells with high mountains or deserts have a very low liveableness index and therefore they tend not to be peopled by the expanding Indo-Europeans, whereas the Indo-Europeans expand to cells with rivers, sufficient rain, and other factors favoring agriculture and animal husbandry. Some cells are sea cells and these cells are not easy to penetrate but, given certain conditions, they can be traversed by the expanding Indo-Europeans by using ships to reach territories on the other side of the sea.

Time is divided up into years and the process of Indo-European expansion goes on for a certain number of years until it ends because the Indo-Europeans have expanded to most of habitable Europe. One first result is that the time taken by our simulated Indo-Europeans to expand in Europe roughly corresponds to the time taken by the real Indo-Europeans: 3 or 4 thousand years. Figure 22 shows six successive snapshots of the simulated expansion of Indo-Europeans into Europe.

But the more interesting result concerns the directions taken by the expansion process. We have assumed that at the beginning Europe was not inhabited, which of course is not true. This means that the Indo-European expansion was not necessarily an expansion of people but it may have been a cultural expansion, that is, the adoption of Indo-European cultural traits by indigenous people that came into contact with the Indo-Europeans. In other words, the invasion of an empty cell can mean not an invasion of people but an invasion of cultural traits. The cultural trait on which we have the best historical evidence is language and this historical evidence mainly concerns how similar are the languages spoken in Europe today but also those spoken in the past for which we have historical data. Linguists construct a "language tree" with roots

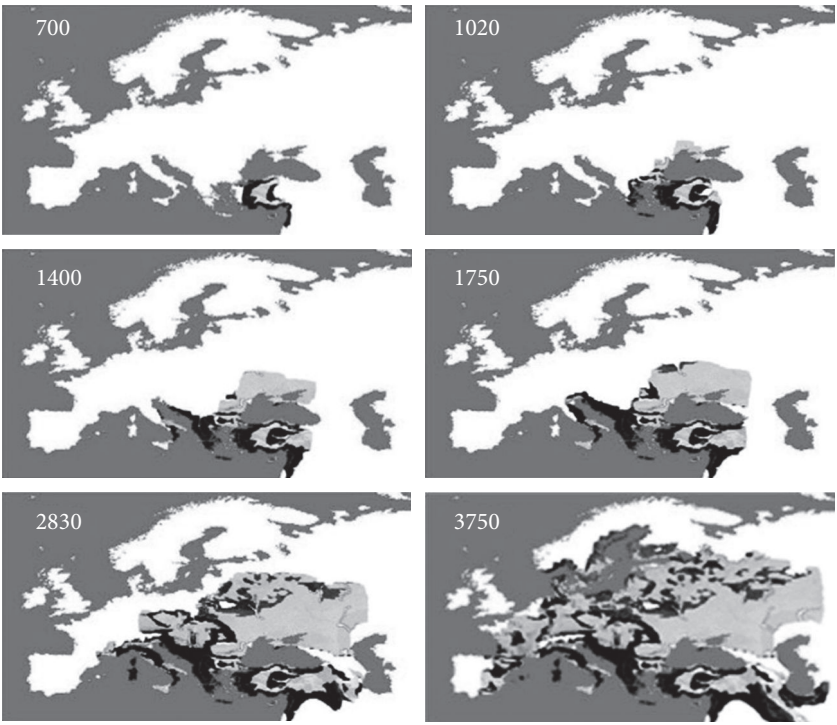


Figure 22. Six successive snapshots of the simulated expansion of Indo-Europeans into Europe. The number in each figure indicates the number of years after the beginning of the expansion in ancient Anatolia (Turkey)

which represent the original language spoken by the people living in Anatolia and each successive ramification representing the moment in time in which two languages separated from a common ancestor language and became different both from the ancestor language and from one another. Our simulation generates an artificial "language tree" which describes when and where different groups of artificial Indo-Europeans separated and we find that the "language tree" of our simulated Indo-Europeans resemble the real "language tree" of Indo-European languages (see Figure 23).

These "Indo-European robots" are interesting because they do not only reproduce an important historical phenomenon and they also show how the nature of the environment in which human beings live can influence what happens to their cultures and languages. (This is also true for some of the robots described in Section 7.) They are also interesting because they show how to reproduce not only cultural convergence and cultural homogenization but also cultural divergence and the creation of new cultures. As we have seen in Section 6, given two separate networks, the two networks will have different cultures but, if we add a sufficient number of inter-network links,

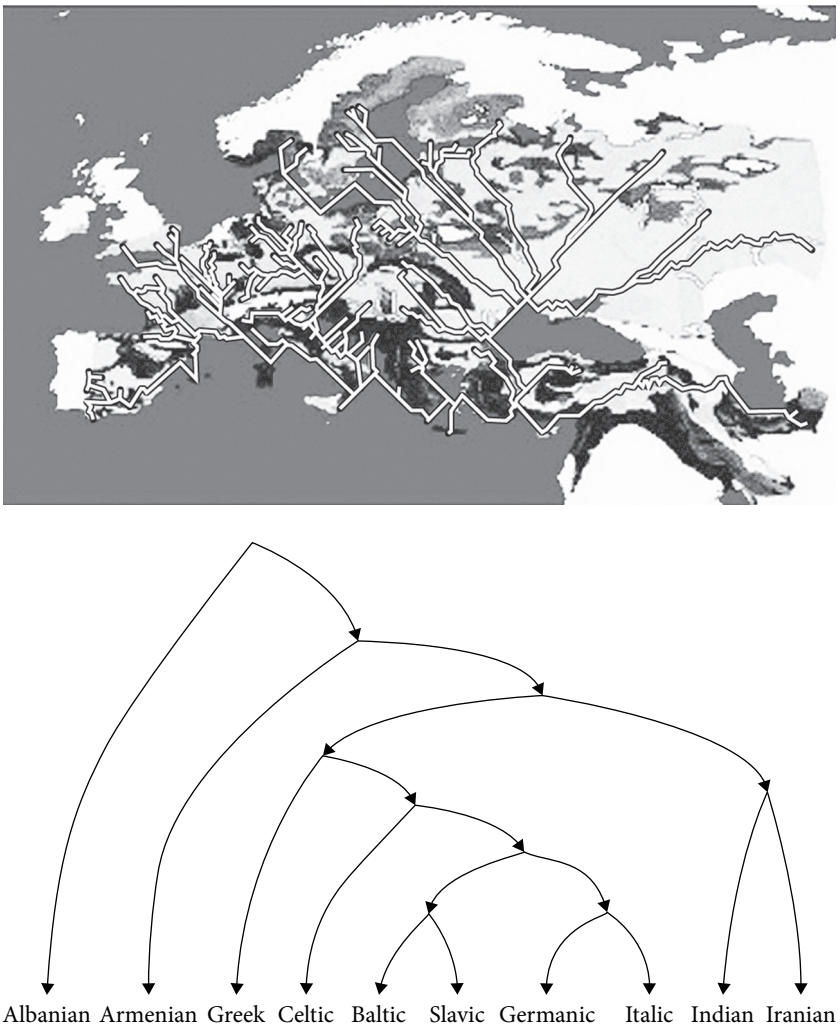


Figure 23. Simulated (a) and actual (b) language tree of Indo-European languages.

a single, unified, culture will appear in both networks. But in the past what has more often happened is that a single culture has given origin to separate cultures, and this can be reproduced by constructing a single network of robots with a single culture and then progressively cutting some of the links in such a way that the single network gives birth to two separate networks with two different cultures.

And now a final comment on robots that have culture. Constructing robots that learn from one another and develop cultures is very important if we want to understand human beings by constructing robots because cultures are important human

phenomena and they determine to a very large extent how human beings behave and how they live. Human beings are the only animal species which constructs the environment in which it lives – both its physical environment and its cultural environment. Other animal species adapt to the environment as it is. Human beings create the environment in which they live and, therefore, they and their environment change together and co-adapt together. Future robots will reproduce how the environment of human beings has progressively changed from a natural environment to the artificial environment of our cities and our technologies. These are phenomena which are very important to understand because, today, human beings are changing their environment at an increasing speed and this may pose special problems to them.

CHAPTER 9

Robot that own things

Another thing that strikes ME about human beings is that they do not immediately use or consume whatever good they are able to procure but they put these goods into external stores to which they only have access so that they can use and consume them whenever they want. External stores allow human beings to survive in difficult times and in difficult environments, to exchange their goods with other human beings, and to have a more long-term perspective on their life. And ME is convinced that external stores explain many aspects of human life and human societies.

The external stores of human beings are not individual stores but they are family stores to which all the members of a family have access. And this, ME thinks, is an important factor that explains why human beings live in families. Family stores make it possible the inheritance of goods. When parents die, the goods contained in their family stores do not disappear but they are transferred to the family stores of their offspring. And family stores also explain why in human societies there are very large differences in wealth, where wealth is the number and value of goods contained in one's external store. Another consequence of having external stores is that external stores tend to be physical structures located in specific locations in the environment, and this explains the emergence of sedentary life and of villages and cities.

Family stores are stores to which many individuals have access but these individuals must have the same genes – typically, parents and their offspring. The reason is that putting one's goods into a common store means to make one's goods accessible to other individuals and, like all animals, human beings generally do not want to give their goods to other individuals unless these other individuals have their same genes.

But ME also finds that large communities of non-genetically related human beings do have a common or central store to which all the members of the community give some of their goods – they pay taxes – and which redistributes these goods to all the members of the community. The central store makes it possible to produce goods which no single individual would be able to produce, to punish behaviours that damage others, and to make wars with other communities to take possession of their goods and to defend one's community from such attempts by other communities.

1. External stores

Some animals hoard the food they find in the environment to eat it later but most animals immediately eat the food they are able to find. Human beings are different because they possess external stores in which they put not only food but all sorts of goods so that they can use these goods whenever they want. The expression "external store" is intended in a completely general and abstract sense and it captures the notion of *owning*. What is in the external store of an individual is something that the individual owns in the sense that no other individual has access to it. And the same generality applies to the notion of "goods". A good is anything which human beings want to have because it increases their survival and reproductive chances and their well-being. (For goods, see Chapter 11 on robotic economies.) External stores are an important human adaptation and they are at the origin of many of the differences that exist between nonhuman animals and human beings. External stores mitigate the consequences of temporary scarcity because, when the environment contains little food, human beings can eat the food contained in their stores. Movable external stores make it possible to travel, migrate, and re-settle elsewhere by bringing one's food with oneself. External stores as physical structures located in specific places are at the origin of sedentary life and the development of houses, villages, and cities. Storing goods allows an individual to specialize in the production of one good and to obtain other goods from other individuals through the exchange of goods. External stores make it possible to inherit goods and not only genes and cultures, to produce new goods by using the goods that one already has, to adopt a more extended temporal perspective which includes predicting and planning for the future, and to have the physical and mental freedom to dedicate oneself to such speculative enterprises as art, religion, philosophy, and science. In fact, man could be called "the animal that stores/owns" as much as he is called "the animal that knows (homo sapiens)" or "the animal that speaks (homo loquens)" or "the animal that constructs artefacts (homo faber)".

This is why constructing robots that have external stores is important if robots must help us to understand human beings. In this Chapter we describe robots that have external stores in which they put the food which they find in the environment. The robots are our usual robots that live in an environment containing food tokens that they must eat to remain alive. But, unlike our other robots, when a robot reaches a food token, it does not immediately eat the food token but it puts the food token in an external store. We describe different robots that have three different types of external stores: individual stores, family stores, and central stores. Individual stores are stores in which only one robot puts the food that it is able to find in the environment and only that particular robot can eat the food contained in the store. Family stores are stores in which a group of kin-related robots – a family of robots – put their food and all the members of the family eat the food contained in the family store. Central stores are stores in which all the members of a community of robots put (some of) the food

tokens they find in the environment and the food tokens contained in the central store are redistributed to all the members of the community.

Human beings have family stores, not individual stores, because human beings live in families, and we construct robots that have individual stores only to better understand the adaptive value of external stores. (But in today's societies an increasing number of individuals live alone, and these individuals can be said to have individual stores.) Central stores are (political) states. The move from family stores to central stores is a crucial event in the history of human societies – what anthropologists call the transition from family-level societies to state-level societies – but central stores do not replace family stores and modern human societies have both family stores and a single central store.

2. Individual stores

To understand the consequences and the advantages of having external stores we compare robots that do not have external stores with robots that have external stores and we determine what the second type of robots can do that the first type cannot do. We begin with individual stores, which are stores in which a single robot puts the food tokens that it is able to find in the environment and only that particular robot can eat the food tokens contained in the store. As we have said, individual stores are not very frequent among human beings because human beings live in families and they have family stores rather than individual stores. But we begin with individual stores because individual stores are simpler than family stores and they can better illustrate the advantages of having external stores.

The robots that do not have external stores are the robots that we already know. They live in an environment containing food tokens and each food token provides a certain quantity of energy which is added to the energy contained in the robot's body when the robot eats the food token. The robots' behaviour is controlled by a neural network with visual neurons encoding the location of the nearest food tokens, one touch neuron which is activated when the robot reaches a food token, and one hunger neuron whose level of activation reflects the current level of energy in the robot's body. When a robot's body touches a food token, the food is considered as eaten.

We now construct another population of robots which are different from the preceding robots because they have an external store in which they put the food they find in the environment. When the robots without an external store touch a food token, they automatically eat the food token. The new robots do not eat the food token but they put the food token in their external store (see Figure 1).

The neural network of these robots has an additional motor neuron which encodes the behaviour of eating one of the food tokens contained in the robot's external store – provided that the external store is not empty (see Figure 2).

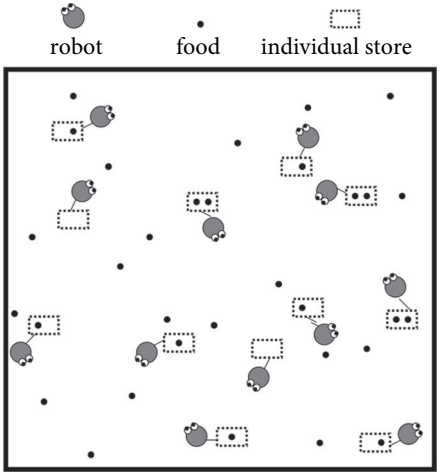


Figure 1. Robots with individual stores which may contain different numbers of food tokens because the robots are more or less able to collect the food tokens contained in the environment.

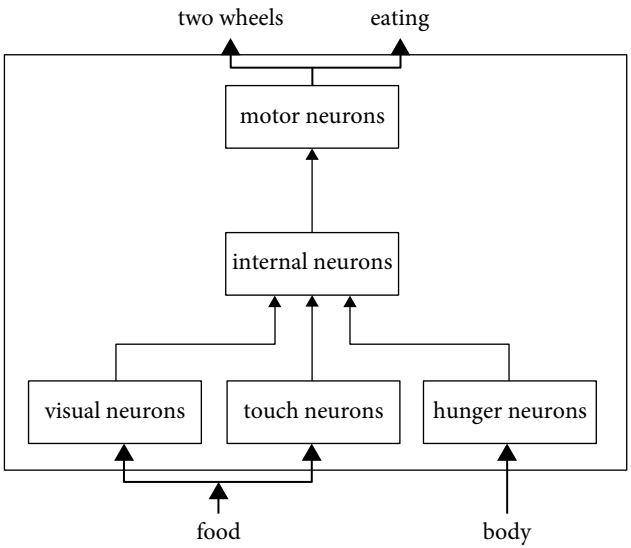


Figure 2. The neural network of robots which have individual stores. When a robot finds a food token, it automatically puts the food token in its external store and eats the food token only when its body signals hunger.

The two populations of robots live in identical environments. The energy in the robots' body can go from 1 (maximum energy) to 0 (no energy) and, when it reaches the 0 level, the robot dies. At birth the energy is at its maximum level but it is reduced by a fixed quantity in each cycle. All food tokens contain the same quantity of energy but if, when a robot eats a food token, the bodily energy of the robot is near to its maximum level, some of the energy contained in the food token is wasted.

How the robots which possess an external store compare with the robots which do not possess an external store in terms of the length of their life? The answer is in Figure 3. The robots which have an external store live longer – have more fitness – than the robots which do not have an external store.

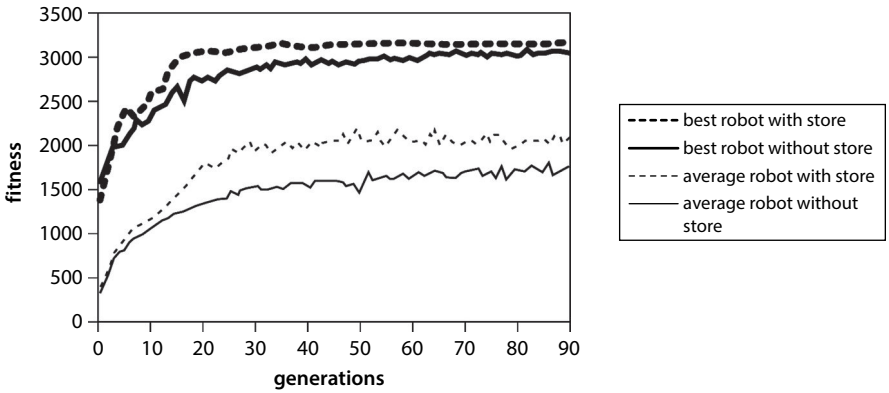


Figure 3. Fitness (length of life) of robots with and without individual external stores.

Why possessing an external store makes the robots live longer? The answer is that the store functions as a safety net. The robots can find themselves in one part of the environment with little food. If the robots do not have an external store, the energy in their body may go to zero and the robots die. In contrast, a robot which has an external store can eat the food contained in its external store and in this way it may be able to reach other parts of the environment with more food. The external store makes the robot less dependent on the environment.

There is also another advantage of possessing an external store. As we have said, if a robot eats a food token when the level of energy in its body is near its maximum level, one part of the energy contained in the food token may be lost. A robot knows how much energy is contained in its body because it has the hunger neurons, and this information can influence the robot's behaviour. And, in fact, we find a difference in how the two types of robots respond to hunger. When a robot without the external store sees a food token and the robot is hungry, the robot approaches and eats the food token. However, when the robot is not hungry, the robot tends *not* to approach the food token because eating the food token would imply a waste of at least some of

the energy contained in the food token. (Overeating could also be dangerous but we do not reproduce the negative effects of overeating in these robots.) If we look at these robots on the computer screen, we see that, when they are not hungry, they almost stop moving or they move in circles or they move around the food tokens without touching them. They only approach and eat the food tokens when they are hungry so that the energy contained in the food tokens is not wasted.

The behaviour of the robots that have an external store is different. These robots always approach the food tokens as soon as they see them. For these robots reaching a food token does not mean to eat the food token but to put the food token in their external store, and they can do this whether they are hungry or not hungry with no fear of wasting the energy contained in the food token. It is only when they are hungry that the robots decide to eat the food tokens contained in their external store and, in fact, the robots' "eating" motor neurons tend to be activated only when they are hungry. The robots with the external store are always active and they explore the environment independently of their hunger. The behaviour of the robots which do not have an external store depends on the state of their body (level of hunger) and this causes them to be overall less active and to visit fewer zones of the environment. The robots which have an external store are less dependent on the state of their body. They are always very active and they are better than the robots without an external store at collecting the energy contained in the environment – the total energy contained in the food tokens. They can always have this energy at their disposal and they do not waste it. (If we were to add a cost of movement, this would be a cost especially for the robots with external stores but the advantages of storing goods would be greater than these costs.)

To better see the differences in the behaviour of the two types of robots, we conduct two laboratory experiments. In the first experiment we put the robots in an empty environment and we measure their exploratory behaviour by counting the number of different parts of the environment (different pixels on the computer screen) which the robots visit during a fixed number of cycles – where visiting more pixels means to be more active and more explorative. The results are that the robots with the external store are more active and more explorative than the robots without the external store. In another experiment we determine how the robots respond to the sight of a food token. The environment contains one single food token and we count the number of cycles taken by a robot to reach the food token. In this second experiment we vary the level of hunger of the robot and we test each robot when it is hungry and when it is not hungry. What are the results? The robots that do not have the external store are very fast at reaching the food token when they are hungry but they hesitate and take much more time when their hunger is low. In contrast, the robots with the external store are always very fast at reaching the food token whatever their level of hunger. This clearly demonstrates that the robots with the external store are more autonomous from the state of their body and they are more effective in collecting the energy contained in the environment. And these are typically human characteristics.

Since our robots are neuro-robots, we can determine if the two types of robots are different not only in their behaviour but also from the point of view of what happens in their brain. As we have said, in addition to the visual neurons that allow the robots to see the food tokens, the robots' neural network has one touch neuron and one hunger neuron. The touch neuron is activated when the robot reaches (touches) a food token. The activation level of the hunger neuron reflects the quantity of energy currently contained in the robot's body. It is 1 when the body is full of energy (no hunger) and its activation progressively decreases with the consumption of energy (increasing hunger). We measure the level of activation of these two neurons in the robots with the external store and in the robots without the external store, and we take these measures not in the laboratory but when the robots are in their natural environment.

Figure 4 shows the results for two robots, one with the external store and the other one without the external store. The robot which does not have the external store reaches the food tokens only when its bodily energy is near zero and the robot is very hungry. The robot with the external store reaches the food tokens – its touch neuron is activated – *independently* of its level of hunger. As a result, the robot with the external store collects more food tokens (8) than the robot without the external store (2). For the robot with the external store, Figure 4 also shows the level of activation of the motor neuron encoding the act of eating one of the tokens contained in the robot's store. This neuron is activated only when the hunger neuron signals hunger. This means that the robot eats the food tokens contained in its store only when it is hungry.

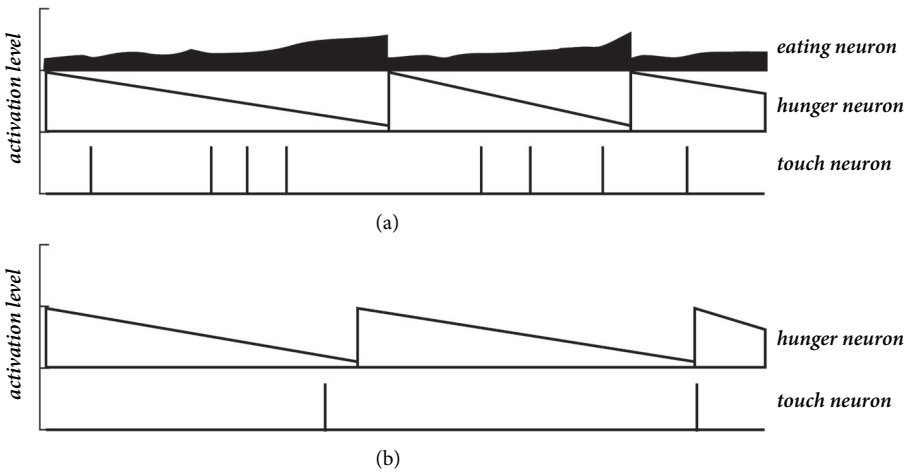


Figure 4. Activation level of the hunger and touch neurons of the neural network of a robot with an external store (a) and of a robot without an external store (b). For the robot with the external store we also show the activation level of the eating neuron.

Possessing an external store makes the robots live longer and this is why external stores are useful and why we expect that, given the appropriate pre-conditions, a population of robots which initially do not have external stores in a succession of generations will become a population of robots with external stores. (The development of external stores in human beings is due to cultural, not biological, evolution, although it may have biological pre-conditions.) However, the adaptive value of any behaviour – in this case, the behaviour of storing food – is not absolute but relative to the particular environment in which the animal lives. So we ask: In which environment are external stores adaptive? Our prediction is that the possession of external stores is adaptive in environments which have an *intermediate* level of difficulty while it is not adaptive in environments which are either *too easy* or *too difficult* to live in. Given the number of food tokens that it contains, the environment of our robots has an intermediate level of difficult and, as we have seen, in this environment possessing external stores increases the length of the robots' life. But if the environment in which the robots live contains more food tokens, external stores may not evolve because there is no need for them. The robots find food whenever they need it. On the contrary, in environments with very little food, external stores may not evolve because a robot must immediately eat all the food which it is able to find and cannot store its food.

To test this prediction we evolve the robots in environments which contain the same number of food tokens as the environment of the preceding robots but have *different* sizes. The environment of our preceding robots was a square of 1000x1000 pixels. We now evolve robots with and without the external store in an environment which is smaller (500x500 pixels) or larger (5000x5000 pixels) than the original environment. Since both environments contain the same number of food tokens as the original environment, in the smaller environment food is more dense and in the larger environment less dense than in the original environment. Therefore, the smaller environment is easier to live in than the original environment while the larger environment is the most difficult one.

The results are shown in Figure 5 and they are in agreement with our prediction. External stores are useful in environments of intermediate difficulty but they are not useful in environments which are very easy or very difficult to live in. In both the very easy and the very difficult environments the length of the life of the robots with an external store and of the robots without an external store is more or less the same and, therefore, in these environments having an external store does not make any difference.

This may tell us something interesting about the environments in which human beings have developed external stores. These environments were environments with a temperate climate, while there was a smaller pressure to develop external stores in very warm environments with much food and in very cold environments with very little food.

There may be other advantages of having external stores. The robots described so far live in an environment in which the food tokens can be found in all parts of the

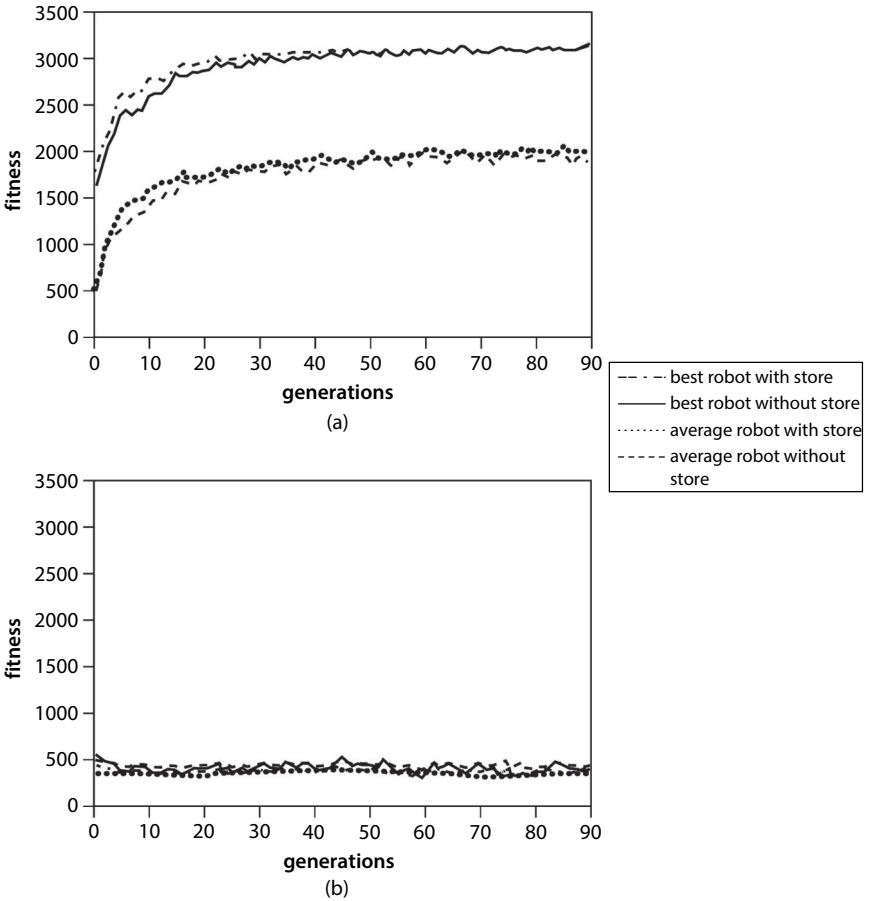


Figure 5. Fitness of robots with and without the external store in an environment which is easier (a) or more difficult (b) than the environment of Figure 1. In the difficult environment, the robots' life remains very short during the entire process of evolution.

environment. But in the real environment food may not be not equally distributed in all parts of the environment but it can be found in some parts and not in others. Our next robots show that having external stores can be especially useful in these more complex environments.

The new robots live in an environment with two zones containing food separated by a large space in which there is no food, and to move from one food zone to the other food zone a robot must traverse the large empty space between the two zones. As before, when a robot reaches a food token and, according to the type of robot, either immediately eats the food token or puts the food token in its external store, the food token disappears and it is not replaced by a new food token. Therefore, the number

of food tokens in a zone is progressively reduced until the zone becomes completely empty and, therefore, if it wants to eat, the robot must go to the other zone. Clearly, this environment is more difficult than the preceding environment in which food could be found in all parts of the environment.

We evolve the same two populations of robots, one with external stores and the other one without external stores, in the new environment and the results show that possessing an external store is even more advantageous in this environment than in the environment in which food can be found everywhere. In the environment with the two food zones the difference between the length of life of the robots with an external store and of the robots without an external store is greater than it was in the preceding environment (see Figure 6).

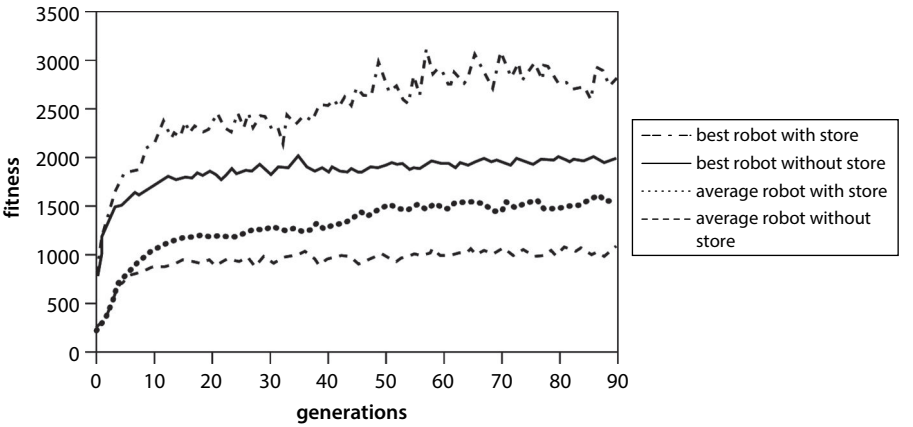


Figure 6. Fitness of robots with and without the external store living in an environment in which food is in two zones separated by an empty space.

If we count the number of robots which eat at least one of the tokens contained in the food zone in which they were *not* born, we find that almost no robot without an external store is able to reach the other food zone and eat the food tokens contained in the other food zone. In contrast, for the robots which store their food the number of robots which are able to reach the other food zone progressively increases across the successive generations until at the end more than 20% of the robots are able to traverse the empty space between the two food zones and reach the other zone.

The robots which do not have an external store are unable to reach the other zone because to reach the other zone they must traverse a large space without food. Without an external store containing food that they can eat while they are crossing the empty space, they die before reaching the other food zone. This influences the capacity of the robots to explore the environment. As we have done with other robots, we measure the capacity to explore the environment by counting the number of pixels visited by a

robot in an empty environment and we find that the robots with the external store are more explorative than the robots without the external store. This is like what we have found with our preceding robots but now the difference is much greater and it may not have the same cause. Inside a food zone there are many food tokens and therefore, until a robot remains inside a food zone, there is no great pressure to develop a capacity to explore the environment. But, as we have seen, the robots without the external store are unable to go outside the food zone in which they were born and, therefore, the behaviour of exploring an empty environment does not have the possibility to evolve in these robots. On the contrary, the possession of an external store makes it possible for the other robots to traverse the empty space between the two food zones and therefore these robots evolve a capacity to explore the empty space. This also may have happened to human beings in the distant past.

So, possessing an external store is not equally useful in all environments. But the usefulness of external stores depends not only on the intrinsic characteristics of the environment but also on the adaptive pattern of the robots. Imagine a population of robots living in an environment with food tokens and water tokens and, to remain alive, the robots must both eat the food tokens and drink the water tokens. The robots have two internal sensors, one telling them what is the current level of energy in their body and the other one what is the current level of water, and the robots' brain uses this information to guide a robot towards the tokens which are more needed by its body in that moment. We evolve the robots in two different environments. In one environment the food and water tokens are distributed in the entire environment. In the other environment the food tokens are in one zone of the environment and the water tokens are in another zone, and the two zones are separated by an empty space.

The results show that possessing external stores, one for food and the other one for water, is even more advantageous for the robots (see Figure 7). If the two types of tokens are distributed in the entire environment, a robot which does not have an external store and needs to eat rather than to drink looks for the food tokens and ignores the water tokens, and the opposite for a robot which needs to drink and not to eat. (These robots are like the robots described in Chapter 2.) But a robot which has an external store for food and another external store for water can put in its stores both the food tokens and the water tokens and eat or drink them according to its needs – which is a more efficient behavioural strategy.

The advantages of having external stores are even greater for the robots that live in the other environment, the environment in which the food tokens are in one zone and the water tokens are in another zone. In this environment, a robot which does not have an external store is in serious trouble when it is in the zone of the environment which contains the food tokens but its body needs water, or vice versa. To get a water token, the robot must cross the empty space between the two zones and it risks dying before it reaches the water zone. On the contrary, a robot with the two external stores can survive by drinking the water tokens contained in its store while crossing the empty space.

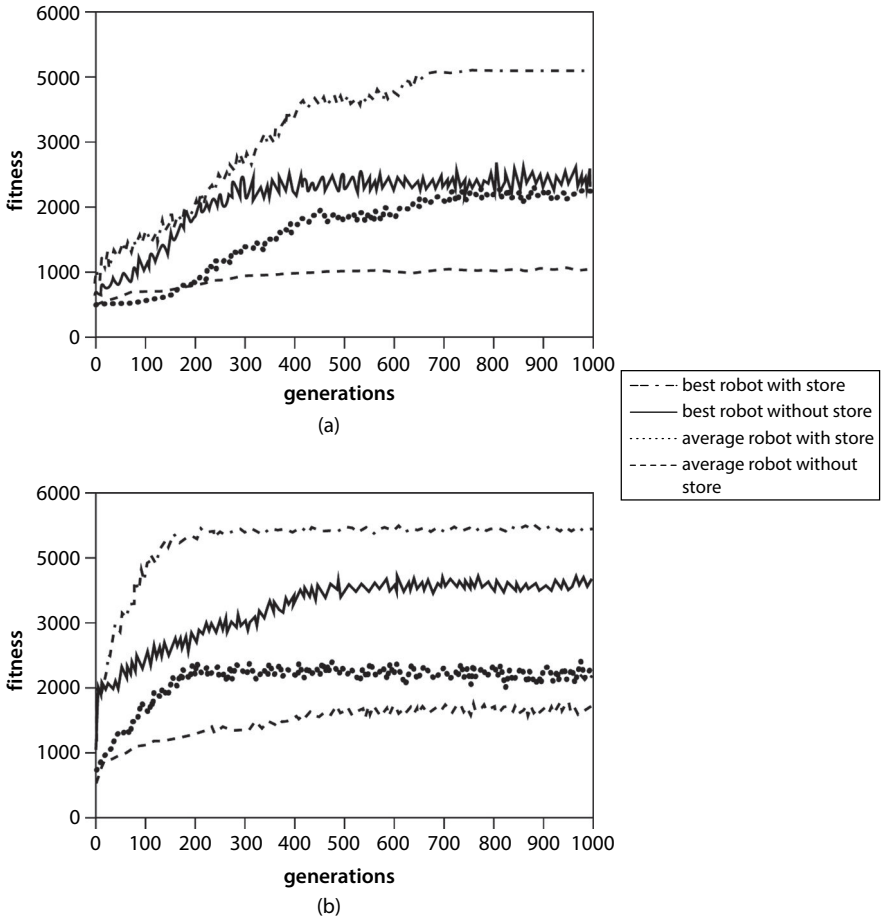


Figure 7. Fitness of robots with two external stores, one for food and one for water, and for robots with no external stores. The food and water tokens are randomly distributed in the entire environment (a) or they are in two separate zones of the environment (b).

So, possessing an external store is especially useful if the food tokens and the water tokens are in different zones of the environment separated by an empty space. Clearly, how large is the empty space between the two zones is a critical factor and the possession of external stores should become progressively more advantageous as the distance between the two zones increases. This is what we find if we let the robots evolve in an environment in which the food zone and the water zone are more distant than in the preceding environment. In this environment the advantages of having an external store for food and an external store for water are even greater and the life of the *average* robot which has the two external stores has the same length of the life of the *best* robot which does not have external stores.

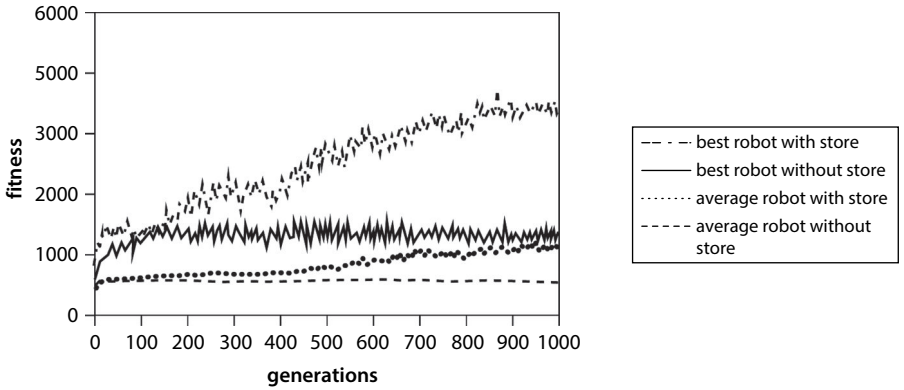


Figure 8. Same robots of Figure 7b but living in an environment in which the distance between the food zone and the water zone is greater.

The external stores of the robots described so far are individual stores: each individual robot has its own store and only that robot can eat the food contained in its store. But human beings do not have individual stores. They have family stores. The next section is dedicated to robots that have family stores.

3. Family stores

Family stores are stores in which all the members of a family of robots put the food they are able to find in the environment and all the members of the family eat the food existing in their family store (see Figure 9). A family is made of two or more robots which have similar genes. Hence, it can include parents and offspring, siblings, grandparents, grandchildren, cousins, uncles, etc.

But there is an important difference between the robots with individual stores and the robots with family stores. While the robots with individual stores automatically put the food they find in the environment in their external store, the robots with family stores can decide whether to put the food they find in the environment in their family store or to immediately eat the food. The robots' neural network has an additional output neuron – a "food storing" neuron – which encodes the behaviour of storing food in the family store, and only if this neuron has an activation of level higher than 0.5, a robot puts the food tokens it finds in the environment in its family store. If the "food storing" neuron has an activation level which is less than 0.5, the robot immediately the food token (see Figure 10).

Since the food which a robot puts in its family store can be eaten by another member of the robot's family, we let the robots decide to behave or not to behave in this altruistic way.

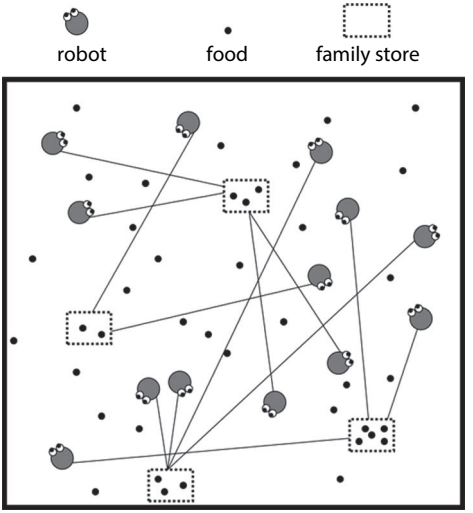


Figure 9. Robots with family stores. The robots put the food they find in the environment in their family store and, when they are hungry, they eat the food contained in their family store.

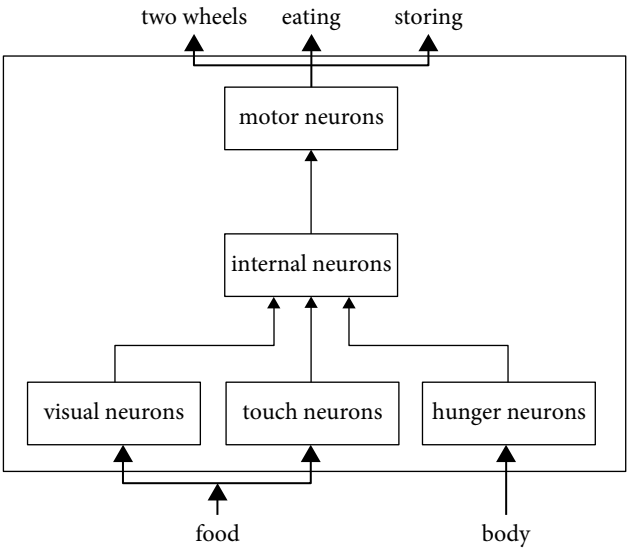


Figure 10. Neural network of robots with family stores. These robots autonomously decide to store or not to store in their family store the food they find in the environment.

The new robots are different from the preceding robots in other respects. All our robots, until they remain alive, generate one offspring at regular intervals. However, while the offspring of the robots with individual stores are virtual robots and they become real robots only when all the robots of the preceding generation have died, for the robots with family stores, when an offspring robot is born, the new robot is immediately put in the environment together with the other robots, and this means that the offspring robots live with their parents, siblings, and other relatives. This has an important consequence. While for the robots with individual stores population size is fixed and the robots that live longer generate a greater percentage of the robots of the next generation, for the robots with family stores population size can vary during the course of evolution. We start with a certain number of robots whose age is decided randomly and, therefore, some robots are younger and some older. Then the number of robots increases or decreases depending on when the robots die and on the number of offspring that they generate. This changes how we measure fitness. For the preceding robots fitness is a property of the individual robot and it is the percentage of the robots of the next generation that are its offspring. With the new robots fitness is the size of the population of robots. If the robots are more numerous, this means that the robots behave better and are better adapted to the environment. As always, the neural networks of the robots of the initial population have random connection weights and therefore the robots are not very good at finding food. They have short lives and generate few offspring and, as a consequence, the size of the population is drastically reduced. Then, gradually, the selective reproduction of the best robots and the addition of random mutations to the robots' inherited genes lead to robots which are better at finding food and, therefore, live longer and have more offspring. This causes an increase in the number of robots until this number reaches a stable value which reflects the "carrying capacity" of the environment – the maximum number of robots that can live in the environment given the total quantity of food present in the environment.

These robots have family stores, not individual stores. Family stores function on the basis of a simple rule: parents share their store with their offspring. One robot and its offspring put the food that they find in the environment in the same external store – the family store – and they eat the food contained in the family store. We evolve different populations of robots with different types of family stores. We begin with robots which have the same family store for the entire length of their life. Offspring robots never leave the family store of their parents. They do not create their own independent family store but they continue to use their parent's store even when they become adult and have their own offspring. And their offspring also use the family store which is the family store of their grandparent, or was the family store of their grandparent because their grandparent is dead. What happens with these robots?

Each robot of the initial population has its own family store but then most family stores disappear because most robots are unable to find food and they die without leaving offspring. This means that most genetic lineages become extinct and, after a certain number of generations, all living robots descend from one or a few ancestors

which were members of the very first generation of robots. In terms of family stores, this means that at the end there are only one or very few family stores and these family stores are very large because they are used both by robots which are closely related genetically – parents and offspring – and by robots which are more distantly related genetically – uncles, cousins, second cousins, etc.

The robots live in a seasonal environment. When a robot reaches a food token, the robot can decide to either eat the food token or put the food token into its family store but in both cases the food token disappears. The number of food tokens is progressively reduced and new food tokens only appear when the new season begins. In this environment external stores are very useful because at the end of a season there is little food in the environment but the robots can eat the food tokens contained in their external store. But the robots with family stores are different from the robots with individual stores. While the robots with individual stores automatically put the food they find in the environment into their individual store, the new robots are free to put the food into their family store or to immediately eat the food. What do these robots do? Do they put the food into their family store – which implies that the food can be eaten by other robots – or do they immediately eat the food?

The behaviour of the robots changes during the course of evolution. Initially, when there are many small family stores, the robots tend not to immediately eat the food they find in the environment and to put the food into their family store. But when family stores become fewer in number and larger in size, the robots prefer to immediately eat the food they find in the environment. And, since the robots cease to put their food in their family store, family stores disappear.

Why do the robots behave in this way? As we have said, after a certain number of generations, all the robots tend to be genetically related because they all descend from one or a few robots of the initial generation. And, as we have also said, this implies that there are only one or a few very large family stores in the entire population. However, the robots that share these few family stores are very distantly related genetically and, therefore, if a robot puts its food in its family store, the robot behaves altruistically towards robots whose genes are rather different from its genes. According to the theory of kin selection, altruistic behaviour which reduces one's reproductive chances and increases the reproductive chances of someone else will only be exhibited towards *closely* genetically related individuals. The behaviour of these robots is in agreement with kin selection theory. When the family stores become very large and they include many robots which are only *distantly* genetically related, a robot ceases to put its food into its family store and it prefers to immediately eat the food it finds in the environment.

Figure 11 shows the percentage of robots which, when they find a food token, do not eat the food token but put the food token into their family store. The percentage at first increases and, after 50 generations, 85% of the robots store their food in their family stores which are still small and numerous. But then, when the family stores have become very large and very few in number, the percentage goes rapidly down and, at

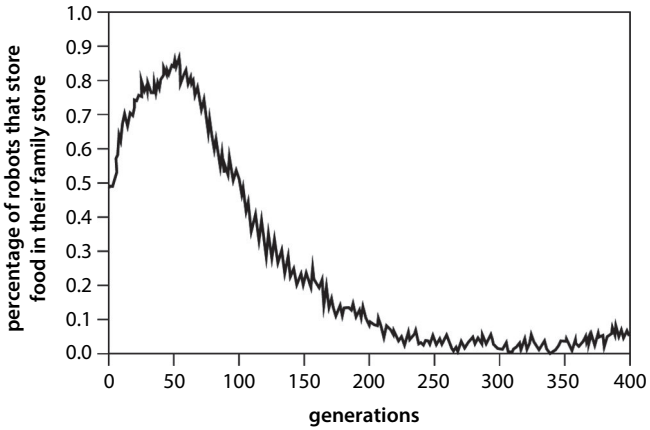


Figure 11. Percentage of robots that store their food in their family store across a period of 400 generations.

the end, less than 10% of the robots put their food tokens into their family store while the remaining 90% of robots prefer to immediately eat the food tokens.

These robots tell us is that, if the offspring robots use the family store of their parents for the entire length of their life and they do not create their own family store when they become adults and have their own offspring, after a while there is only one family store for the entire community of robots. An external store which is shared by all the members of a community of robots can be called a community store. But the members of the community, even if they descend from the same ancestors, are very distantly related genetically and, therefore, kin selection theory predicts that community stores should not exist. And this is what we find. When the family store has become a community store, the robots cease to put their food in the community store and the community store disappears. (Human beings do have large community stores which are called "states", and this raises the question of how human beings can have states. We will return to this question later in this chapter and in next chapter.)

But this poses a problem because, unlike our robots, human beings do have family stores. So, how should we change our robots so that they have family stores? We construct another population of robots in which an offspring robot does not use the family store of its parent for the entire length of its life but, when it becomes an adult (at the age of 1000 cycles, which is half the maximum length of its life: 2000 cycles), the robot leaves the family store of its parent and it creates its own family store. Now the situation changes radically. The family stores remain small and, until the end, there are a considerable number of them. At the end the percentage of robots that put their food in their family stores rather than eating the food immediately is almost 90% (see Figure 12) and family stores become a permanent feature of the population of robots.

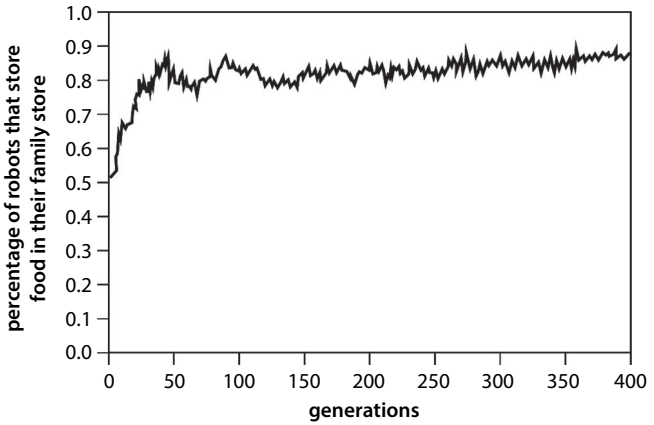


Figure 12. Percentage of robots that store their food in their family store which is small because it is shared only by closely genetically robots.

This result also is explained by the theory of kin selection. The theory of kin selection predicts that an individual will be less and less disposed to behave altruistically towards individuals that are increasingly distant genetically from them but it will behave altruistically towards closely related individuals. The robots which leave the family store of their parent when they become adults and create their own family store have family stores which not only are small but – what is more important – are shared by closely kin-related robots because only parents and their pre-adult offspring share the same family store. This convinces the robots to put their food into their family stores. Parents help their pre-adult offspring and pre-adult offspring help their parents and their siblings.

External stores are useful because, when there is little food in the environment, a robot can eat the food contained in its external store and this reduces the probability of dying for lack of energy. This leads to the prediction that small family stores will lead to larger population sizes – which is one measure of fitness or adaptation. As we have seen, community stores disappear because the robots are not disposed to contribute to a store which is shared by individuals which are only distantly genetically related to them. Small family stores do not disappear and, since external stores are a useful adaptation because they allow a robot to survive by eating the food contained in its external store, population size should be larger. This is what we find (see Figure 13). After the usual initial oscillation, population size reaches a value of 75–80 individuals for the robots that, when they become adults, create their own family store, while for the robots that never leave their original family store the corresponding value is around 30 individuals.

The robots which, when they reach adulthood, leave the family store of their parents and create their own family store, have small family stores because only parents and their pre-adult offspring share the same family store. But is the restricted number

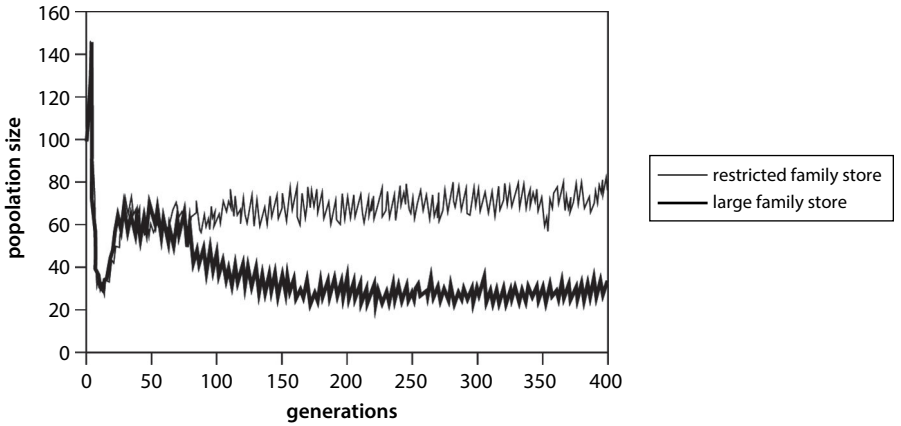


Figure 13. Population size of robots that share the same family store for the entire length of their life (large family stores) and for robots that, when they become adults, leave the family store of their parents and create their own family store (restricted family stores).

of robots that share the same store, by itself, which explains the adaptive value of a shared store and its continued existence, or the adaptive value of a shared store requires that the store be shared by genetically related robots? To answer this question we evolve another population of robots and in this population, when a new robot is born, the robot is not assigned to the family store of its parent but to the family store of another, randomly chosen, robot. In other words, the newly born robot is "adopted" by another robot. Like the preceding robots, these robots create their own family store when they reach adulthood and, therefore, a store is shared by the same restricted number of robots. But now the stores are adoptive family stores because the robots that share the same store are genetically unrelated robots. Do the robots – both parents and adoptive offspring – put their food in the shared store although the store is not a (genetic) family store and the food which they put in the store can be eaten by robots which do not have their same genes? The answer is in Figure 14, which shows the tendency to put one's food in a shared store when the store is shared by the same restricted number of robots but in one case the robots are genetically related and in the other case they are genetically unrelated.

The difference is clear. The robots that share a store with a restricted number of other robots which are their kin have a greater tendency to put their food in the shared store than the robots which share the store with the same restricted number of other robots but these robots are not their kin. Genetic family stores are different from adoptive family stores.

But it is interesting that these robots do not completely ignore their adoptive family store. When the store is a community store, the tendency to put one's food in the store is almost zero. When the store is an "adoptive" family store, the tendency to put one's

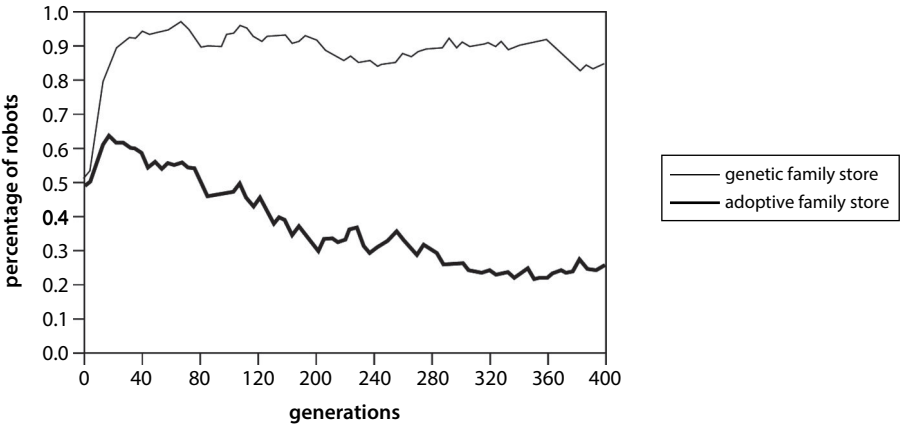


Figure 14. Percentage of robots that put their food in their family store when it is their genetic family store and when it is an adoptive family store.

food in the store is much lower than if the store is a genetic family store but it is not zero. External stores are useful – whatever their specific nature – because they help to survive in difficult environments. Although the adopted robots and their adoptive parents in many cases do not put their food in their adoptive family store, still they appear to be under a pressure not to let their adoptive family store disappear completely. Why? The size of the store may be the crucial factor. Our robots do not put their food in the community store because the community store is very large and this implies that the food that a robot puts in the community store will probably be eaten by another robot. Adoptive family stores are small and, therefore, a robot which puts its food in its adoptive family store is likely to be the same robot which eats the food. Therefore, adoptive family stores may still function as a "safety net" for the robots. There may be other explanations of a more "psychological" nature for the behaviour of treating one's adoptive offspring as natural offspring and one's adoptive parent as natural parent, but to test these explanations, we need to construct more "psychologically" sophisticated robots with motivations and emotions (see Chapter 2). However, we predict that these more "psychological" explanations will also be based on kin selection theory.

So, family stores are useful adaptations only if they are restricted to individuals which have very similar genes. But why human beings have family stores instead than individual stores? As we have seen, individual stores function as a "safety net". So we should ask: Are family stores better than individual stores as a "safety net"? Since population size is our measure of adaptation, to answer this question we look at the size of the population of robots in the two cases. We evolve two populations of robots, one with individual stores and the other with (small) family stores, and we determine what is at the end the size of the two populations. What we find is that there is a small advantage for the robots which have individual stores rather than family stores (see Figure 15).

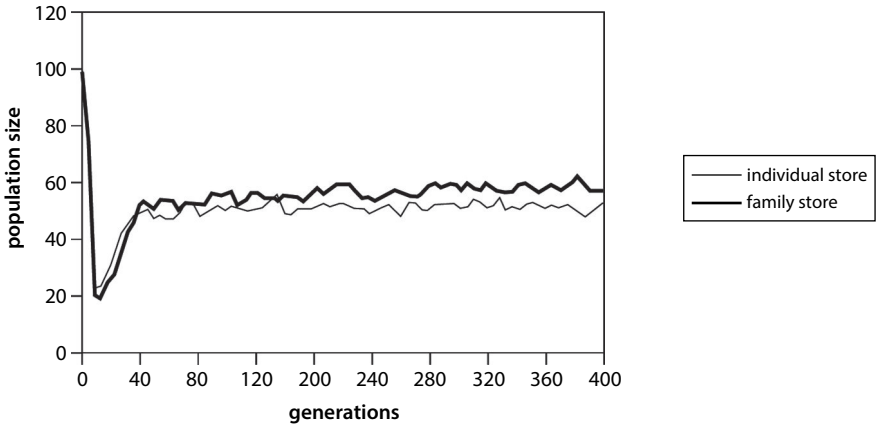


Figure 15. Population size of robots that have individual stores and robots that have family stores.

This result is interesting and it may reveal an important fact about family stores and, more generally, about shared external stores. If the robots have family stores, a robot which is not very good at finding food might equally survive because it can eat the food contained in its family store. This is impossible for the robots which have individual stores. A robot which is not very good at finding food will die even if it has an individual store because its individual store will be empty. This implies that the robots which have family stores are exposed to a weaker selective pressure compared to robots with individual stores, and this weaker selective pressure translates into a lower level of the capacity to find food and, therefore, into a smaller population size. To test this hypothesis we bring the robots of the two populations into an experimental laboratory which contains a certain number of food tokens and we measure the capacity of the robots to reach the food tokens. The results are that the robots with family stores reach fewer food tokens than the robots with individual stores. Having a family store implies to rely on others for survival and this reduces the selective pressure for evolving a good food-finding capacity. (We return to this phenomenon in the next section dedicated to central stores.)

But these results, clearly, raise a problem. If individual stores and family stores have more or less the same adaptive value, or individual stores are even slightly better than family stores, why all human communities have family stores and not individual stores? The answer is in one characteristic which is uniquely human. For a long period after birth, human infants are unable to feed themselves and, therefore, if left alone, they die. They must be fed by someone else and who feeds them are their parents. According to the theory of kin selection, parents feed their dependent offspring because this is the only way to insure the presence of their genes in future generations. For human beings it is not sufficient to have offspring. To leave their genes to future generations, they must feed their offspring for a long period after birth because, otherwise, the offspring

will die before reaching the age in which they themselves can reproduce – and the genes shared by parents and offspring will be lost. (We have already described robots that must feed their offspring to leave their genes to future generations in Chapter 2 and in Chapter 7.)

To test this idea, we evolve two new population of robots which, when they are born, for a period of 300 cycles cannot move and therefore cannot feed themselves. Their only chance to survive and reach the age in which they are able to move and look for food is that they eat the food contained in their family store. We compare two populations of robots, one with family stores and the other one with individual stores, and the results are very clear. The robots with family stores and dependent offspring have a larger population size than the robots with individual stores and dependent offspring (see Figure 16). This tells us that the existence and the importance of family stores for human beings is the long period after birth during which they depend on their parents. And they may also suggest why the relation between parents and offspring is not a symmetrical one. (On this, see below.)

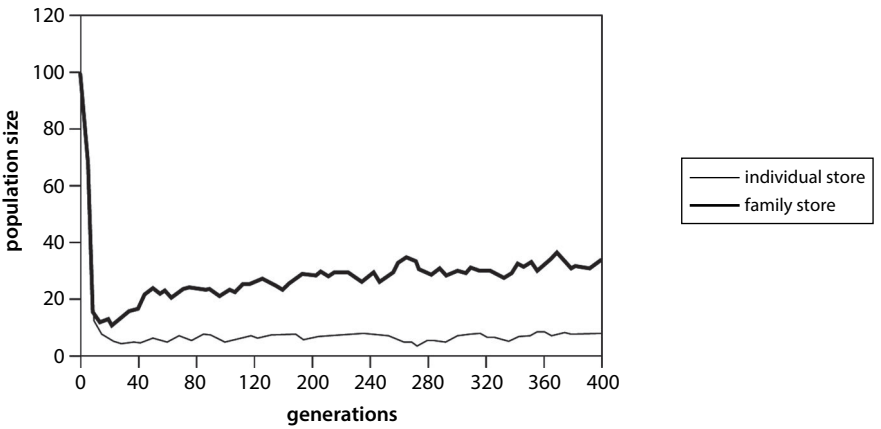


Figure 16. Population size of robots that have family stores and robots that have individual stores when, for a period of 300 cycles after birth, offspring robot are unable to feed themselves.

As the beginning of this Chapter we have said that external stores are an important human adaptation because they are at the origin of many typically human phenomena and are necessary to explain many aspects of human societies. One of these phenomena is economic inheritance. If the robots do not have external stores in which they put their food, when a robot dies nothing special happens because the robot leaves no food. But if the robots have external stores, when a robot dies the food contained in its external store can be inherited by the robot's offspring – transferred to the family store of the offspring. In the robots described so far, when a robot dies and all its offspring already have their own family store, the food contained in the parent's store

simply disappears. We now construct another type of robots. When a robot dies and its store is not empty, the food contained in its store is moved to the stores of its adult offspring – although this only happens when all the robot's offspring have become adults and have left the family store of their parent to create their own family stores.

Economic inheritance has many important consequences for the organization and functioning of human societies but here we only consider its consequences for the difference in wealth among the robots. Wealth is the number of food tokens contained in a robot's external store. Family stores can contain more or less the same number of food tokens (small differences in wealth) or some family stores may contain many more food tokens than other family stores (large differences in wealth: rich and poor families). The difference in wealth between two family stores is originally due to a difference in the capacity to find food between the robots of the two families. But economic inheritance can also have an impact on the differences in wealth in a population of robots. What is this impact? The answer depends on how economic inheritance is realized. The food contained in a robot's external store when the robot dies can be equally distributed among all the robot's offspring or it can be inherited by one single offspring – say, the eldest one. What are the consequences of these two different inheritance schemes? If the food left by a dead parent goes to only one of its offspring, the differences in wealth among the family stores of the robots are larger than if the food is equally divided among all the offspring. The selective inheritance of the parent's food by the eldest offspring creates a small number of rich robots and a large number of poor robots. In contrast, if the inherited food is equally distributed among all the offspring of a dead robot, economic inheritance has no impact on the distribution of wealth, and wealth has the same distribution which it has in a population of robots without economic inheritance – or it can even make the distribution of wealth more egalitarian (see Figure 17).

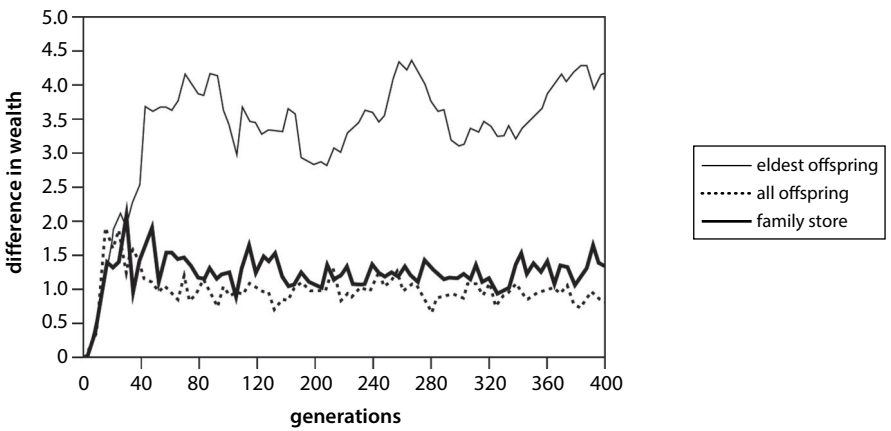


Figure 17. Differences in wealth in populations of robots in which the eldest offspring inherits all the parent's food, or the parent's food is equally distributed among all the robot's offspring, or there is no form of economic inheritance.

In the past, many human societies have adopted an inheritance scheme according to which all the wealth of a deceased parent goes to the eldest (male) offspring but in modern Western-like societies this is less true and law favors more egalitarian inheritance schemes. This raises some interesting questions. Since in modern human societies the distribution of wealth is far from egalitarian, there must be other factors in these societies that produce a non-egalitarian distribution of wealth. One factor is that modern societies have reduced the number of offspring per individual. Human beings live longer but they have fewer offspring and, therefore, even egalitarian inheritance schemes cannot reduce the differences of wealth. This has already happened in economically developed societies and it is happening in emerging societies today.

But the critical factor which explains the increase in the differences in wealth in modern human societies is the increase in total wealth. If there is little total wealth, large differences in wealth are impossible because all individuals will only have what is strictly necessary for their survival. But an increase in total wealth causes an increase in the differences in wealth even with egalitarian inheritance schemes. The offspring of a rich individual all equally inherit much wealth when their parent dies and the offspring of a poor individual all inherit the little wealth of their poor parent, so that the differences in wealth remain or are even increased. Modern societies have more total wealth than past societies for a variety of reasons but the most important reason is that in modern societies goods are used to produce new goods. For our robots food is the only good but for human beings many more things are goods. If our robots were able to produce new goods by using the goods stored in their family store, economic inheritance, however realized, would increase the differences in wealth because the robots that inherit more goods from their parents would be able to produce more goods so that wealth becomes cumulative and wealth differences larger.

Economic inheritance is the transfer of food from the family store of a parent robot to the family store of its offspring when the robot dies. But some food can be transferred from the store of a parent robot to the store of its adult offspring (without obtaining anything in exchange) even if the parent robot is still alive. Imagine that an offspring robot has no food in its store and, therefore, the robot risks dying together with its pre-adult offspring. Will some food be transferred from the parent's store to the offspring's store? The offspring robot already has its own family store but, since it has the same genes of its parent, this may push the parent robot to give some of its food to its adult offspring. To test this idea, we compare two populations of robots. In one population, when a robot's family store is empty, nothing happens and both the robot and its pre-adult offspring die. In the other population, when the family store of a robot is empty, some food is automatically transferred from the family store of the robot's parent to the family store of its offspring so that its offspring – and the offspring of its offspring (its grandchildren) – can remain alive. We compare the size of the two populations and the results are those of Figure 18.

The robots which help their adult offspring when they need help have a larger population size than the robots which ignore their adult offspring which need help.

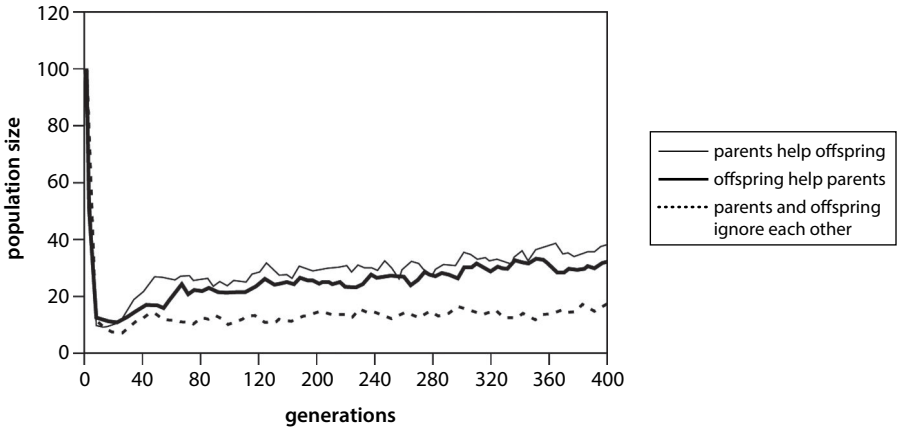


Figure 18. Population size for three populations of robots: parents help their adult offspring, adult offspring help their parents, parents and offspring ignore each other.

This also is predicted by the theory of kin selection. The ultimate "goal" of an individual is that its genes will continue to exist in future generations. Since the genes of its offspring are a copy of its genes (aside from mutations), by giving some of its food to its offspring which need food to remain alive, a robot guarantees that its genes will continue to be part of the genetic pool of the population. And by giving its food to its adult offspring in need, the robot does not only help its offspring but also its grandchildren to remain alive – which also is predicted by kin selection theory. (On this, see Chapter 7 dedicated to robotic families.)

Population size is our measure of fitness for these robots and, since the robots that give their food to their adult offspring in need have a larger population size than the robots which do not do this, we expect that the robots will evolve the behavior of giving their food to their adult offspring in need. Now we ask the opposite question: Do offspring robots which already have their own family store help their parents which risks dying because they don't have food in their store? To answer this question we also compare two populations of robots. In one population, when an offspring robot leaves the family store of its parent and it creates its own family store, all links between the two stores cease to exist and parents and offspring ignore each other. The second population is the mirror image of the preceding population in which parents give their food to their adult offspring in need. Now, when a robot's store is empty, the adult offspring of the robot give some of their food to their parent. The results are that, in this case too, the behavior of helping one's parent in need leads to an increase in population size (see Figure 18), although the increase is somewhat smaller than in the population in which parents help their offspring in need.

Kin selection theory predicts that parents will help their adult offspring in need. But why should an adult robot help its parent in need? This question is interesting

because most human beings do help their parents in need and they tend to do so until their parents die. Kin selection theory explains this behavior too. An adult robot may help its parent in need for three reasons: (1) the parent might help its offspring in some future occasions; (2) the parent might generate other offspring which have the same genes of the robot (they are its siblings) and the robot is interested in the continuing existence of the copies of its genes possessed by its siblings; (3) the robot can keep in life its already existing siblings which still are in the parent's family store because they do not have reached adulthood. Kin selection theory directly generates these predictions but, in this case too, there may be other explanations of a more "psychological" nature which are more indirectly derived from kin selection theory. For example, to increase the probability that it will be helped by its parent, an offspring robot must show that it cares for its parent and, to show that it cares for its parent, the robot must help its parent in need.

Something similar happens among sibling robots. Imagine two sibling robots which already have their family stores. If one of the two robots has no food in its store, will its sibling give some of its food to the robot to keep it alive? Kin selection theory predicts that in this case too siblings will help each other, although the picture is more complex because of sibling rivalry: siblings compete with each other for their parent's food.

To compare what happens with the three different types of help – from parent to offspring, from offspring to parent, from sibling to sibling – we adopt another, more direct, approach. We construct three different populations of robots, we add to the robots' genotype a gene which encodes the probability that the robot will help a relative in need, and we see how the value of the gene changes in a succession of generations. The results are that the evolved value of the gene is highest in the populations in which parents help their offspring, somewhat lower in the populations in which offspring help their parent, and much lower in the populations in which help is among siblings – although it is not zero even in this last case (see Figure 19).

So far, external stores have been abstract entities which are intended to reproduce the phenomenon of ownership. A robot or a family of robots has an external store if the robot or the members of the family can eat the food present in the external store and no other robot can eat that food. But real external stores tend to be physical structures such as houses and storehouses which exist in specific places and cannot be easily moved from one place to another place. If external stores are abstract entities, they do not impose any limitation on the movements of the robots. The only problem for the robots is to search for food in the environment and, when they find food, they automatically "put" the food in their external store. But if external stores are physical structures – such as houses – they limit the movements of the robots. The robots must move in the environment searching for food but they must also continuously return to their house to store the food or to eat the food contained in their house – and this, clearly, has a cost in terms of both time and energy. In fact, if the robots could see not only the food tokens but also their house, we expect that the robots will look for food

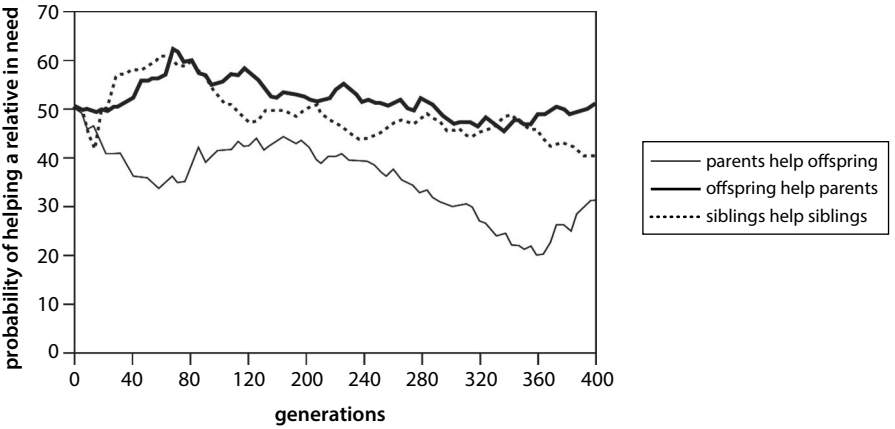


Figure 19. Evolutionary change in the value of a gene encoding the probability to help a relative in need. Parents help their offspring somewhat more than vice versa. Siblings help each other much less, although even in this case the probability is not zero.

only in the space which is around their house – unless there is little food around their house. For these robots the environment ceases to be a homogeneous space. It is an environment which is divided into the environment near to their house and the rest of the environment (see Figure 20).

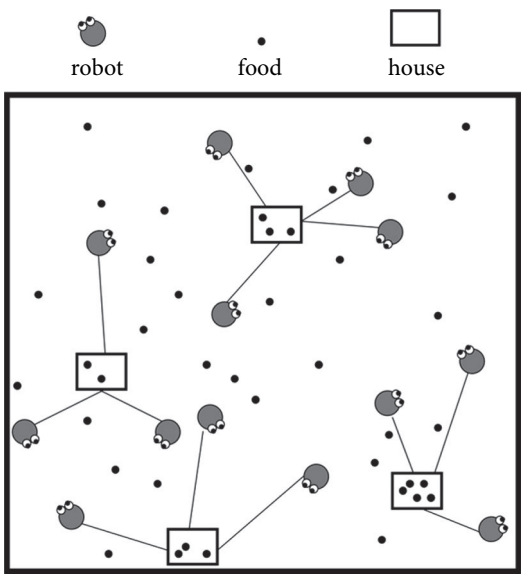


Figure 20. Robots whose family stores are houses tend to live near to their house.

This has various important consequences. In Chapter 7 dedicated to robotic families, we have seen how robots acquire the capacity to use physical landmarks to return to their house, and this may create a more articulated map of the physical environment in their brain. But other consequences are of a more social and political nature. If family stores are too close to one another, this may result in overcrowding and the depletion of food around one's family store. Therefore, we predict that, if we add to the robots' genotype a gene which encodes where a robot locates its family store – its house – when it reaches adulthood and we let the value of the gene evolve, the gene will encode a sufficient distance from other existing family stores. Another consequence concerns the ownership of land. The space around a robot's house becomes the robot's space and ownership is extended to pieces of land. Our robots are hardwired by us to respect ownership. They cannot eat the food contained in the external store of another robot and they cannot look for food in the piece of land around the house of another robot. But if they were left free to decide whether to respect or not to respect ownership, this would pose a problem. Nonhuman animals directly defend the things they own, and this was also true for ancient human beings. The owner of some good is personally in charge of preventing other individuals from taking its good. In more recent human societies, the defense of ownership is one of the most important goods produced by the central store (the state) and distributed to all the members of the community. We have described robots that are punished by the state if they damage other robots in Chapter 7. In the last section of this Chapter we describe robots that have a central store – a state.

As we have said, if external stores as physical structures, they can lead to geographical dispersal. The robots locate their external stores far from the external stores of other robots in order not to interfere with, and not to be interfered with, by other robots and, therefore, the robots prefer to live in different places in the environment. But if external stores are family stores, family stores as physical structures can, on the contrary, lead to geographical grouping. Our robots help their adult offspring that risk dying from lack of food and this means that, if an offspring robot has no food in its family store, some of the food existing in the family store of its parent is transferred to the offspring's store. And the same for the robots that help their parents or their siblings in need. If external stores are abstract entities, the transportation of food from one store to another store is purely virtual – as it is for our robots. But if external stores are physical structures located in specific places, transporting food from one store to another store can be costly and it requires that the physical distance between the two stores is not too great. To solve this problem, we might add another gene to the robots' genotype which specifies how far from its parent's family store an offspring robot will place its family store when the robot reaches adulthood. Since, as we have seen, helping and being helped by one's parent increases fitness, if being near in physical space is a precondition for helping and being helped, we predict that the gene will evolve a value which dictates a small distance between the two family stores. And this extends to the

family stores of sibling robots because the family stores of sibling robots are near to the family store of their parent and, therefore, they are all near to one another. This leads to the emergence of robotic villages, where a village is a group of family stores (houses) of genetically related robots which are located near to one another in physical space.

But external stores as physical structures can have other implications of a geographical nature. If we start from a certain number of robots randomly distributed in the entire environment, the family stores of these robots will also be randomly distributed in the entire environment. However, with the successive generations, many of the robotic lineages become extinct and at the end there are only one or a few lineages, where a lineage is a group of robots which all descend from a single robot of the initial population. The fact that the offspring robots place their family stores near to the family store of their parents has the consequence that, at the end, a robotic village will be located in one particular zone of the environment, the zone near to where the original ancestor of the lineage lived and had its family store. Again, this imposes a structure to an initially unstructured environment: the environment is divided into a part which is inhabited by the robots and the rest of the environment which is not inhabited by the robots. But this condition can change. Overcrowding can cause some robots to move to parts of the environment which are still not inhabited, these robots may create their family stores far from the family stores of their parents and their siblings, and their family stores progressively develop into a new village. The new village is initially inhabited by robots which are genetically related to the robots living in the original village but then, with the successive generations, the genetic distance increases. (Cf. the simulation of the expansion of Indo-Europeans and their languages described at the end of the preceding chapter.)

This does not necessarily mean that the two villages ignore each other. They can interact in a variety of ways. They can exchange goods. (For the exchange of goods, see Chapter 11 on robotic economies.) Or one village can try to enter into possession of the goods of another village by the use of force – war – or two villages can be become allied or fuse together for making war against a third village. (For robots that make wars, see Chapter 10 on political robotics.) And, if the robots are male and female, new family stores may be created by a male robot and by a female robot which come from different villages so that the two villages become genetically linked in this indirect way. In some cases a village becomes larger and larger and it is transformed into a small town or a city, and this is inevitably accompanied by a progressively increasing genetic distance among the robots that live in the town or city.

External stores may have been both cause and effect of many human adaptations. A capacity to predict and plan for the future may have been a *pre-condition* for constructing external stores but it may also have been an *effect* of having external stores because the advantages of having an external store have been a pressure to develop and exercise a capacity to predict and plan for the future. (On predicting and planning, see Chapter 5 dedicated to mental life.) Specialization in the production of goods may have

been a *cause* for constructing external stores in which to put the goods one is specialized in producing in order to exchange them with other goods but they may also have been an *effect* of having external stores because external stores make it possible to have goods that can be exchanged with other goods. (On specialization and exchange, see Chapter 11 dedicated to robotic economies.) The greater number of different things which are goods for human beings compared to nonhuman animals may have been the *cause* for the emergence of external stores in which to put all these things but the cause/effect relation can also go in the opposite direction and the increase in the number of different things which are goods for human beings may have been the *effect* of having external stores. (For the much greater number of things which are goods for human beings compared to nonhuman animals, again see Chapter 11 dedicated to robotic economies.) External stores may have favoured the development of technology since it does not make much sense to invest in the construction of artefacts if one cannot store them for future use. But, again, the development of technology may have been a cause for the development of physical stores – pottery, houses – because physical stores are technological artefacts. Economic inheritance, and its many consequences for the economic and political organization of human societies, may have been a cause of the emergence of external stores but it might also have been an effect of having external stores. And finally, living in an environment of physical external stores – houses – may have led to what has been called the "domestication" of the human species, with changes in perception of both the social and non-social environment and the division of the social world in a public world and a private world. And, at the same time, these "psychological" changes may have led to the construction of physical stores as private and protected environments in which to live. To put it shortly, and strongly, external stores have made us human.

4. Central stores

Our family stores are shared by a small number of robots which are closely related genetically. Ants have very large family stores but ants have many offspring and, therefore, their large family stores are also shared by closely related individuals. Human beings have a small number of offspring and, therefore, their family stores are small.

But human beings do have external stores which are very large (in modern societies, millions of individuals) and are shared by genetically unrelated individuals. These very large social stores are called "states" and, if our robots must be human robots, we must construct robotic "states": communities of genetically unrelated robots which have a single shared store. We call this single shared shore the "central store" of the community and constructing communities of robots which have a central store is the first step towards the development of a political robotics. (The next chapter is dedicated to political robotics.)

In the preceding Section we have seen that large stores of genetically unrelated robots tend to disappear because the robots do not to put their food into an external store to which a large number of genetically unrelated robots have access. So, how can states arise and remain in existence? To answer this question, we must first look at the advantages of possessing a central store to which all the members of a community of robots give some of their food and which redistributes this food to all the members of the community.

We compare two populations of robots, one with a central store and the other one without a central store. The environment is divided into cells and some of these cells contain a food token (see Figure 21).

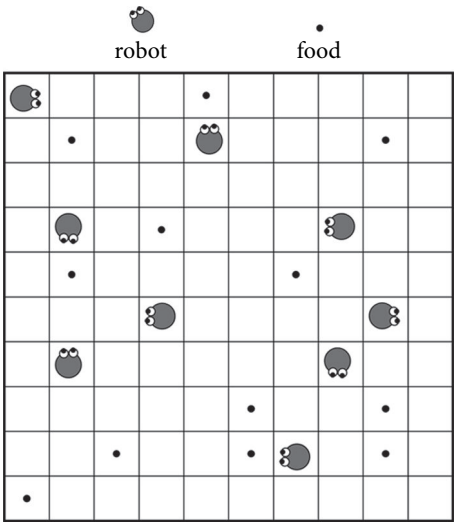


Figure 21. The environment of the robots is made of cells and a robot or a food token occupy one cell.

The robots do not have wheels and we do not reproduce the physical movements that allow a robot to displace itself in the environment. With their visual neurons the robots see the food token which is nearest to them and with their motor neurons they respond in four possible ways. They move to the cell which is front of them, they turn 90 degrees to the left or to the right, or they do not do anything. When a robot enters into a cell which contains a food token, the food token disappears because the robot the robot takes possession of it. Therefore, the number of food tokens is progressively reduced but the environment is a seasonal one and at the beginning of each season new food tokens appear in the cells of the environment. In all other respects the robots are like our preceding robots. They have a certain amount of energy in their body which is consumed by a fixed quantity in each cycle and they die if their energy reaches the

zero level. As always, in the initial generations the number of robots decreases because the robots are not very good at procuring food and, therefore, they die early and they have few offspring. Then, the selective reproduction of the best robots and the addition of random mutations make the robots acquire the capacity to find the food tokens and their number rises again until it reaches a stable level.

As we have said, we compare two populations of robots. The robots of one population do not have any type of external store and they immediately eat the food tokens they are able to find in the environment. The robots of the other population have a central store to which the robots give some of the food which they are able to find in the environment (see Figure 22). But the robots do not decide whether to give or not to give their food tokens to the central store. We hardwire them to eat only 1 out of 4 food tokens they find in the environment and to give the other 3 food tokens to the central store – this is like a tax rate of 75%. And the central store periodically redistributes the food tokens equally to all the members of the community.

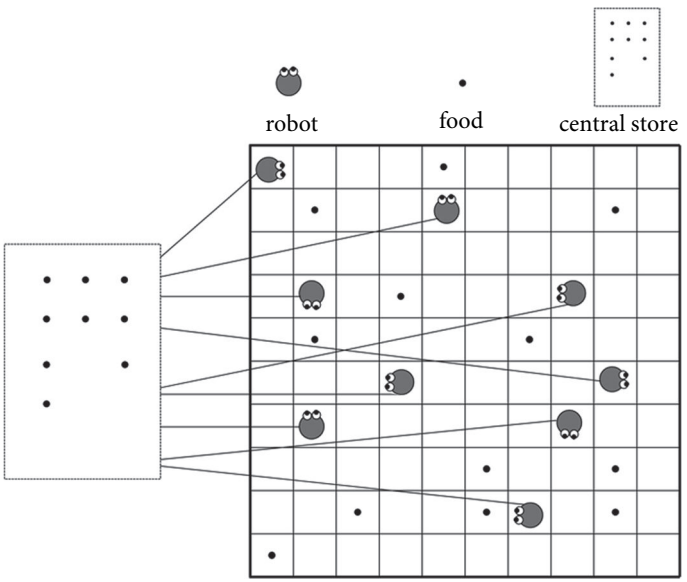


Figure 22. Same robots of Figure 21 but with a central store to which the robots give some of the food they are able to find in the environment.

What are the results? Population size is the measure of fitness for these robots and Figure 23 shows how population size changes across time for the two populations of robots.

After the initial drop due to the inability of the robots to find the food tokens, the number of robots increases in both populations until it reaches a stable value which depends on the quantity of food contained in the environment. But the increase is

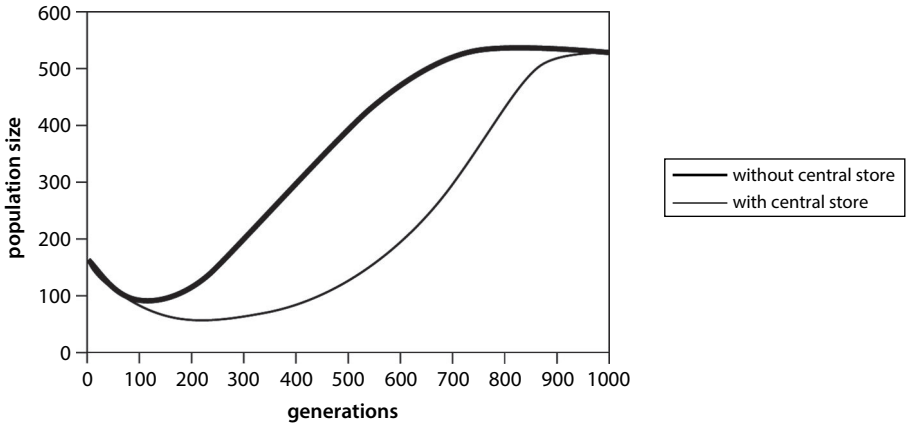


Figure 23. Evolution of population size for robots that do not have the central store and for robots that have a central store to which they give 75% of their food. The central store redistributed equally the food to all robots that are members of the community.

different for the two populations of robots. It is faster for the robots which *do not have* the central store than for the robots which have the central store, although at the end population size is the same for both populations of robots. How can these results be explained? The explanation is that the existence of the central store reduces the selective pressure on the robots. If the robots do not have the central store, the probability that a robot will generate an offspring depends entirely on the robot's capacity to find food. These robots do not have any form of external help and they can only rely on themselves. On the contrary, if the robots have a central store and they mostly live with the food redistributed by the central store, the reproductive chances of a robot depend on the *other* robots and on *their* capacity to find food. (As we have seen, to some extent this is also true for family stores.) A robot is less responsible for its destiny and the selective pressure for developing a good food-finding capacity is reduced. Even the robots that are not particularly good at finding food can survive and generate offspring because they can eat the food redistributed by the central store. And, since there is less selective pressure, the capacity to find food of the robots which have the central store takes more time to evolve compared to the robots which do no have the central store.

This is what we also find if we bring the robots into an experimental laboratory and we measure their capacity to find food. The laboratory contains a certain number of food tokens and we count the number of food tokens reached by a robot during a fixed period of time. The capacity to find the food tokens increases more rapidly across the successive generations in the robots that do not have the central store compared to the robots which do have the central store – although in this case too, at the end the two types of robots are equally good at finding food (see Figure 24).

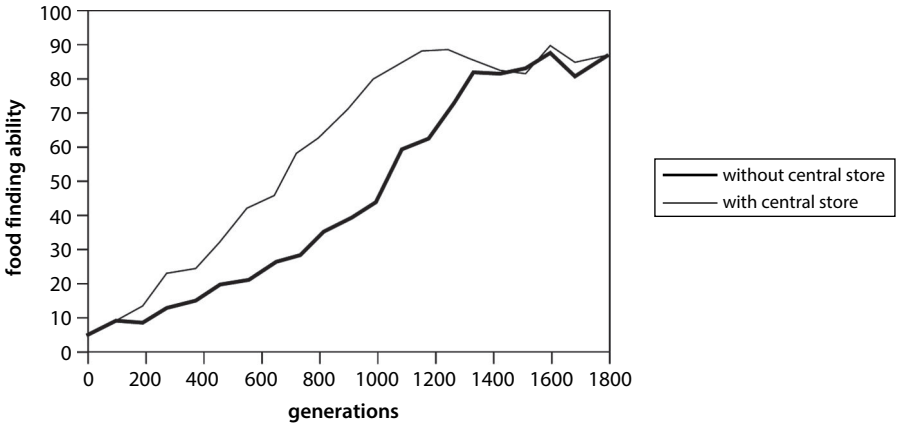


Figure 24. Number of food tokens eaten in an experimental laboratory by the robots of the successive generations for robots that have and for robots that do not have a central store.

This result might not seem to be so important because at the end both the size of the population and the capacity to find food are the same for the robots without the central store and for the robots with the central store. But this conclusion would be wrong. Time is important. If the two populations of robots were to compete with each other for the food present in the environment, the robots without the central store would win the competition because for many generations they are better at finding food than the robots with the central store. They eat more and they have more offspring. Therefore, the robots without the central store would survive while the robots with the central store would become extinct.

So, an individual survival strategy (no central store) is better than a strategy based on the existence of a central store because the individualistic strategy creates a stronger selective pressure which leads to a more rapid evolution of the capacity to find food. But if this is true, how can we explain that most human beings live in communities which have a central store – the state? One answer is that, like all external stores, central stores are important in difficult environments and in difficult times. We do another simulation in which, after the robots have evolved the capacity to reach the food tokens and their population size has reached a stable state, the environment changes and it becomes a less good environment: less food appears at the beginning of each season and, therefore, at the end of a season, there is very little food in the environment. While in the past the robots were able to survive during the lean periods at the end of a season, now they find themselves with very little food in these periods and they risk extinction. This is when having a central store turns out to be useful. The individualistic robots cannot do much to avoid extinction. If the environmental changes are really bad, a population of individualistic robots without a central store is very likely to become extinct. In contrast, the robots which have a central store

can protect themselves from the negative effects of the changes in the environment because during the harsher winters they can survive by eating the food redistributed by the central store.

This is what we find. After the environment has worsened, population size changes for both the robots without a central store and the robots with a central store. But the robots without a central store cannot do much in the new unfavourable environment and they become extinct. In contrast, the robots with a central store avoid extinction. Their population size is reduced because of the inferior carrying capacity of the new environment but they are able to survive (see Figure 25).

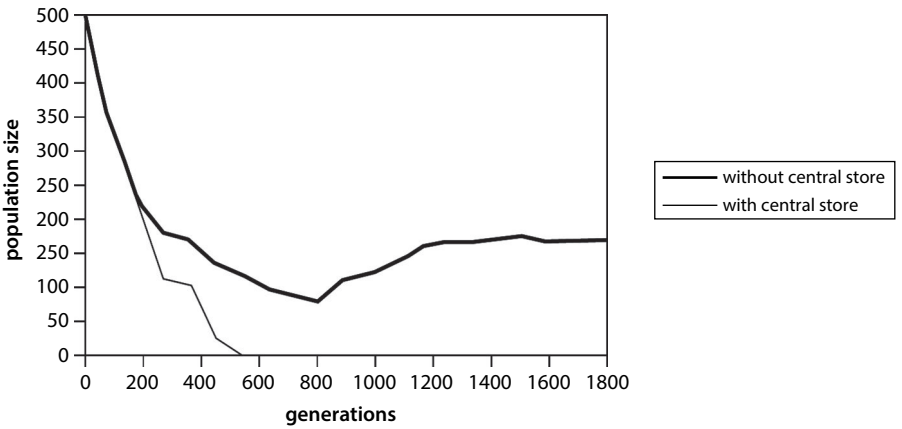


Figure 25. When the environment deteriorates, the robots with a central store become fewer in number but they survive while the robots without the central store become extinct.

So, a possible cause of the emergence of central stores is that central stores make it possible to survive in difficult environments and in difficult times. Central stores may be a useful adaptation if a population migrates from an original, more favourable, environment to a new, less favourable, environment – which is what may have happened to some ancient human populations. Or the new environment may not be intrinsically worse than the original environment but it may pose new problems because it is new and the population must adapt to the new environment. Central stores may help to survive during this period of adaptation to the new environment. Or central stores may be useful to traverse the empty space between zones of the environment containing food because the robots can eat the food redistributed by the central store while traversing the empty space. As we have seen, this can also be done with individual stores but central stores are much larger and richer than individual stores and they make it possible for the robots not only to traverse an empty space between two zones with much food but also to settle in the empty space by adopting new ways of obtaining food which take time to develop – for example, irrigation agriculture.

But if central stores are useful adaptations, will they actually emerge and become a stable property of a population or robots? The existence of a central store clearly depends on the willingness of the robots to give some of their food to the central store. Our robots give 3 out of 4 of their food tokens to the central store but they do not autonomously decide to give their food to the central store. We hardwire them so that they give their food to the central store. But hardwiring can only be a provisional research strategy because in biology and in behaviour nothing is hardwired but everything is evolved or learned. So our next question is: What happens if we let the robots themselves decide whether to give or not give their food to the central store?

To let the robots decide, we add a "social gene" to the robots' genotype which encodes a robot's propensity to give or not to give its food to the central store. (This gene is like the gene of the robots with family stores.) The value of the gene varies between 0 and 1 in different robots, where 0 means that, when the robot finds a food token, there is 0 probability that the robot will give the food token to the central store – the robot is completely selfish – while a value of 1 means that there is 100% probability that the robot will give the food token to the central store – the robot is completely altruistic. In more human terms, the robots with a high value for their "social gene" pay their taxes to the state while the robots with a low value for their "social gene" tend not to pay their taxes. They are free riders: they take from the central store but they do not contribute to the central store.

The robots of the initial population have a random value for their "social gene", which means that some robots are more socially responsible and they are more likely to give their food to the central store while other robots are less socially responsible and they are less likely to give their food to the central store. When a robot generates an offspring, the offspring robot inherits the value of the "social gene" of its parent with some random mutations that can make the offspring more or less socially responsible compared to its parent, that is, more or less disposed to give its food to the central store (to pay taxes).

Our prediction is that the robots which live longer and have more offspring are the less social ones, that is, the robots which have a lower value for their "social gene" and, therefore, give less of their food to the central store. Why do we make this prediction? If a robot is not socially responsible and it does not give its food to the central store, the robot eats both the food it finds in the environment and the food redistributed by the central store. Therefore, the robot will have more offspring compared to another, more socially responsible, robot which gives its food to the central store. The results confirm this prediction. The asocial robots generate more offspring than the social robots and, after a certain number of generations, almost all robots do not give their food to the central store and, therefore, the central store disappears. The robots become a population of individualistic robots without a central store.

This is very similar to what we have found with the robots that never leave the family store of their parents and create their own family store. In a few generations family stores become very large and they are shared by robots which are only very distantly

related genetically and, as a consequence, the robots cease to put their food into their family store because their family store has become a community store. And this similarity is not surprising. All human communities are made up of individuals that can be said to be genetically related because all human beings have a common genetic origin. But their genetic relatedness is so vanishingly small that they tend to treat each other as genetically extraneous. In our robots family stores appear and remain in existence because giving one's food to the family store increases the survival chances of other robots that have the same genes of the giver. But with central stores the situation is different. If the robots are left free to decide whether to give or not to give their food to the central store, central stores do not emerge because giving one's food to the central store increases the survival chances of genetically extraneous robots and most robots don't want to give their food to genetically extraneous robots.

But there are other problems with central stores. The body of our robots consumes a fixed quantity of energy in each cycle and, therefore, for these robots looking or not looking for food does not make much of a difference because their consumption of energy is the same both when they look for food and when they do nothing. (Remember that these robots have the possibility to respond to the sight of a food token by doing nothing.) We now construct another population of robots which consume their bodily energy *only* when they are looking for food. If they do nothing, the energy in their body remains the same. Again, we compare two populations of robots, one with a central store and the other one without a central store, and in both populations the robots do not decide to give or not give their food to the central store but they are hardwired by us to give some of their food to the central store. What we want to know is if they look for food (they work) or they prefer not to look for food (they do not work).

What do the robots do? The answer is that the robots which do not have the central store rarely choose the option of not working because their life depends entirely on their capacity to find food and, if they don't work, they do not eat and they die. The situation is different for the robots that have the central store. Many robots choose the option of not working because in this way they avoid consuming their bodily energy but they can still eat the food redistributed by the central store. The problem is that, when an increasing number of robots choose the option of not working, little food is captured by the robots in the environment and, therefore, the central store becomes empty and it has no food to redistribute. The final result is that the central store disappears and the robots become extinct.

So a central store can disappear for two reasons. If the robots are left free to decide whether or not to give their food to the central store, they do not give their food to the central store – they do not pay their taxes – and they eat both their food and the food redistributed by the central store. If the robots are hardwired by us to give their food to the central store but looking for food is costly, the robots do not to look for food because they can eat the food redistributed by the central store without paying any costs. In both cases, these selfish behaviours progressively penetrate the population of robots and the central store disappears. This leads to a paradox. In the last 5000 years

an increasing number of human beings have lived in communities possessing a central store – a state to which all the members of the community give some of their goods and which redistributes these goods to all the members of the community. If, as our robots tell us, central stores cannot exist, how can this have happened?

The paradox is particularly disturbing if we consider that, for a community of robots, having a central store is very advantageous. Central stores make it possible for certain robots to survive which would otherwise die. A robot may be intrinsically unable to find food or it may be unable to find food when it reaches old age. Or the robot may incur some illness which prevents the robot from moving in the environment looking for food. In all these circumstances, if someone else does not provide the robot with food, the robot dies. The robot's relatives can give some of their food to the robot and they will generally be ready to do so. (See the section on family stores.) However, in many cases the problem cannot be solved by the robot's relatives because the robot in difficulty may not have living relatives or the robot's relatives may not have enough food to give to the unable, old, or ill robot.

Another advantage of central stores is that central stores can facilitate the development of specialization in the production of goods. For our robots food is the only good. But human beings have many different types of goods, not only food. If some robots specialize in the production of one good and other robots specialize in the production of other goods and if the robots give the goods they produce to the central store, the central store can redistribute the goods among the robots so that all the robots have all types of goods. The central store functions as a mediator for the exchange of goods – which may have been an important motivation for the emergence of ancient states. (We describe robots that specialize in the production of different goods in Chapter 11 dedicated to robotic economies.)

But the most important advantage of having central stores is that central stores can produce goods that no robot would be able to produce by working alone. Examples are systems of rules that make the behaviour of other robots predictable and reliable, routes and other infrastructure, health and educational systems, mechanisms for defining, discovering, and punishing behaviours that damage others, and the capacity to make offensive or defensive wars. The central store collects goods – under the form of taxes – from all the members of the community and it uses them to produce these new goods which are then made available to all members of the community.

Given all these advantages of central stores, how can the obstacles to the emergence and maintenance of central stores be overcome so that there may be communities of robots which, like all (modern) human communities, have a central store – the state? The answer is the existence of "chief" robots that have the power to make all the robots of the community do what they must do to keep the central store in existence. We describe robotic communities which have a "chief" in the next chapter dedicated to political robotics.

CHAPTER 10

Political robotics

ME knows that central stores – or "states", as human beings call them – exist because they produce important goods for all the members of a community. But it also knows that central stores have a problem. They must function according to rules, and there must be someone who decides these rules and guarantees that they are applied by punishing those who do not respect them. Human beings have found a solution to this problem: they have created political chiefs. Political chiefs decide what the rules of functioning of the central store are and they make all the members of the society behave according to these rules.

But this poses another problem: who are the chiefs? In the course of their history, human societies have experimented with many solutions to this problem and in modern Western societies and in societies that want to be like Western societies the chief is chosen by all the members of the society. The problem is that chiefs must be leaders. They must know what is the common good and they must want to realize the common good. ME finds that often chiefs are not leaders in this sense, and this is especially true in today's society which are so complex that it is difficult to know what is the common good and have such sophisticated communication technologies that candidate chiefs can easily persuade the members of the society to choose them even if they are not leaders.

When it arrives on the Earth, ME discovers that human beings are experimenting with an alternative solution to this problem: using the new digital technologies of communication to create societies which do not have chiefs and in which all the members of the society directly contribute to finding the best rules of functioning for the society. ME thinks that this solution, at least for now, has its problems because it may be impossible for a digital society to know what the common good is and to do some of the things that only chiefs can do such as punishing behaviours that damage others and doing wars with other communities.

1. Chief robots

In the preceding chapter we have seen that for a community of robots it is very advantageous to have a central store to which all the members of the community give some of their food and which redistributes this food to all the members of the community. But we have also seen that there are two obstacles to the existence of central stores. If left free to give or not to give their food to the central store, most robots do not give their food to the central store – they try not to pay taxes – and, if left free to look or not to look for food, most robots do not look for food – they do not work – and, therefore, they have no food to give to the central store and they live with the food redistributed by the central store. In both cases, the central store becomes empty and it disappears. How can these obstacles be overcome so that there may be communities of robots with a central store?

The answer is chief, or governing, robots. A chief robot is a robot which has the power to make all the members of a community of robots behave as they must behave in order to keep the central store in existence – and this explains why all human communities, beyond a certain size, have chiefs. Power is an important phenomenon in the life of all – or almost all – social animals but it is especially important in the social life of human beings. A makes B do what A wants B to do by using a variety of different means. This is *social* power in general. The power of chiefs is *political* power. Political power is power on *all* the members of the community and it is based on punishing the members of the community who don't do what chiefs want them to do.

But chief robots have another task to accomplish and this task derives from a characteristic of central stores which distinguishes central stores from family stores. Unlike family stores, central stores function on the basis of explicit rules. What percentage of its food must a robot give to the central store? Must all the members of the community give the same percentage of their food to the central store or rich robots must give a higher percentage than poor robots? To whom must the food collected by the central store be re-distributed? Must it be re-distributed equally to all robots? The rules for collecting and redistributing food are only one example of rules that chief robots must define and implement if they must be human chief robots. The life of modern human communities are regulated by all sorts of explicit rules because only if a community functions on the basis of explicit rules, the members of the community can predict and rely on the behaviour of the other members of the community and of the central store itself, and both the central store and the entire community can function appropriately. Chief robots have the task to define these rules and to guarantee that they are respected by punishing the members of the community which do not respect them.

Finding the best rules of functioning for a community is an intrinsically difficult task but the task is made more difficult because the central store must function for the common good – it must benefit all the members of the community, including future members. And what is the common good is not easy to determine and there may be

different views on what is the common good. For human beings food is not the only good. Human robots must give their money to the common store – they must pay taxes – and the common store must use this money to produce new goods: hospitals, schools, roads and other infrastructure. The robots may disagree on which goods should be produced by the central store and distributed to all the members of the community and which goods should be produced by "private" organizations of robots and sold to those members of the community that can buy them. (On private organizations of robots that produce goods, see the next chapter on robotic economies.) The robots may have contrasting interests and, therefore, contrasting views on the rules of functioning of the central store. For example, the robots which have more food in their family stores (the rich robots) may prefer that the members of the community give a lower percentage of their food to the central store (lower tax rate) and that all robots give the same percentage (same tax rate), while the robots with little food in their family stores (the poor robots) prefer high tax rates for all and higher tax rates for rich robots than for poor robots. Other contrasts may arise because there may be different categories of robots and each category has its own interests which do not coincide with the interests of the other categories. For these and other reasons, the robots will have different visions of what is the common good so that the responsibility of defining the common good is left to the chief robot.

Defining the common good has always been a difficult task but, today, this task is much more difficult because today's societies are more complex than past societies. They are made of a greater number of different categories of individuals, activities, and goods, and they are very complex networks of cause-effect relations. An example is deciding the fiscal regime. In modern societies there are many different categories of tax-payers and many different types of taxes and, for a chief robot, defining the fiscal regime is a very complicated task since each decision has both good and bad consequences. (At the beginning of XX century the American fiscal rules were contained in a book of 400 pages. Today, the American fiscal book is more than 70.000 pages.) This, of course, is only one example of the greater complexity of modern societies and, to understand the problems associated with this complexity, we use a very simple and abstract model.

In a society a certain number of possible events can happen and, of these events, some are good and some are bad. The events have cause-effect links, which means that, when one event happens, other events will also happen. Therefore, if the chief robot decide to make an event happen, all the direct and indirect consequences of that event will also happen. Past human societies were simpler. There were fewer possible events and fewer cause-effect links between events. Today's societies are more complex because the number of possible events and of cause-effect links among events has become much greater (see Figure 1).

What are the consequences of living in a simpler society with fewer possible events and fewer cause-effect links among events and living in a more complex society with

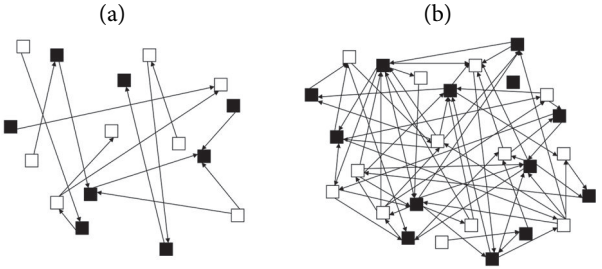


Figure 1. Some of the events that may happen in a society are good (white squares) and some bad (black squares) and the events are linked by cause-effect links (arrows). In past societies (a) there were fewer events and cause-effect links than in today's societies (b).

more possible events and more cause-effect links? Imagine a robotic society in which the chief robot decides to make a good event happen. If we repeat this for a certain number of times and each time we choose randomly the good event that the chief robot makes happen, we find that in a society in which the possible events and the cause/effect links between events are fewer in number, it is not so difficult for the chief robot to decide what to do so that what it does has more good than bad consequences. But in today's more complex societies which contains many more possible events and many more cause/effect links among events, whatever the chief robot does, there are more or less the same number of good and bad consequences.

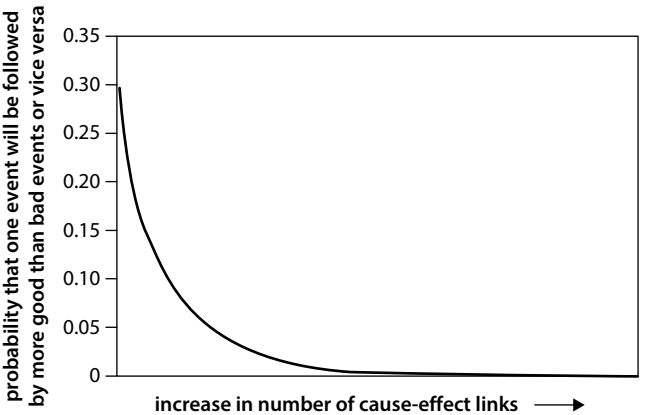


Figure 2. With the increase in the number of events and cause-effect links between events, the probability that an event will be followed by more good than bad events or vice versa becomes zero and, therefore, for a chief robot it is difficult to decide what to do.

This is a very abstract model of human societies but it tells us that, in today's complex societies, it is often impossible to solve one problem without creating other problems and it is very difficult to choose the rules of functioning for the central store that make the central store function for the common good. This is why choosing the appropriate individual as the chief of the community is so important in today's complex societies. The problem of choosing the chief of a community is a recurrent problem because chiefs die or they go out of office and, when one chief dies or goes out of office, another chief must replace the preceding chief. In the course of history human societies have experimented with a variety of solutions to this problem. Chiefs can be individuals who become chief by the use of force. Or they may be the son or, more rarely, the daughter of the preceding chief, or they may be chosen by the preceding chief. Another, more recent, solution is that all the members of the community choose the chief. A certain number of individuals declare what they will do if they become the chief of the community, all the members of the community beyond a certain age vote for one of these individuals, and the individual who is chosen by the largest number of voters becomes the chief of the community. This solution to the problem of choosing the chief of a community is called democracy.

All these methods for choosing the chief of a community have their limitations and they do not guarantee that the chief will adopt the best rules of functioning of the central store – the rules that realize the common good. The non-democratic methods have many defects but even adopting democracy as the method for choosing the chief of a community has its limits, and democracy has three limits.

The first limit is that candidate chiefs *want* to be chosen as chief. Being the chief of a community satisfies the human desire to have power – to make others do what one wants them to do – and it has so many other advantages that candidate chiefs will do all sorts of things to be chosen as chief. Instead of trying to find the rules for the functioning of the society that realize the common good, candidate chiefs will propose programs that favour the interests of one sub-group if this increases the probability that they will be chosen as chiefs and, once they are elected chiefs, they will adopt decisions in favour of the sub-group which has elected them. Or the candidate chiefs will use all sorts of techniques to convince the members of the community to vote for them. Why this is a problem? This is a problem because a candidate chief may convince voters of the goodness of rules for the functioning of the society which are *not* in their interest. Inducing others to do things which are not in *their* interest is a common strategy in the social life of human beings – and modern communication technologies and marketing techniques make this strategy very powerful. Candidate chiefs will be chosen on the basis of their capacity to convince the members of the community to vote for them rather than on the basis of their capacity to find the best rules of functioning for the society.

The second limit of democracy is that, even if candidate chiefs propose well-thought programs to voters, most voters do not really understand these programs and, therefore, they cannot choose the best among them. Today's societies are very

complex and most of their members do not have the time, motivation, and background knowledge which are necessary to understand what are the best rules of functioning for the society. Voters vote without knowing and understanding how the society functions and, therefore, they are unable to distinguish among the candidate chiefs on the basis of their programs. They vote on the basis of their interests, their preconceived ideas and ideologies, and the capacity of the candidate chiefs to convince them.

The third limit of democracy is that the members of the society do not directly contribute to finding the best rules of functioning of the society, they often do not know if what a candidate chief has promised to do has been actually done after he or she has been elected, and they cannot do much if it has not been done.

Notwithstanding these limits, democracy is considered today as the best method for choosing the chief of a community. This may be true but it is not clear why we should not try to go beyond its limits. In the last chapter of the book, we return to these problems of democracy as the method for choosing the chief of a community and we propose to use human robots to solve them.

2. Chiefs and leaders

But the most serious limit of democracy as a method for choosing the chief of a community has to do with the distinction between chiefs and leaders. As we have said, candidate chiefs want to be chosen as chiefs and, to be chosen as chiefs, they will not necessarily propose the rules of functioning of the central store that would lead to the common good but they will propose the rules they think the people who choose them, their electors, want. In this way they are chiefs but not leaders because they do not *lead* but are *led*. But if the central store must function for the common good, chiefs must be leaders. In his book *On Politics* the Greek philosopher Aristotle wrote that chiefs must be individuals who "can anticipate with intelligence" and we interpret this as meaning that chiefs must know what the common good is and they must lead the community towards the common good. Our next robots show how a robot which – literally – sees the common good can be useful to a group of robots.

A group of robots live in an environment with a prey which is so large and strong that all the robots of the group must be near to the prey to attack and kill the prey. (The prey is an object that does not move.) For these robots, being all near to the prey is the common good because it is the only way for them to eat. The robots have visual sensors that allow them to see the prey but they can see the prey only when they are near to the prey. Therefore, when they don't see the prey, the robots can only move randomly in the environment until they happen to see the prey. (The behaviour of approaching the prey when a robot sees the prey is hardwired in these robots.) The life of the robots is a succession of episodes. At the beginning of each episode the robots are everywhere in the environment and the episode terminates when all the robots are near to the prey so that they can kill the prey (see Figure 3).

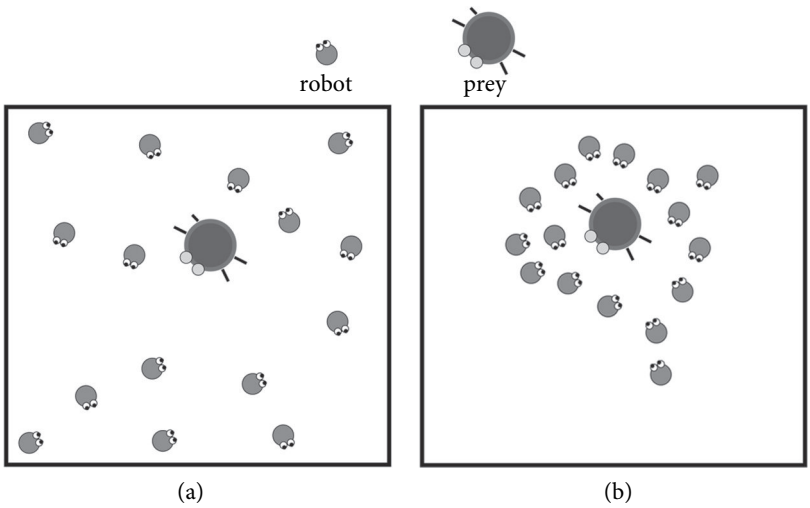


Figure 3. The robots live in an environment with a single large prey (a) and, to attack and kill the prey, the robots must all be near to the prey (b).

The robots take much time to kill the prey because, to attack the prey, they must wait until the last robot of the group has reached the prey. If we measure success as the number of cycles taken by the last robot of the group to reach the prey, we find that these robots, on average, need 5427 cycles to kill the prey.

This is our first type of robots. A second type of robots do not all have the same visual capacities but one robot of the group can see the prey from a greater distance (see Figure 4). What are the consequences for the group of including a member which has a better sight? The results are that there are no advantages. The single robot which can see the prey from a greater distance generally reaches the prey earlier than the other robots but, as a group, the robots take the same long time to reach and kill the prey as the preceding robots: 5459 cycles.

In a third type of robots, one robot is the chief of the group. The preceding robots saw the prey but not each other. These new robots still do not see each other but they see the chief robot and, therefore, they follow the chief robot in its movements in the environment. (The robots are hardwired to follow the chief robot.) What happens with these robots? Even if the chief robot sees the prey at the same distance as the other robots, the robots of the group reach the prey in less time than the robots without a chief: the last robot takes only 1800 cycles to reach the prey. Therefore, for these robots it is advantageous to have a chief even if the chief robot is not better than the other robots and it sees the prey at the same short distance. The advantage is that the robots do not explore the environment one independently of the others, like the robots without a chief, but they explore the environment together, as a group. The robots do not see each other and they only see the chief but they move together because they all follow

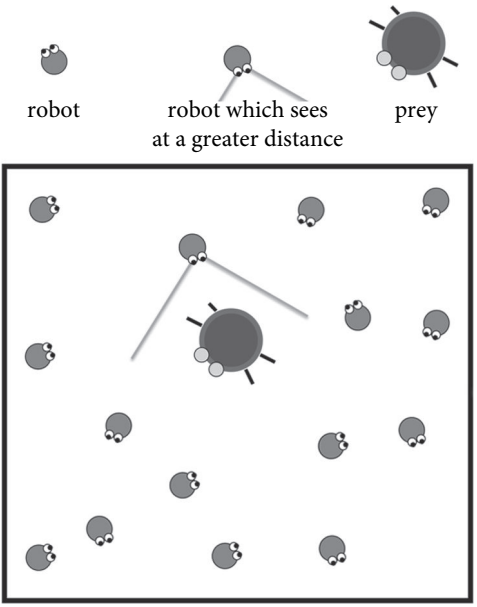


Figure 4. One robot sees at a greater distance than the other robots.

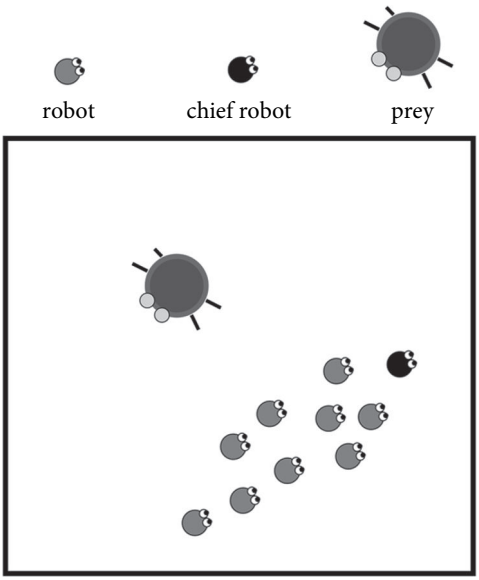


Figure 5. One robot is the chief of the group and is followed by the other members of the group although it sees at the same distance as the other robots.

the chief. Moving as a group is advantageous because, when one of the robots of the group, not necessarily the chief, sees the prey, the other robots are near and they also will soon see the prey (see Figure 5).

Do these robots take less time to reach the prey because they have a chief or because they move together as a group? Our robots do not see each other but, if they could see each other, they could move as a group even without a chief. Some animals do exactly this. An individual responds to the sight of another individual by remaining in proximity to the other individual and, as a result, all the individuals move together as a group even without a chief. For these animals moving together is adaptive because it makes it easier to find and kill a prey – as in these robots – or to avoid being killed by a predator – as we have seen with the groups of robots that function as "information centres" described in Chapter 6.

But, if it is possible to move as a group without a chief, why have chiefs? To answer this question we construct a fourth type of robots. Like the second type of robots, one robot of the group has better visual capacities than the other robots of the group but now *this* robot is the chief of the group. The other robots follow this robot because it is their chief but now the chief sees the prey from a greater distance (see Figure 6).

These robots obtain the best results: the last robot of the group reaches the prey in only 657 cycles. The chief robot quickly discovers the prey because it is also a leader in the sense that it sees the prey from a greater distance. The other robots follow the chief/leader so all the robots of the group are able to reach the prey in a much shorter time.

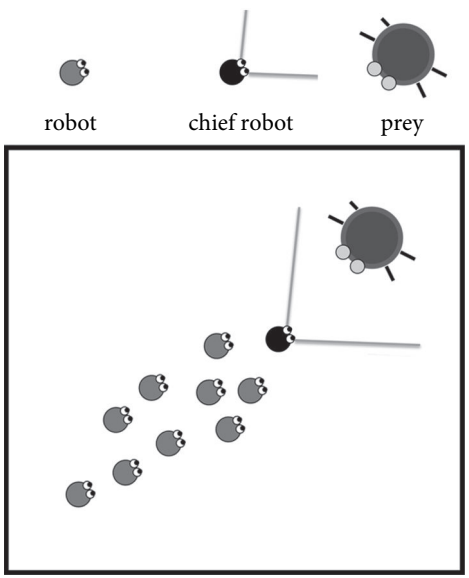


Figure 6. The robot which sees at a greater distance than the other robots is the chief of the group.

So Aristotle was right: chiefs must be leaders. Our robots would not need to have a chief if all the robots of a group could see the prey from a great distance. But in human beings it rarely happens that all the members of a group can see at a great distance and can "anticipate with intelligence". Our robots do not all see the prey from a great distance and, if the robots do not all see the prey from a great distance, having a chief is particularly useful if the chief is a leader – it sees at a greater distance than the other members of the group. And what is important is not that the group has both a chief and a leader but that the chief and the leader are the same robot.

For these robots we decide if they have a chief/leader and which robot is the chief/leader. The next step is to construct robots that autonomously acquire the behaviour of acting as leaders or as followers. The new robots are predators like the preceding robots but they are different from the preceding robots because they live in an environment which contains many small preys which also are robots and which run away from the predator robots to avoid being killed by them (see Figure 7).

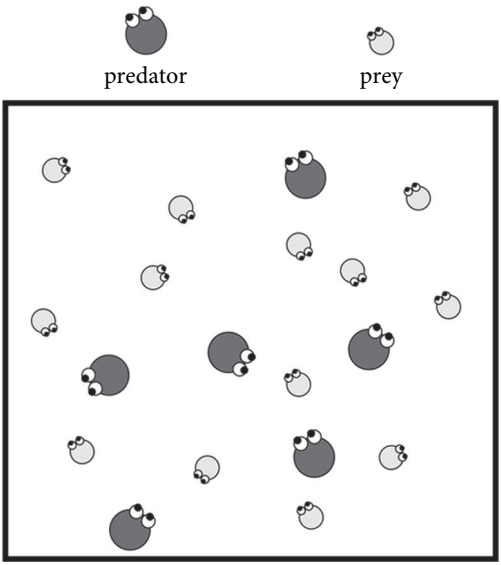


Figure 7. The predator robots live in an environment which contains many smaller prey robots that the predator robots must be able to reach and kill to remain alive and reproduce.

Both the predator robots and the prey robots evolve – co-evolve – their behaviour in a succession of generations. The predators evolve the capacity to reach and kill the preys and the preys evolve the capacity to flee away from the predators – although the best predators reach and kill the preys which are not very good at fleeing away. The fitness of a predator is how many preys the predator kills during its life while the fitness of a prey is how long the prey is able to remain alive without being killed by a predator. The preys have visual sensors which allow them to see the predators. The predators

have visual sensors which allow them to see both the preys and the other predators. But the visual sensors of both the predators and the preys have a limited range and, therefore, it can happen that a predator does not see any preys but it only sees other predators – its conspecifics.

What we find at the end of evolution is that, when a predator sees a prey, it ignores the other predators and it tries to reach the prey. But when a predator does not see any preys and it only sees one or more of its conspecifics, we observe an interesting phenomenon. Some predator robots ignore the other predators and autonomously explore the environment looking for preys, while other predators follow those predator robots – which therefore act as leaders. This is like what happened with our preceding robots which had a leader but what is new is that the behaviour of being a leader or being a follower is not hardwired by us but it is the spontaneous result of a process of evolution. The predator robots do not only evolve the behaviour of reaching and killing the preys when they see them but, in addition, some robots evolve the behaviour of autonomously exploring the environment in search for preys (leaders) and other robots evolve the behaviour of following a leader (followers).

We see these differences among the predator robots if we look at their behaviour in their natural environment but we determine more precisely if a predator is a leader or a follower by testing the robots in an experimental laboratory. We do two experiments. In one experiment we put a single predator robot into a completely empty environment and we determine how much of the environment the robot explores by counting the number of different parts of the environment – pixels on the computer's screen – that are visited by the robot. Some robots move in a way that does not allow them to explore much of the environment while other robots explore more of the environment. This is the first experiment. In the second experiment we put a robot of the first type (non-explorer) together with a robot of the second type (explorer) into the same environment and what we find is that the non-exploring robot follows the exploring robot in its movements in the environment – with the result that the non-exploring robot also explores the environment like the exploring robot (see Figure 8).

Being a leader is useful because leaders autonomously explore the environment and, therefore, they are more likely to find the preys. But why being a follower is useful? Imagine a robot which, at a certain time, does not see any preys. If the robot reacts to this situation by not exploring the environment, the robot dies because it may never see a prey and it cannot eat. Being able to explore the environment when they do not see any preys is so important for our robots that, if they lived alone, all the predator robots would need to evolve the behaviour of exploring the environment. But our predator robots do not live alone. They live with their conspecifics and this allows them to find a social solution to the problem. Some predator robots evolve the behaviour of leaders: they ignore their conspecifics and autonomously explore the environment in search of preys. Other predator robots evolve the behaviour of followers: they follow the leaders and in this way they also explore the environment. Of course, there is also a third type of unlucky robots which neither autonomously explore the environment

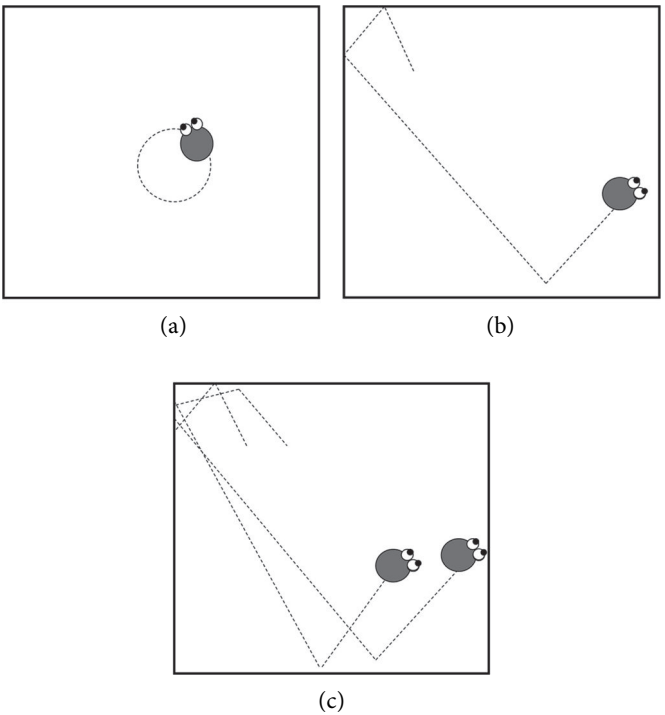


Figure 8. One robot explores a little part of an empty environment (a) while another robot is much more explorative (b). If we put the two robots together, the first robot follows the second robot and, therefore, it also explores the environment (c).

nor do they follow a conspecific which explores the environment – but these robots are unlikely to have a long life.

So, even if they autonomously evolve their behaviour and their behaviour is not hardwired by us, some robots turn out to be leaders and other robots followers. These robots cannot be said to have a common good because the robots compete with each other for the number of preys which each robot is able to eat and the followers follow the leaders because they find this behaviour useful. But if the robots had a common good, the problem would be to choose a chief robot which is also a leader.

3. No chiefs

But perhaps there is another solution to the problem of realizing the common good: abolishing chiefs and replacing them with communication among all the members of the community. To explore this possibility, we return to the robots which, to attack and kill the large prey, must all be near to the prey. These robots do not see each other and do not communicate with each other. When one robot sees the prey, it limits itself to

approaching the prey. But now we ask: What happens if, when a robot sees the prey, the robot communicates the location of the prey to the other robots? With communication, can the robots achieve the common good – which, for these robots, is that they are all near to the prey so that they can attack and kill the prey – with no need for a chief?

To explore the possibilities offered by communication to achieve the common good, we go back to the robots which all have the same limited visual capacities and do not have a chief with better visual capacities that they can follow so as to reach the prey more rapidly. These robots take much time to reach the prey because the only thing which they can do is move randomly in the environment until they happen to see the prey. Since to attack and kill the prey all the robots must be near to the prey, the robots must wait until the last member of the group reaches the prey – and this takes time. We now add communication. When a robot sees the prey, the robot communicates the location of the prey to the other robots so that the other robots can cease to move randomly in the environment and they can immediately approach and reach the prey. (The robots are hardwired to approach and reach the prey if they know the location of the prey.) We construct two different types of robots which have one of two different types of communication system, a more ancient or a more modern one. The ancient communication system depends on spatial proximity: one robot can communicate with another robot only if the two robots are near to one another in physical space. The modern communication system has no need for spatial proximity: two robots can communicate with each other independently of where the two robots are in physical space (see Figure 9).

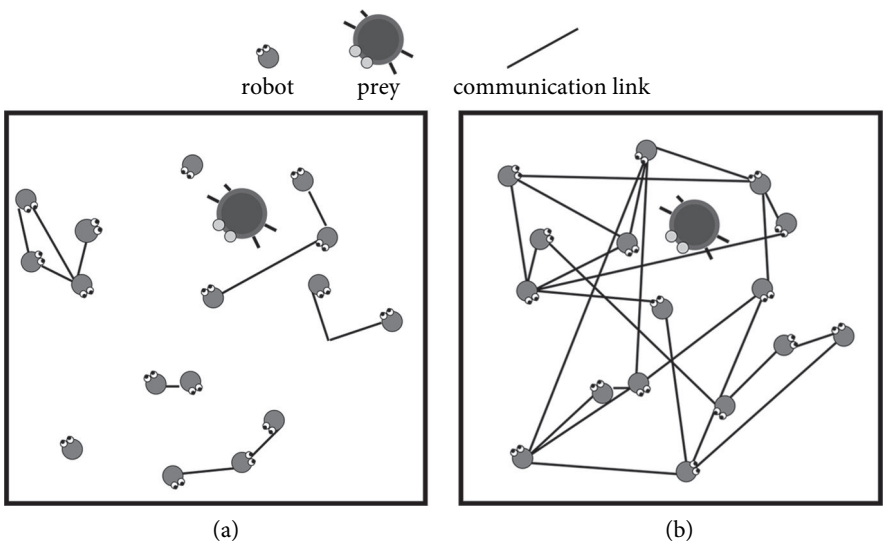


Figure 9. (a) Ancient communication system: the robots communicate only with robots that are near to them. (b) Modern communication system: the robots can communicate with all other robots independently of space.

Is communication useful? Is one communication system better than the other? For these robots success is measured as the time taken by the last robot of the group to reach the prey and, as we have seen, for the robots without communication the last robot needs 5427 cycles to reach the prey. Communication drastically cuts down this number. If the communication system is the more ancient one which allows the robots to communicate only if they are physically near to one another, the last robot of the group reaches the prey after only 720 cycles. If the robots use the more modern communication system which is independent of space, the time is even less: 95 cycles.

Given these results, we might conclude that communication makes chiefs superfluous. If there is no communication, when for purely chance reasons one robot sees the prey, this is not particularly useful to solve the collective problem of being all together near to the prey. But if there is communication, the robot can communicate to the other robots the location of the prey and, in this way, all the robots can quickly reach the prey with no need for a chief. It is the structure of the group which changes. If there is a chief, the group has a vertical structure. The chief is at the top of the structure and it tells the other robots what they must do. As we have seen, this is useful to the group, especially if the chief is a leader and it sees at a greater distance than the other robots. On the contrary, if there is no chief, the structure of the group is a purely horizontal one. No robot has power on other robots and no robot needs to see at a greater distance than the other robots because communication replaces chiefs and replaces the differences in visual capacities. Communication is useful in all cases but if the communication system is the modern one which is independent from physical space, the robots solve the problem of being all near to the prey at the same time really quickly. And what is interesting is that chiefs can even be damaging to the group if all robots communicate with all other robots. In groups of robots in which each robot can communicate with all the other robots and there is no chief, the time taken by the last robot to reach the prey is only 95 cycles. If we add a chief, the time becomes 679 cycles. In a society in which everyone can communicate with everyone, chiefs are not only useless but harmful.

This is why the Internet can be seen as the basis of a purely horizontal society without chiefs in which the rules of functioning of the central store are directly decided by the citizens: e-democracy. But e-democracy also has its problems. Today's societies are very complex and most citizens do not have the time, motivation, and knowledge which are necessary to identify the best rules that allow the society to function for the common good. As we have said, this problem also exists for representative democracy – democracy with chiefs – but it may be more serious for an Internet-based horizontal society in which the Internet is used – as it is used today – only as a communication and coordination tool. Our robots live in a simple environment in which it can easily happen that one of them finds the prey even if the robots explore the environment randomly – and the robot tells the other robots where is the prey. But today's human beings live in a much more complex environment and, in this more complex environment, finding the rules that make the central store function for the

common good is difficult, especially because, as we have seen in Section 1, in complex societies all decisions tend to have both good and bad consequences. If Internet is only a communication and coordination tool but it does not provide knowledge and understanding to all citizens, an Internet-based horizontal society cannot function because zero knowledge plus zero knowledge is zero knowledge. (In Chapter 15 we discuss how human robots can help human beings to solve this problem.)

Purely horizontal societies have other problems. Behaviours that damage others are inevitable in human societies and, as we have seen in Chapter 6, if human societies are unable to contain these behaviours, they risk dissolving because the disadvantages of living together become greater than the advantages.

To link this problem to the problem of horizontal vs vertical societies, we change the robots that must all be near to the large prey to kill the prey. For the new robots it is sufficient that a certain number of them, not necessarily all the robots of the group, are near to the prey to kill the prey. However, the food obtained by killing the prey is distributed to all the robots of the group, including those that have not participated to the killing of the prey. The other change is that, unlike the preceding robots, for these robots moving has a cost in terms of energy and this cost reduces a robot's fitness. Therefore, to save their energy some robots tend not to move and not to go in search of the prey because they can equally eat the food distributed to all the members of the group when the prey is killed. This is a problem because, if too many robots do not move, too few robots go in search of the prey, the prey is not killed, and no robot can eat. This problem is the same problem that we have discussed in Chapter 9 and the problem can only be solved if there is one robot which discovers and punishes the robots that do not go in search of the prey. The robot which discovers and punishes the robots which do not go in search of the prey is the chief of the group. A purely horizontal societies without a chief has more difficulty discovering and punishing socially damaging behaviours, and this is why purely horizontal societies appear to be, at least for now, impossible.

Chief robots are also necessary for another reason: war. When one hears of robotic wars, what immediately comes to mind are drones, soldier robots, robots that find and defuse mines, robots which do reconnaissance tasks. These are all technologies which help human beings to conduct *their* wars but it is not what we have in mind here. Human beings have engaged in wars for millennia and, therefore, if we construct robots to better understand human beings, we must construct communities of robots that engage in wars with other communities of robots.

Wars are attempts by a community of individuals at taking possession of the goods of another community of individuals by the use of force – offensive wars – or defending one's community from such attempts – preventive and defensive wars. If a single individual tries to take the goods of another individual, the probability of success depends on the physical strength and on the skills of the two individuals – and this is what generally happens among nonhuman animals. But when an entire community of

individuals tries to take possession of the goods of another community of individuals by the use of force, other factors determine which of the two communities wins the war.

Imagine two communities of robots that live in two different zones of the environment, and each zone contains one type of goods. If the robots want both types of goods, the two communities can exchange their goods. One community gives some of its goods to the other community and, in exchange, the other community gives some of its goods to the first community. (In the next chapter dedicated to robotic economies, we describe robots that exchange their goods with other robots and merchant robots that buy the goods of one community and sell these goods to another community.) But the two communities can also adopt another strategy: they can try to take possession of the goods of the other community without giving anything in exchange. This requires making war to the other community. The result of the war may depend on a variety of factors such as the war tools available to the two communities, the nature of the environment in which the two communities live, and the size of the two communities, with larger communities being more likely to win a war against smaller communities. This last factor represents a pressure to increase the size of communities, and such a pressure may have played an important role in the growth of the size of human communities from families to chiefdoms, and then to states and empires. (See the final section of this chapter.)

But how wars have originated? To answer this question we compare two populations of robots. The robots live in an environment with randomly distributed food tokens but, while the robots of one population are free to look for food in the entire environment, the environment of the other population is divided into four zones and the robots that are born in one zone cannot go into the other zones and eat the food which is found there. The robots generate one offspring at regular intervals and, since the robots that live longer, generate more offspring, the size of the two populations varies in the succession of generations. What is, at the end, the size of the two populations of robots? The robots that are free to look for food in the entire environment have a larger population size than the robots that cannot look for food outside the territory in which they were born (see Figure 10).

The explanation is the following. Since the food tokens are randomly distributed in the environment, there may be periods in which there is little food in one zone and more food in another zone. When this happens, the robots that can look for food in the entire environment can move from a zone with little food to a zone with more food and, in this way, they may continue to eat and remain alive. The robots that cannot go outside the zone of the environment in which they were born do not have this possibility and, if they happen to live in a zone which temporarily has little food, they die early and generate few offspring. Therefore, the total number of robots that live in the four small zones is smaller than the number of robots that live in a territory which coincides with the entire environment. The simulation stops here but we expect that, since living in a larger territory represents an advantage, the robots will try to enlarge

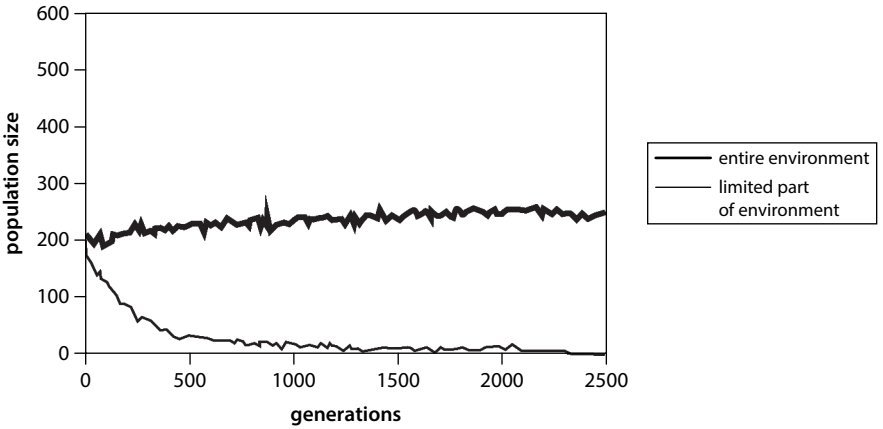


Figure 10. Population size of robots that can look for food in the entire environment and of robots that can look for food only in a limited part of the environment.

their territory and, if the environment around their territory is already occupied by other robots, to make war with the other robots to take possession of their territory. (See the simulation of the changes in human settlements in ancient Southern Etruria described in the last section of this chapter.)

We are talking of wars here because wars necessarily require chiefs who coordinate the activity of the members of the community which is engaged in a war against another community. The community's chief must make and change war plans with the necessary speed and it must be immediately obeyed by all the members of the community which is engaged in the war. This is why horizontal societies without a chief cannot make wars – and this would be an important advantage of horizontal societies.

There is another link between wars and horizontal societies. Making war to take possession of the goods of another community is a form of socially damaging behaviour in which the damaging and the damaged agents are not members of the same community but are two different communities. But while the other-damaging behaviours that occur inside a community can be punished by the chief of the community, this is not possible for the socially damaging behaviours of one community towards another community because there is no chief that has the power to punish a society for attacking another society. And this is why human societies will continue to make wars for a long time – until all human beings become members of a single community. (But human beings also make civil wars between two different parts of the same community.)

Wars have played, and continue to play, an important role in human societies and, if we want to understand what causes wars, how two communities can become allies in a war against a third community, what determines the final results of a war, how

wars have changed in the course of human history, what are the alternatives to wars, and what is the future of wars, we must construct communities of robots that make wars. This is a task for the future but we have made an attempt at reproducing some of the effects of wars in ancient societies and we describe this attempt in the next and final section of this chapter.

4. Historical changes in human settlements in ancient Southern Etruria

Ancient Southern Etruria is one part of Italy located north of Rome and with an extension of around six thousand square kilometres. Towards the end of the second millennium B.C. there was a change in the human settlements in this area. It was a transition from a large number of small settlements – villages – to a small number of large settlements – proto-urban centers – with the progressive displacement of settlements to positions with higher defensive potential and better territorial control.

We have reproduced these historical phenomena in the computer. We have divided the entire area of ancient Southern Etruria into square cells and, based on actual geographical data, we have characterized each cell by two variables: presence of resources and defensive potential. The value of both variables varies between 0 and 1, where 0 means that there are no resources in that particular cell and the cell has geographical characteristics that make it impossible defend the cell from external attacks, while 1 means that there are a lot of resources in that cell and the cell can be easily defended from external attacks.

In the simulated ancient Southern Etruria there are a certain number of human settlements, each with a given number of inhabitants and a "zone of control" with a size which depends on the number of inhabitants and the presence of resources. The resources collected in the zone of control are divided by the number of inhabitants of the settlement to calculate the quantity of energy available to each inhabitant. Based on archaeological data, each settlement has a specific geographical location. At the beginning of the second millennium B.C. there are a large number of small settlements located in positions with limited or no defensive potential, while at the end of the millennium the number of settlements is drastically reduced and the settlements are located in positions of high defensive potential – summit plateau heights often completely isolated from the valley floor and defended by steep slopes. The simulation is a succession of cycles, with each cycle corresponding to two years. In each cycle, each settlement counts the number of its inhabitants, determines its zone of control, and computes the energy available to each of its inhabitants by dividing the resources collected in the zone of control by the number of inhabitants. And, based on these data, the settlement decides what to do.

What do the settlements do? For some settlements there is an increase in both the number of inhabitants and the size of its zone of control, which means that the energy available to each inhabitant remains the same or it increases. For other settlements the

number of inhabitants increases but the size of their zone of control remains the same because near their zone of control there is the zone of control of another settlement. For these settlements the situation is critical. The energy available to each of their inhabitants decreases and, therefore, the number of inhabitants also decreases until the settlement disappears. To avoid a complete collapse, a settlement has two options. Its inhabitants can migrate to a new location with more available resources or the settlement – perhaps by forming an alliance with other settlements – can make war to the other settlement to enter into possession of the area of control of the other settlement.

The result is that in the course of the simulation the number of settlements decreases and the average number of inhabitants in each settlement increases, and this corresponds to the historical data (see Figure 11).

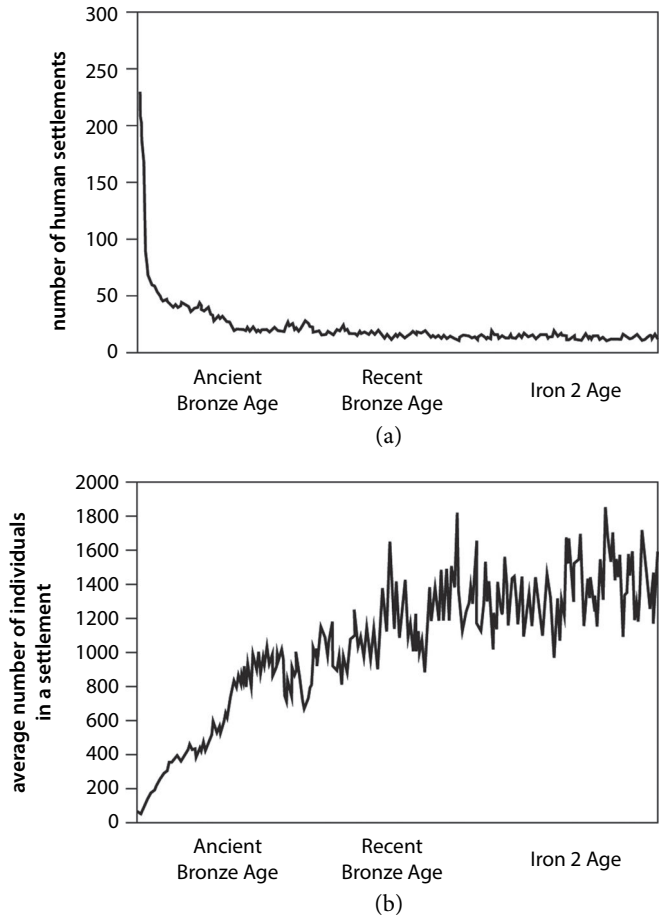


Figure 11. (a) Reduction in the number of settlements in ancient Southern Etruria. (b) Increase in the average number of inhabitants in a settlement.

Our simulation also reproduces the changes in the geographical location of human settlements in ancient Southern Etruria at the end of the second millennium B.C. As we know from archaeological data, there is a tendency for these settlements not only to decrease in number and to have a greater number of inhabitants but also to move to geographical locations with higher defensive potential. This is reproduced in our simulation (see Figure 12).

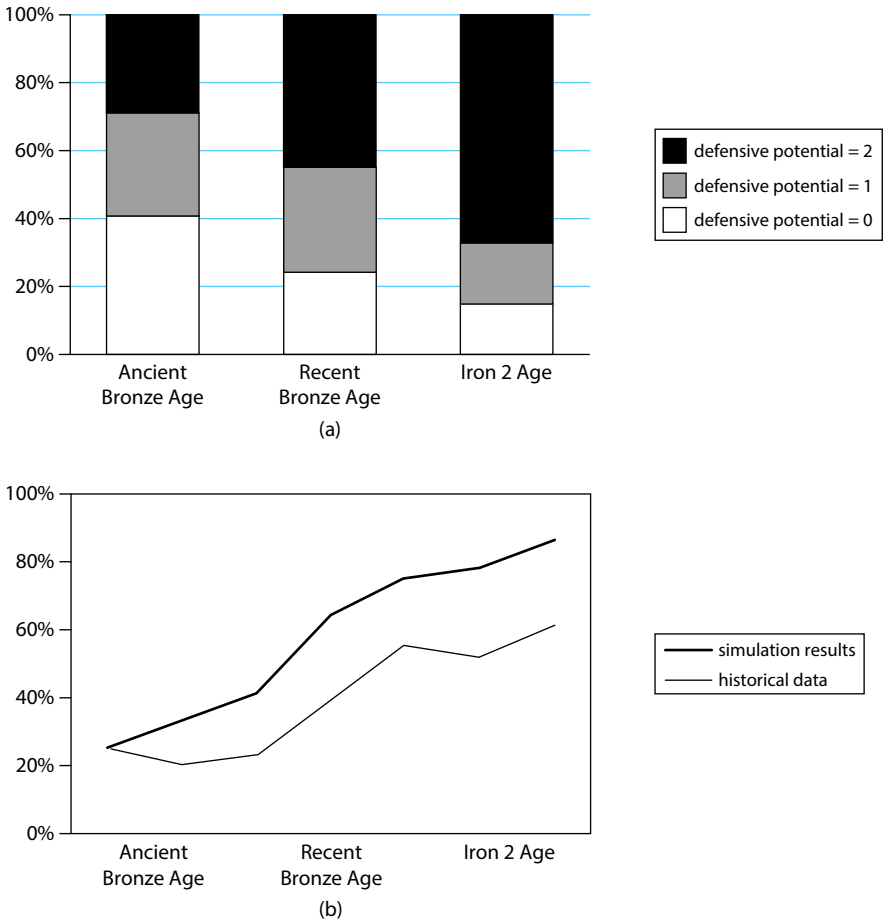


Figure 12. Increase in the defensive potential of human settlements from Ancient Bronze Age to Recent Bronze Age to Iron 2 Age (a) and comparison between historical data and simulation results (b).

What is interesting is that at the end of the second millennium the entire simulated Southern Etruria is subdivided into four large zones of control which are located around the four historical proto-urban centres of Orvieto, Tarquinia, Cerveteri, and

Veio of the real Southern Etruria (see Figure 13). This suggests that wars have been an important factor in the history of the human environment because the necessity for one centre to defend itself from attacks from other centre pushes the centre to move to locations with higher defensive potential.

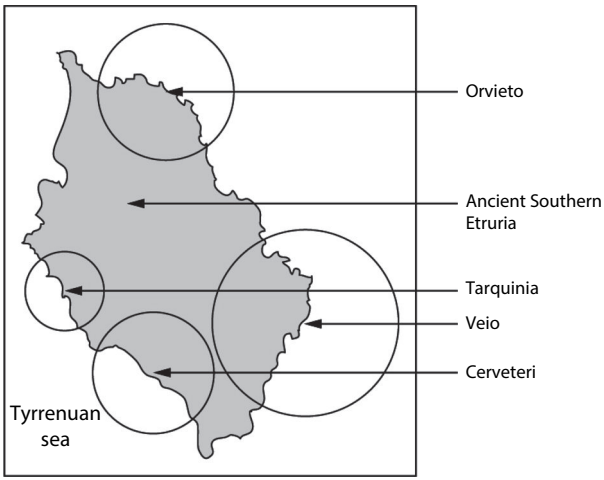


Figure 13. The four proto-urban centers in Ancient Southern Etruria at the end of the second millennium BC.

But this simulation is also interesting because it allows us to make an important methodological point about trying to understand reality by constructing artifacts that reproduce reality. In the simulation the tendency to locate settlements in areas with high defensive potential increases during the course of the second millennium – and this corresponds to the historical data. But in the simulation the increase is very fast – and this does not correspond to what has actually happened in that place at that time because the archaeological data tell us that only towards the end of the second millennium there was a rapid decrease in the number of human settlements and a rapid increase in the number of their inhabitants. Why is this interesting? As we have said in Chapter 1, the behaviors of an artifact are the predictions derived from the theory which has been used to construct the artifact – and these predictions must correspond to the actual empirical data. If we find that the artifact does not behave exactly like reality, what we must do is change the artifact by introducing other factors, increasing or decreasing the weight of the different factors, and in other ways, until the artifact behaves in ways that more closely match the real phenomena. If we succeed in doing so, we can be more confident that our artifact makes us really understand and explain reality. What is missing in our simulation of the historical changes in the human settlements in ancient Southern Etruria? We can test various hypotheses by changing our simulation until the simulation matches the historical data.

This applies to all the robots described in this book. The robots described in this book reproduce some very basic phenomena concerning the behavior of human beings and the functioning of human societies. But if what the robots do is different from what human beings do, we must not abandon the project of reproducing human beings with robots but we must change our robots until they behave exactly like human beings.

CHAPTER 11

Robotic economies

When ME looks at human beings, one thing that strikes ME is how many more things human beings want to have compared to nonhuman animals. Nonhuman animals only want what they need to remain alive and have offspring. Human beings want all sorts of goods and the number of goods they want has increased exponentially in the course of human history. The goods of nonhuman animals are limited to what already exist in nature while human beings produce always new goods and they create organizations of individuals that produce goods that no single individual would be able to produce by working alone. And human beings exchange their goods and, in this way, they can specialize in producing only one good because they can obtain all the other goods from other human beings.

The exchange of goods has many consequences for human societies. Some human beings – merchants – specialize in facilitating the exchange of goods by obtaining the goods from those who produce them and bringing these goods to those who want them. And human beings have invented money, a good which is not "used" like all the other goods but it is only exchanged in all exchanges and which facilitates the exchange of goods in that every human being always wants to have money because with money he or she can obtain (almost) all goods. They have also invented the borrowing of money which allows them to produce and to buy many more goods than they would be able to produce or buy if each individual or each organization of individuals could only use its money. But what particularly strikes ME is a more recent invention: marketing. The number of things that human beings want to have has enormously increased recently because sellers have developed all sorts of techniques for creating in buyers the motivation to buy the goods they sell.

1. Goods

All behaviour is intrinsically economic because all behaviour is concerned with goods. A good is anything which animals and human beings try with their behaviour to have. The goods of nonhuman animals are restricted to what is strictly necessary to their survival and reproduction – food, water, a sexual partner, protection from predators and other dangers. For human beings most of the things which are goods for them are only indirectly related to survival and reproduction and this explains why such a

large number of different things are goods for human beings compared to nonhuman animals. This is an important difference but our definition is completely general and purely behavioural: if an animal – nonhuman or human – tries with its behaviour to have X, X is a good for the animal, whatever X is and whatever the reasons that explain why X is a good for the animal. Animals and human beings are said to *need* or *want* the things that are goods for them, and we ourselves will use these words in some occasions. But our definition is based on behaviour and it does not make any reference to needs or wants.

When one talks of goods, what inevitably comes to mind is the science of economics because the science of economics is the science of goods. But economists are concerned with the goods of human beings and not with the goods of other animals, with goods that can be sold and bought and whose value can be measured in terms of money and not with other types of goods, and with modern market economies and not with other types of human economies. Understanding the economic behaviour of human beings requires a general science of goods which applies to both nonhuman animals and human beings, to all sorts of goods, and to all types of human economies. Human beings are different from nonhuman animals because nonhuman animals obtain their goods from nature while human beings obtain most of their goods from other human beings through the exchange of goods. Therefore, unlike the economy of nonhuman animals, the economy of human beings is a social economy and human societies are complex organizations for regulating the exchange of goods. But if we want to understand human economies, we must adopt a larger perspective than the perspective of the science of economics.

The idea of a robotic economy sounds completely unfamiliar and it does not make much sense for today's robots because nothing is a good for today's robots. Today's robots do not behave as they behave because with their behaviour they can have this or that thing. They behave as they behave because *we* decide how they should behave. Which things are goods for an animal or for a human being – which things they try with their behaviour to have – depends on their motivations, and today's robots do not have motivations (see Chapter 2). They do what they do to satisfy *our* motivations. Therefore, today's robots cannot have goods and cannot have economies, and the construction of robotic economies is a task for the future. In this Chapter we begin to address this task. We begin with robots which obtain their goods from nature and, therefore, are more like nonhuman animals than human beings. And then we move to human economic phenomena: the exchange of goods, specialization in the production of different goods, the emergence of money and the price of goods, workers and private enterprises, financial economies and marketing (not market) economies. A general theme of the chapter is to understand the enormous increase in the number of things which, today, are goods for human beings compared to nonhuman animals.

2. Value

As we have said, if an individual tries with its behaviour to have X, X is a good for the individual. Now we say: if X is a good for an individual, X has *value* for the individual. What is added by saying that a good has value for an individual? The answer is that value can be measured. It can be measured in a comparative sense: two things are both goods for an individual but one good has more value than the other good for the individual. And the value of a good can be measured in an absolute sense if there is a measurement tool – money. Money presupposes the exchange of goods because the absolute value of a good is the number of money units which an individual is disposed to give to another individual to obtain the good from the other individual. We describe robots that exchange goods and develop money later in this Chapter. The robots of this section do not obtain their goods from other robots but they obtain their goods from nature – but these goods have different values for them.

To measure the value of a good with respect to another good, we look at the robots' preferences. If a robot can choose between good X and good Y and the robot chooses good X rather than good Y, good X has more value for the robot than good Y. We evolve a population of robots in an environment in which there are two types of food tokens – grey and black – and the two types of food tokens do not contain the same quantity of energy. The grey tokens contain one unit of energy and the black tokens contain two units of energy (see Figure 1).

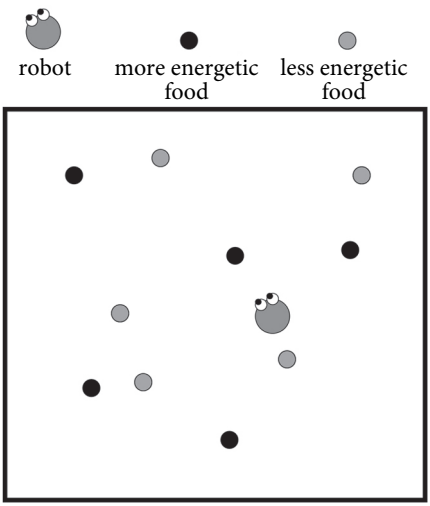


Figure 1. The robots live in an environment with black and grey food tokens. The black tokens contain two unit of energy while the grey tokens contain only one of energy.

At the end of evolution, we find that the robots eat more black tokens than grey tokens, and this means that they prefer the black tokens to the grey tokens. But to have a better measure of the robots' preferences, we look at how the robots behave in the experimental laboratory. In the experimental laboratory a robot is exposed to one black token and one grey token which are at the same distance from the robot and we find that the robot approaches and reaches the black token rather than the grey token. This is a better measure of the robots' preferences for the following reason. The environment of the robots contains the same number of grey and black tokens but, if the grey tokens were more numerous than the black tokens, during their life the robots would probably eat more grey tokens than black tokens. However, our prediction is that in the experimental laboratory they will still prefer the black tokens to the grey tokens.

But the value of a token is not only determined by the quantity of energy contained in the token. Saying that X is a good for an individual is the same as saying that the individual has the motivation to have X but, as we have seen in Chapter 2, the strength of a motivation depends on the strength of the individual's other motivations. Therefore, the value of a good for an individual depends not only on the strength of the motivation to have X but also on the strength of the individual's other motivations. Another motivation of our robots is the motivation to save their time. As we have seen in Chapter 2, if the robots need to both eat and drink to remain alive and they have the same level of hunger and thirst, in the experimental laboratory they prefer to go to the food token rather than to the water token if the food token is nearer to them than the water token because going to the nearer token allows them to save their time. We continue these experiments here. We evolve a population of robots in an environment in which the black tokens and the grey tokens contain the *same* quantity of energy – one unit – and, as expected, we find that in their natural environment the robots eat more or less the same number of black and grey token. However, in the experimental laboratory they prefer a grey token to a black token if the grey token is nearer than the black token, and vice versa if the black token is nearer than the grey token. And the preference becomes progressively stronger if we keep one token always at the same distance from the robot and, in different trials, we progressively increase the distance of the other token (see Figure 2).

These robots are interesting because they demonstrate the generality of the notion of good. For the robots not only the food tokens are goods but also their time is a good ("time is money"). They behave in ways that allow them to save their time – and this, according to our definition, makes time a good. The robots prefer to go to the tokens which are nearer to them because this allows them to save their time which they can use to eat other tokens. Time viewed as a good reminds us that a good is not only something which a robot wants to have but also something which a robot does not want to lose. A robot can do something that causes the robot to have X but it can also do something that allows the robot not to cease to have X. In both cases, X is a good for the robot.

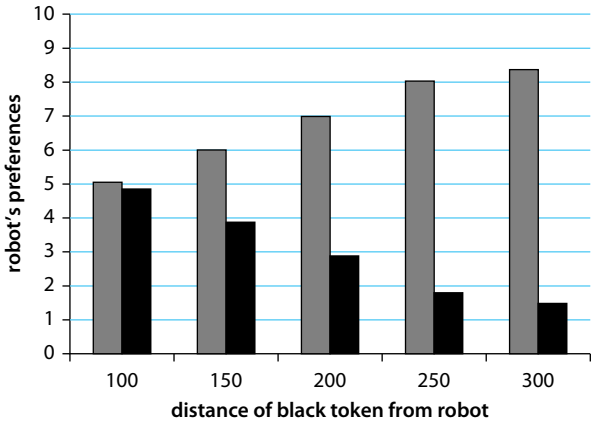


Figure 2. If the black and grey tokens contain the same quantity of energy, the robots show no preferences if the two tokens are at the same distance from them. But they prefer the grey token to the black token as the distance of the black token is progressively greater.

In determining the value of a good for an individual the different motivations of the individual interact together and, to see how they interact together, we return to the robots that live in the environment in which the black tokens contain two units of energy and the grey tokens only one unit of energy. These robots also have the motivation to save their time and this is shown by the fact that, in the experimental laboratory, the robots' preference for the more energetic black token slightly decreases if we progressively increase the distance of *both* tokens from the robot and, therefore, the time needed to reach the food tokens (see Figure 3).

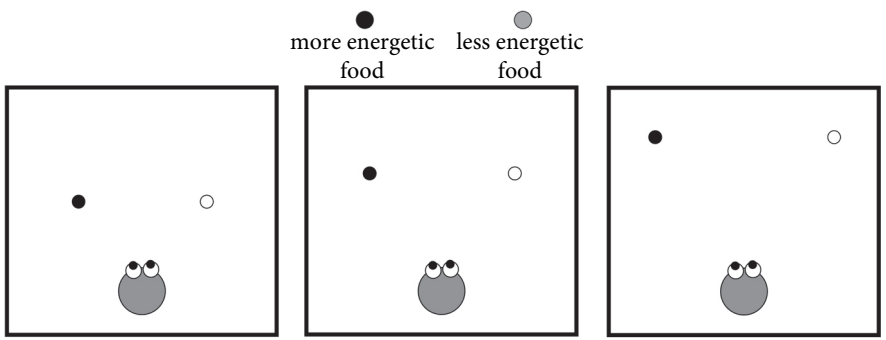


Figure 3. The more energetic black token and the less energetic grey token are at the same distance from the robot but in different trials we progressively increase this distance.

The increase in the time needed to reach the two types of tokens reduces the importance of the different quantity of energy contained in the tokens in determining their value for the robot (see Figure 4).

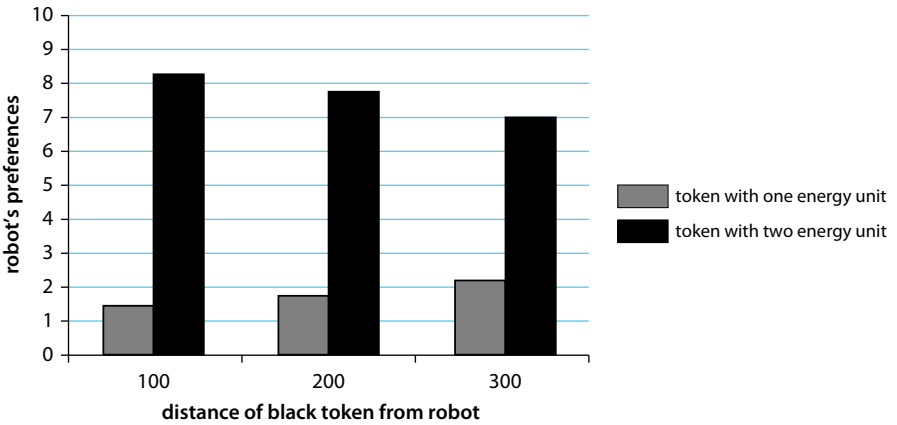


Figure 4. The robots clearly prefer the more energetic black token to the less energetic grey token but, if we increase the distance of both tokens from the robot, the preference for the more energetic black token is progressively reduced.

The importance of the motivation to save one's time is shown by another experiment in which the less energetic grey token is always at the same distance from the robot but, in different trials, we progressively increase the distance of the more energetic black token. The robots prefer the black token even if it is more distant than the grey token but this is only true if the distance of the black token does not become too great. If the black token is very distant from the robot, the preference for the black token disappears and the robot goes half of the times to the more energetic black token and half of the times to the less energetic grey token. (This is called the "equilibrium point".) If we further increase the distance of the black token, the preferences of the robots are inverted and the robots prefer to go to the less energetic grey token than to the more energetic black token (see Figure 5).

Another interesting problem is if it is the *relative* difference or the *absolute* difference in the energy contained in the black tokens and in the grey tokens which is more important in determining the value of the tokens. We compare two populations of robots that live in two different environments. One environment is like the preceding one: the grey tokens contain *one* unit of energy and the black tokens *two* units of energy. In the other environment the grey tokens contain *three* units of energy and the black tokens *six* units of energy. The relative difference is the same – in both cases the black tokens contain *twice* the energy contained in the grey tokens – but the absolute difference is different: it is one unit (2 – 1 = 1) in one environment and three units (6 – 3 = 3)

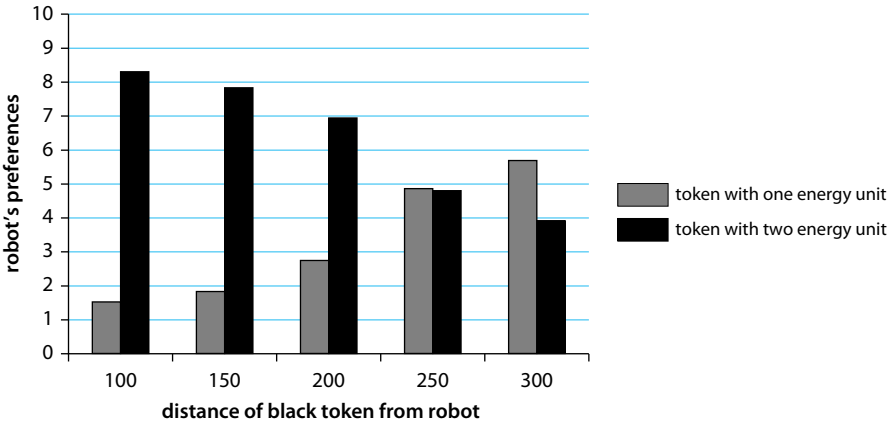


Figure 5. If the more energetic black token is not too distant from the robot compared to the less energetic grey token, the robot prefers the black token. But when the distance of the black token becomes too great (300 pixels), the robot prefers the grey token.

in the other environment. The robots that live in the second environment live longer than the robots that live in the first environment because in their environment the total quantity of energy is greater than in the environment of the other robots. But, although they live longer, the robots that live in the second environment eat *fewer* tokens than the robots that live in the first environment. Since their only problem is to remain alive, the robots which live in the richer environment can survive by eating fewer food tokens.

The robots described so far are like nonhuman animals because with their behaviour they try to have only what they need to remain alive and have offspring. Human beings – especially, today's human beings – are different because, for most of them, the problem is not to remain alive and have offspring but to have as many goods as possible and to have more goods than other human beings. Our next robots are more like human beings. The difference is in the criterion of fitness. The criterion of fitness of the preceding robots was the length of their life – longer life, more offspring. For the new robots life has the same length for all robots and the robots that reproduce are those that accumulate more energy during a life which has the same length for all robots.

With these more human-like robots we ask the same question: What is more important, in determining the value of the tokens, the relative or the absolute difference in the energy contained in them? To answer this question, we compare our robots for which the grey tokens contain *one* unit of energy and the black tokens contain *two* units of energy with a new population of robots that lives in an environment in which the grey tokens contain *five* units of energy and the black tokens contain *six* units of energy. The absolute difference is the same – one unit ($2 - 1 = 1$; $6 - 5 = 1$) – but the relative difference is much smaller. For the first type of robots, it is 100% (2 is 100% more than 1) while, for the new robots, it is only 20% (6 is only 20% more than 5). So, what is the

answer to our question? The answer is that the value of two types of tokens depends on the *relative* difference rather than on the *absolute* difference in the quantity of energy. If eating the black tokens gives the robots only 20% more energy than eating the grey tokens, the black tokens are still preferred to the grey tokens but the preference is not very strong. This is interesting because it shows that the robots' preferences depend on their overall adaptive pattern. For these robots, to generate more offspring than other robots a robot must not eat *much* but it must eat *more* than the other robots. So their preference for the black tokens is stronger if the black tokens provide them with 100% rather than only 20% more energy than the grey tokens.

Other factors can influence the robots' preferences and, therefore, the value of the tokens for the robots. One is the total number of the tokens contained in the environment. The environment of the robots described so far contains ten tokens: five grey tokens and five black tokens. We evolve another population of robots in an environment that contains twenty tokens: ten grey tokens and ten black tokens. What are the consequences for the robots' preferences of this greater abundance of food? Contrary to our expectations, the robots' preference for the more energetic black tokens become stronger, and this is true both if count the number of grey and black tokens eaten by the robots in their natural environment and if we test the robots in the experimental laboratory. The reason is that, when food is more abundant, the robots can be choosier in their behaviour. If there is little food in the environment, the robots tend to eat any type of food and their preference for the more energetic black tokens cannot be too strong. If the environment is richer in food, they can choose which food to eat and they have a stronger preference for the black tokens. This may also be true for the robots for which the criterion of fitness is the length of their life – animal robots – but it is typical of robots whose problem is to accumulate more goods than other robots – human robots.

Another good whose value can be measured is avoidance of risk. Avoidance of risk is a good because both nonhuman animals and human beings behave in ways that allow them to avoid risk. Our next robots tend to avoid risk. We evolve a population of robots in an environment in which the less energetic grey tokens always contain one unit of energy while some black tokens contain two units of energy and other black tokens do not contain any energy – and the robot cannot know this in advance because the black tokens are all visually identical. If a robot eats a grey token, it is sure to get one unit of energy but, if it eats a black token, it may get two units of energy or zero energy (see Figure 6).

What do the robots do? Do the robots prefer the less energetic but riskless grey tokens or the more energetic but risky black tokens? If 50% of the black tokens contain two units of energy and 50% no energy, the robots do not have preferences. They eat more or less the same number of grey and black tokens and, when we bring them into the experimental laboratory, they go to the grey token or to the black token with more or less the same probability – provided the two tokens are at the same distance from the robot. But the situation is different if we vary the probability that the black tokens

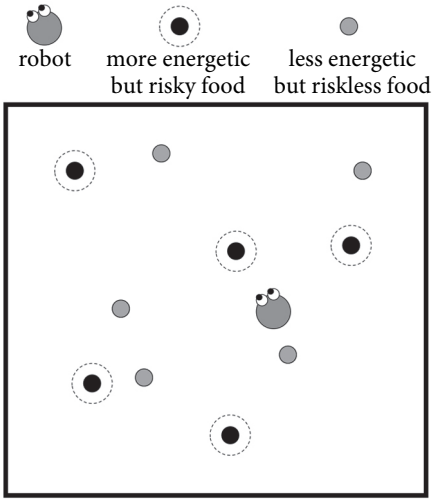


Figure 6. The robots live in an environment with riskless grey food tokens that always contain one unit of energy and risky black food tokens that can contain two units of energy but also no energy.

contain no energy. If 75% of the black tokens contain two units of energy and only 25% do not contain any energy, the robots prefer the more energetic black tokens to the less energetic grey tokens because the black tokens are not so risky. But if the situation is inverted and the probability of obtaining zero energy by eating a black token is 75% (high risk), they prefer the less energetic grey tokens. The robots' behaviour is sensitive to the probabilistic structure of the environment.

Avoidance of risk is a good because the robots tend to avoid risk and, as for any other good, the value of this particular good depends on a variety of factors. We have already seen one of these factors: the probability that the bad event will actually occur. The robots are disposed to risk to get zero energy if the risk is only 50% but not if the risk is higher. Another factor is the quantity of energy contained in the risky black tokens. If the grey tokens contain one unit of energy and the black tokens contain not *two* but *ten* units of energy, in the experimental laboratory the robots prefer the black tokens to the grey tokens even if the probability that a black token contains no energy is 75%. If the reward is very high, our robots are ready to do things which are very risky – and this reminds us of today's financial economy.

But, interestingly, risk can also increase the strength of an individual's motivations and, therefore, the value of a good. This is what we show with the next robots. The behaviour of these robots is not evolved in a succession of generations of robots but it is learned in an experimental laboratory. We put a robot at the beginning of a corridor at the end of which there is a food token but the corridor is a long one and the robot can see the food token only when it is near to the food token. Using another learning

algorithm – the reinforcement learning algorithm – we make the robots learn to run through the corridor to reach the food token. At the beginning the robots do not move or they move very hesitatingly back and forth along the corridor. At the end they go straight through the corridor, reach the end of the corridor, and eat the food token. But we make the robots learn in two different conditions. In one condition the robot always finds a food token at the end of the corridor (see Figure 7).



Figure 7. A robot is placed at the beginning of a corridor at the end of which in all experimental trials there is a food token. The robot must reach the food token but it can see the food token only when it is very near to it.

In the second environment, half of the times at the end of the corridor there is a food token while in the other half there is nothing, and the succession of two different types of trials is random. Therefore, in this second environment, the robots risk not to find any food token at the end of the corridor (see Figure 8).



Figure 8. In half of the experimental trials at the end of the corridor there is a food token (a) and in the other half there is nothing (b). The two types of trials follow each other unpredictably.

If we measure the strength of a motivation as the time taken to satisfy the motivation, we find that the robots take less time to run through the corridor if there is only a probability of 50% to find a food token at the end of the corridor than if the probability is 100%. We conclude that risk increases the robots' motivation to eat. These results replicate the behaviour of real mice in the experimental laboratory. The only difference between mice and human beings is that mice die if they try to avoid risk and do not go to the end of the corridor while human beings do even more risky things because they generally do not die if they fail.

We end this Section with four general comments on goods and their value. Our first comment is a methodological one and it has to do with the relation between

behaviour in the experimental laboratory and behaviour in the natural environment. Psychologists tend to assume that human beings behave in the same way in the experimental laboratory and in real life and they believe that what they find in the experimental laboratory can be extrapolated to real life. This may be true for very simple behaviours and very basic capacities but it may not be true for most behaviours. Behaviour depends on the environment. The natural environment and the experimental laboratory are two different environments and we cannot expect that human beings behave in the same way in their natural environment and in the experimental laboratory – especially because, unlike the experimental laboratory, their natural environment is a social environment. (The male and female robots described in Chapter 7 do not behave in the same way in the natural environment and in the experimental laboratory.) This is an important point to make when we talk of goods and the value of goods because economists and other social scientists have recently adopted the experimental method to study human behaviour and the value of goods, and what they find in the laboratory may not correspond to what happens in the real world. Laboratory experiments – with both human beings and robots – are useful but one must be aware of their limitations.

Our second comment is on inter-individual differences in the value of goods. All robots that live in the same environment tend to have the same preferences but, since no two robots are identical, there are inter-individual differences in the preferences of the robots. One robot gives more value to time compared to another robot, and we see this because in the experimental laboratory the first robot tends to always approach the nearer food token while for the second robot the tokens' distance does not make much of a difference. One robot is more sensitive to the number of grey and black tokens contained in the environment while another robot has more or less the same preferences independently of the number of grey and black tokens. In Chapter 7 we have seen that male robots tend to exhibit risky behaviours more than female robots, and this implies that the differences among the robots in the value of goods – in this case, avoidance of risk – can be not only inter-individual differences but also sexual differences. (On robots that are inter-individually different, see Chapter 12.)

Our third comment concerns what our robots with their few goods can tell us about human beings with their many goods. We make our robots evolve by using two different criteria of fitness – the length of their life or the quantity of energy they are able to collect during a life which has the same length for all robots – and, as we have said, the first type of robots resemble nonhuman animals while the second type of robots resemble human beings. Nonhuman animals only want what they need to remain alive and reproduce. Human beings want to have *more* goods than other human beings. Competition in nonhuman animals is biological (who has more offspring?) while competition among human beings is almost entirely social. Human beings want to have more goods than other human beings, and this is why human beings want always new goods. If we interpret the food tokens as money, human beings are like the robots

which reproduce based on the total quantity of energy they are able to accumulate during their life and, therefore, the robots have an "infinite" desire for energy. In modern economies people have an "infinite" desire for money because they can get all sorts of goods in exchange for money (We describe robots that exchange goods and develop money later in this chapter.)

Our fourth and final comment is that something can become a good for an individual which initially was not a good for the individual. The behaviour of our robots is entirely determined by their genes. What is a good and what is the value of the good for them are both dependent on purely genetic factors and do not change during a robot's life. This is a limitation of our robots because most of the things which are goods for human beings are not originally goods for them but they become goods because of the experiences they have during their life. We have already seen this in Chapter 3 dedicated to robots that both evolve and learn. In the natural environment the robots evolve the behaviour of approaching and eating the black tokens because the black tokens contain energy and, therefore, the black tokens are originally (genetically) goods for them. Then we bring the robots into an experimental laboratory and we expose them to a grey token which they have never seen before and which does not contain any energy. Initially, the robots ignore the grey token and, therefore, the grey token is *not* a good for them. But if, when they touch the grey token, a black food token makes its appearance, they progressively learn to approach and touch the grey token. The grey token has become a good for them. This reminds us that a crucial component of today's human economies is marketing, which is the use of all sorts of methods to make us learn to treat as a good what was not a good for us.

3. Specialization and the exchange of goods

The economy of the robots described in the preceding section is an individual economy. Each robot lives alone in its environment and we can talk of a robotic economy only in the sense that there are things which are goods for the robots and these goods have a value that can be measured by looking at the robots' preferences. Nonhuman animals have an individual economy but the economy of human beings is a *social* economy based on the exchange of goods. Nonhuman animals obtain all their goods from nature. Human beings obtain almost all their goods from other human beings through the exchange of goods. One individual gives its goods to another individual and, in exchange, the second individual gives its goods to the first individual. Some nonhuman animals give their goods to other individuals but this generally only happens if the other individual is genetically related to the giver, for example, is the giver's offspring. And there is no immediate return: the other individual does not give anything to the first individual in exchange for the good it receives from the first individual. This type of giving also occurs between genetically related human beings (see Chapter 7

on robotic families) but what makes human beings unique is that human beings obtain most of their goods from individuals which are not their kin – in exchange for the goods that they give to these other individuals.

This is the economic exchange of goods. Human beings have another – non-economic – form of exchange: reciprocity. An individual gives his/her goods to another individual in the hope to receive some good from the other individual in some future occasion. Reciprocity is different from economic exchange. Reciprocity is based on the hope of return, and what is given in return is not specified in advance. In the economic exchange of goods there is no need for hoping for a return because goods are exchanged simultaneously and the goods involved in the exchanged are well-defined and completely explicit. And there is another difference. In reciprocity one individual gives his/her goods to another individual in the hope that not necessarily that particular individual but other individuals will reciprocate because of his/her reputation as a "giver". The economic exchange of goods is different because it is between two specific individuals and it has little place for one's reputation as "giver". The exchange of goods of the robots described in this Chapter is an economic exchange, not an exchange based on reciprocity.

The exchange of goods is made possible by the existence of external stores. A robot gives its goods to another robot only if it does not need to immediately use its goods and, therefore, it keeps the goods in its external store for future use. Consider the robots that must eat two different types of food to remain alive. If a robot immediately eats the food it finds in the environment – as most nonhuman animals do – the robot cannot exchange one type of food with the other type of food with another robot because it has no food to give to the other robot. This is only possible if both robots have an external store into which they put their food so as to be able to exchange one type of food with the other type of food with the other robot. In Chapter 9 we have said that external stores make us human. One reason why external stores make us human is that they make the exchange of goods possible. And the exchange of goods is very important for human beings not only from an economic point of view but also because it requires social contact and social interaction. For example, when people of different cultures exchange their goods, they do not only exchange goods but they also influence each other culturally. (For robots that have culture, see Chapter 8.)

But external stores by themselves do not explain the exchange of goods. Another pre-condition is specialization. Specialization means that one robot specializes in finding one type of food and another robot specializes in finding another type of food but both robots can eat both types of food because they exchange one type of food for the other type of food. If both robots were to look for both types of food in the environment, there would be no need for exchange. The exchange of goods makes sense only if one robot has more food of one type in its external store because it is specialized in collecting that type of food and the other robot has more food of the other type in its external store because it is specialized in collecting the other type of food. External

stores, the exchange of goods, and specialization are three human adaptations which are strictly linked together and each represents a pressure for the emergence of the others. The exchange of goods presupposes the existence of external stores and specialization. External stores and specialization emerge because they make the exchange of goods possible. This is why all three human adaptations have probably evolved together.

Our next robots reproduce specialization and the exchange of goods. The robots live in an environment with two types of food – black and white food tokens – and the robots need to eat both types of food to remain alive. The environment is a seasonal one. The tokens collected by the robots in the environment are not immediately replaced by new tokens and new tokens only reappear at the beginning of the next season. A robot generates one offspring at fixed intervals and, therefore, the fitness of a robot is the length of its life: the robots that live longer leave more copies of their genes to the next generation. At the beginning the population is very small because the robots are not very good at collecting the food tokens and, therefore, they have a short life and generate few offspring. Then, the selective reproduction of the best robots and the addition of random mutations to the inherited connection weights lead to better robots. Population size increases until it reaches a stable value which depends on the characteristics of the environment – how much food exists in the environment and how long are the seasons. If food is abundant and seasons are short, population size is greater than if food is scarce and seasons are long.

But another factor, beyond the abundance of food and the length of seasons, is how the food tokens are distributed in the environment. We compare two populations of robots. One population lives in an environment in which the black tokens and the white tokens can be found in all parts of the environment. The other population lives in an environment in which the black tokens are in one zone of the environment and the white tokens in another zone, and the two zones are separated by an empty space (see Figure 9).

At the end of evolution, we find that population size is greater for the robots that live in the homogeneous environment than for the robots that live in the environment with the two food zones. If the black tokens and the white tokens are in two separates zones, the robots risk dying when they try to traverse the empty space between the two zones. And, if the environment is really difficult, with little overall food, long seasons, and a very large empty space between the two food zones, the robots become extinct because they die so early in their life that they are unable to generate offspring. The robots that live in the environment in which the two types of food can be found everywhere do not have these problems and for them it is easier to remain alive and have offspring.

These robots do not make exchanges. Like the robots of the preceding section, they have an individual economy. Each robot directly collects the food it must eat to remain alive and, even if the robots have an external store which helps them to survive

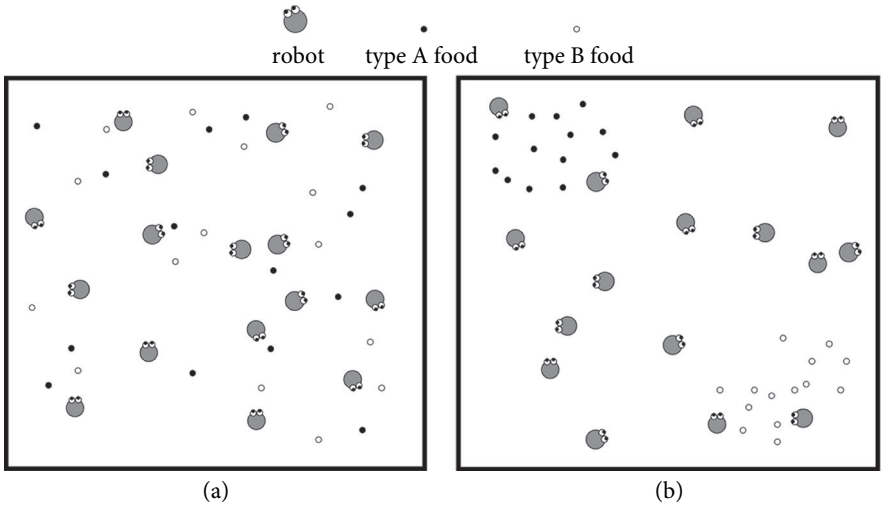


Figure 9. In one environment the black tokens and the white tokens are distributed in the entire environment (a). In the other environment, the black tokens and the white tokens are in two different zones separated by an empty space (b).

when traversing the empty space between the two food zones, life is very difficult for these robots.

Now we construct another population of robots which are like the preceding robots but, unlike the preceding robots, they exchange one type of food for the other type of food with other robots. In each cycle, we randomly select two robots and, if their external stores are complementary, the robots make an exchange (see Figure 10).

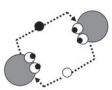


Figure 10. Two robots exchange one type of food for another type of food.

Having complementary external stores means that the external store of one robot contains more black tokens than white tokens and the external store of the other robot contains more black tokens than white tokens. Making an exchange means that one black token is transferred from the store of the first robot to the store of the second robot and one white token from the store of the second robot to the store of the first robot. (We "hardwire" the robots to make an exchange if their stores are complementary. For robots that autonomously decide if and what to exchange, see later in this section.)

What are the consequences of exchanging one type of food for the other type of food? The robots need to eat both the black tokens and the white tokens to remain alive and, therefore, if a robot has only one of the two types of food in its external store, without the exchange of food the robot risks dying. On the contrary, if the robots exchange their food, a robot may remain alive because it can give one token of its type of food to another robot and obtain one token of the other type of food from the other robot. Our expectation is that the robots that exchange their food will have a larger population size and will be protected from the risk of extinction if the distance between the two food zones is very great – and, in fact, this is what we find. At the end, the population of robots which exchange their food is made of 300 individuals, while the robots that do not make exchanges are only 200. In Chapter 9 we have called external stores a "safety net" because, when the robots do not find food in the environment, they can eat the food contained in their external store. The exchange of goods is another and much more effective "safety net". If a robot has only one type of food in its store and, therefore, it risks dying, it can obtain the other type of food from another robot in exchange for its food. Notice that a robot gives a black token to another robot in exchange for a white token (or vice versa) not because it wants that the other robot remains alive but because *it* wants to remain alive. The exchange of goods is a beneficent "invisible hand" that allows both robots to remain alive.

Now we ask another question: Does the exchange of goods lead to specialization? Do we find that some robots specialize in collecting the black tokens and other robots specialize in collecting the white tokens? Specialization is impossible without exchange. The robots must eat both types of food and, therefore, if there are no exchanges a robot cannot specialize in collecting the black token and ignore the white tokens, or vice versa. But does specialization emerge if the robots have the possibility to exchange one type of food for the other type of food?

The robots that live in the environment in which the two types of food can be found everywhere survive both if they exchange and if they do not exchange their food and we do not find any form of specialization among these robots. A robot can have more black tokens than white tokens in its external store, or vice versa, but no robot is really specialized in collecting only one type of food. The situation is different for the robots that live in the environment in which the black tokens are in one zone and the white tokens are in another zone of the environment and there is little food of any type in the space between the two zones. As we have seen, the possibility to exchange the food of one type for the food of the other type is very important in this environment. Therefore, for the robots which do not make exchanges, the situation is a very difficult one. The robots need to continually move from one zone to the other zone and they risk dying while they are traversing the space between the two zones. This is not true for the robots that exchange one type of food for the other type of food. These robots are able to survive because an individual that lacks one type of food can obtain that type of food from another robot. Is the exchange of food accompanied by specialization?

The answer is Yes. Some robots spontaneously specialize in collecting the black tokens and other robots specialize in collecting the white tokens and then they exchange one type of tokens for the other type of tokens (see Figure 11).

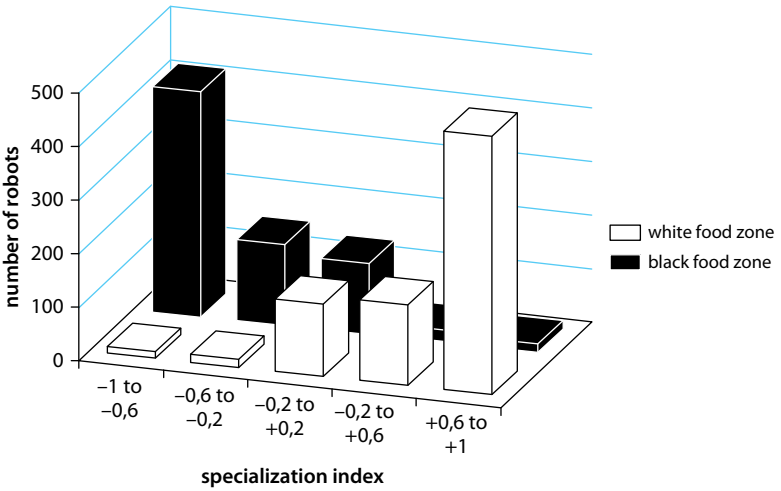


Figure 11. Half robots specialize in collecting the black tokens (specialization index = −1) which are in one zone of the environment and half robots specialize in collecting the white tokens (specialization index = + 1) which are in another zone.

In principle, specialization could be genetic or environmental. Genetic specialization means that a robot is born with a genetically inherited tendency to look for one type of food and to ignore the other type of food. Environmental specialization means that a robot which happens to be born in the zone of the environment which contains one type of food specializes in collecting that type of food even if it does not have any genetically inherited specialization for collecting one or the other type of food. The specialization of our robots is environmental, not genetic, and this is shown by the fact that if we bring the robots into an experimental laboratory which contains one black token and one white token, the robots do not exhibit any specialization or preference but they go indifferently to the black token or to the white token. Therefore, specialization is not encoded in their genes. And, in fact, in human beings specialization is mainly due to learning. Producing one type of food requires specific skills and these skills are learned, not genetically inherited – although, as in all sorts of learning and all sorts of behaviours, the inherited genes can make a difference. Since it is impossible for an individual to acquire all the skills which are necessary to produce all the goods which the individual wants to have, the individual specializes in producing only one type of good and he or she obtains the other goods from other individuals.

The robots we have described live in an environment with two different types of food and the robots need to eat both types of food to remain alive. We now construct a new type of robots which live in an environment in which there is only one type of food and the robots need to eat only this type of food to remain alive. But the environment of these robots also contains a certain number of tool tokens which allow the robots to extract more energy from the food tokens. Eating a food token increases the robot's bodily energy by one unit and reaching (capturing) a tool token does not by itself increase the robot's bodily energy. But, if when the robot reaches a food token, the robot also has a tool token, the robot extracts not one unit but three units of energy from the food token by using the tool token. Therefore, the tool tokens are useful to the robots and we expect that the robots will evolve the behaviour of collecting them. (Like the food-producing tools of the robots of Chapter 7, these tool tokens can be used only once and then they must be thrown away.) These tool tokens resemble the vases of the robots described in Chapter 8. The only difference is that the vases are constructed by the robots by imitating the existing vases, while these tool tokens are found in the environment like the food tokens.

Like the preceding robots, we let the robots evolve in two different environments. In one environment the food tokens and the tool tokens can be found in the entire environment while in the other environment the food tokens are in one zone and the tool tokens in another zone (see Figure 12).

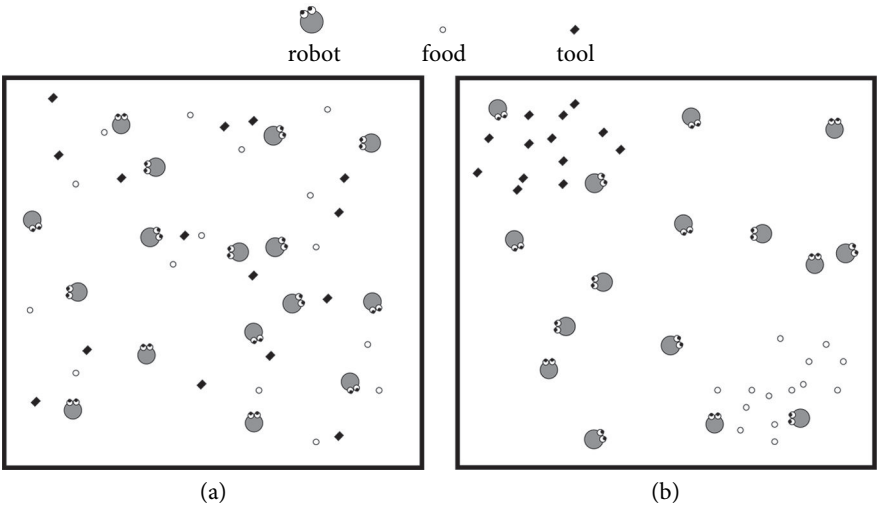


Figure 12. The robots live in an environment which contains both food tokens and tool tokens. In one environment the food and tool tokens can be found everywhere (a) while in another environment they are located in two zones separated by an empty space (b).

In each of the two environments we evolve two different populations of robots. The robots of one population do not make exchanges while the robots of the other population exchange food for tools. The results have many similarities with those of the preceding robots. In the environment in which food and tools are randomly distributed in the entire environment, population size is larger for the robots that exchange food for tools but even the robots that do not make exchanges are able to survive because they can limit themselves to eating the food tokens without collecting the tool tokens. This is also true for the robots that live in the environment in which the food tokens and the tool tokens are in two separate zones but these robots tell us something new and interesting. As always, we replicate the simulation many times and we find that in some replications most of the robots colonize the food zone and are not interested in the tools – they are all peasants – but in other replications some robots live in the food zone – peasants – and other robots live in the tool zone – artisans – and peasants and artisans exchange food for tools. These last robots have the largest population size – and this explains why specialization and the exchange of goods are useful adaptations (see Figure 13).

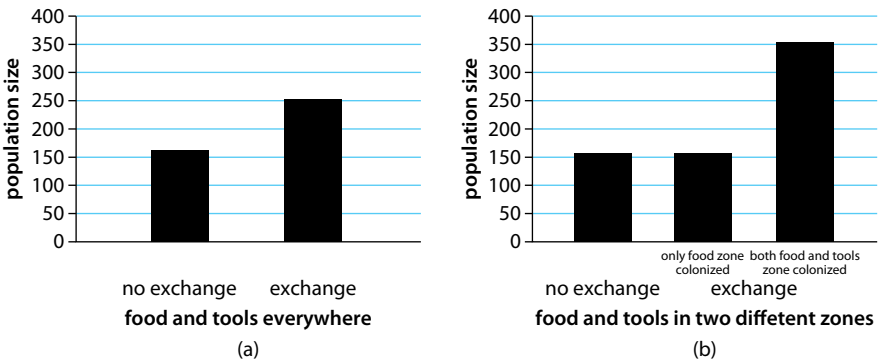


Figure 13. Population size of robots that do not exchange or do exchange food for tools in an environment in which the food and tool tokens can be found everywhere (a) and in an environment in which the food and tool tokens are in two different zones (b).

We now return to the robots that exchange one type of food for the other type of food. In the preceding robots, the exchanges were hardwired by us. In each cycle two robots are randomly selected from the population and, if their external stores are complementary – one robot has more black tokens than white tokens and the opposite is true for the other robot – an exchange between the two robots automatically takes place. But human robots should be able to make exchanges without our intervention. They should be able to autonomously decide which token to offer in exchange for the other token and they should decide whether to accept or reject an exchange proposal. Our next robots evolve this capacity.

The new robots are more abstract than the preceding robots. They do not have a body and they do not live in an environment, and for these robots the procurement of food is implemented in a very abstract way. The robots must eat both the black and the white tokens because the black and the white tokens contain two different types of energy which are both needed by the robots to remain alive. In each cycle one black token or one white token, randomly chosen, is added to the external store of all robots and the robots eat one or the other type of food, also randomly chosen. Even if the procurement of food is abstract and hardwired by us and the robots are not specialized in procuring one or the other type of food, for these robots also it is useful to exchange their food because the exchange of food may protect them from the risk of having zero tokens of one or the other type in their external store and, as a consequence, they may die. What is new is that these robots autonomously propose an exchange to another robot (a black token for a white token, or vice versa) and the other robot autonomously accepts or rejects the exchange proposed. The behaviour of procuring food is abstract and hardwired by us but the behaviour of proposing an exchange and accepting or refusing an exchange is under the control of the robot's neural network.

The neural network of the robots is made of two modules. One module controls the behaviour of the robot when the robot proposes an exchange to another robot. The other module controls the behaviour of the same robot when the robot receives an exchange proposal from another robot and it must decide whether to accept or refuse the exchange proposed. In both cases the robot needs to know how many black and white tokens are contained in that moment in its external store and, therefore, the sensory neurons of both modules encode the number of black and white tokens which are currently contained in the robot's external store. If a robot is (randomly) assigned the role of proposer of the exchange, the robot uses the first module of its neural network and the output neurons of this module specify which of the two tokens the robot is asking from the other robot in exchange for the other token. On the contrary, if the robot is assigned the role of receiver of the exchange, the robot uses its second module. The sensory neurons of the second module encode not only the current content of the robot's external store – like the first module – but also which token the *other robot* is asking for the exchange, while the output neurons encode the decision to accept or reject the exchange (see Figure 14).

The robots of the initial generation have random connections weights in both modules of their neural network and, therefore, the robots are unable to make the appropriate exchange proposals to other robots and to accept or reject appropriately the exchanges proposed by other robots. The robots may ask for a black token even if they have more black than white tokens in their external store and they may accept to give a black token even if they have fewer black than white tokens in their store – which are both maladaptive behaviours. But the connection weights of their neural network evolve in a succession of generations and, at the end, the robots have acquired the capacity to ask for the food type which in that moment they need more and to accept an exchange proposal only if it gives them the food type which they need more.

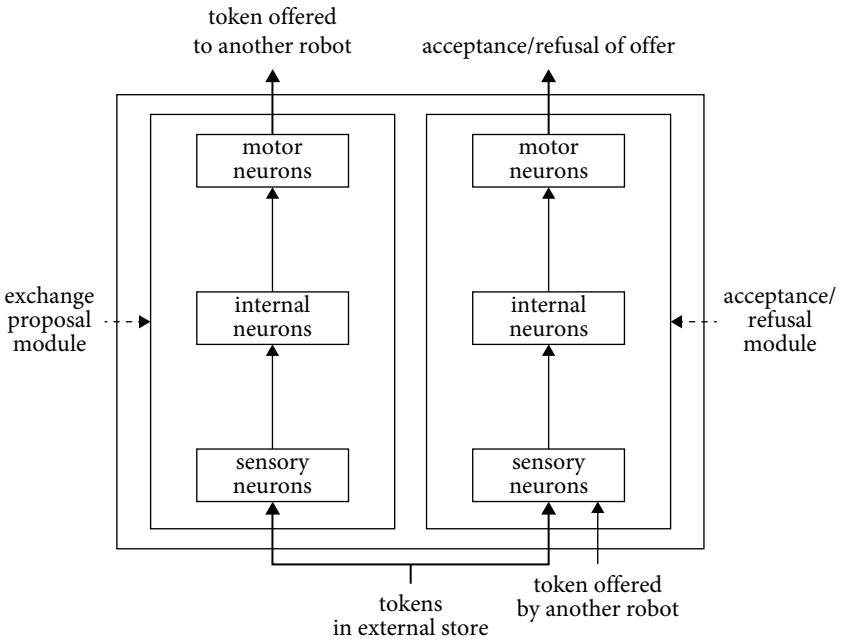


Figure 14. The neural network of the robots is made of two modules, one for proposing to another robot to exchange one type of food for the other type of food and the other one for accepting or refusing the exchange proposed by another robot.

These robots are a first step towards reproducing the capacity to make exchanges. Now we make a second step. The robots decide which token to ask to the other robot, a black token or a white token, and they also decide whether to accept or reject an exchange proposed by the other robot but, in all exchanges, it is automatic that one token of one type is exchanged for one token of the other type. This restriction must be eliminated because an important property of the exchange of goods is the *quantity* of the two goods which are exchanged, and the robots might exchange, for example, *two* black tokens for only *one* white token, or vice versa. So we construct another type of robots whose neural network has output neurons which specify how many tokens of one type they want in exchange for one token of the other type and the robots proceed to an exchange only if they agree on this number. This is important because the exchange of goods is another way to measure the value of goods. In the preceding section the value of goods was measured in terms of preferences. A robot sees a black token and a white token and, if the robot approaches the black token rather than the white token, the black tokens have more value than the white tokens for the robot. But the value of a good can also be measured when the good is exchanged with other goods. If a robot accepts to exchange *two* white tokens for *one* black token, for the robot the black tokens have twice the value of the white tokens. Measuring the value of goods on the basis

of what we might call their "exchange rate" – two black tokens are exchanged for one white token – introduces an important novelty. The value of goods measured in terms of preferences is an individual measure of the value of goods. It tells us something on the particular individual which exhibits the preferences – and different peoples can have different preferences – and it also exists in the individual economy of nonhuman animals. On the contrary, measuring the value of goods on the basis of the exchange of goods tells us something not on particular individuals but on the entire community of individuals – the society. If a robot is offered *two* white tokens in exchange for *one* black token, the robot can reject the exchange proposal and it can go to another robot and accept to proceed to the exchange only if it is offered *three* white tokens for one black token. If all robots do the same, the exchange rate for the black tokens and the white tokens *is* one black token for three white tokens. The exchange rate is a *social* measure of the value of the black and white tokens for the robots.

That the values of goods is linked to the exchange of goods can also be seen if we go back to our robots which live in the environment which contains not only food tokens but also tool tokens that allow the robots to extract more energy from the food tokens. This scenario is interesting because there is an asymmetry between food and tools. For the robots that need two types of food to survive, the two types of food are both necessary. Now the situation is different. If a robot's external store contains only food and no tools, the robot can survive by eating its food tokens, even if it cannot extract much energy from the food tokens because it lacks tools. But if a robot has only tools in its store and no food, the robot dies because it has nothing to eat. The peasant robots can survive without tools but the artisan robots cannot survive without food. The exchange of food for tools is *advantageous* for the peasant robots but the exchange of tools for food is *necessary* for the artisan robots. This has consequences for the value of food and the value of tools. The robots are hardwired by us to exchange one food token for one tool token but if they were free to decide how many food tokens to exchange for how many tool tokens, we predict that the peasant robots would ask and obtain from the artisan robots more tool tokens in exchange for one food token. This is only a prediction because we do not have actually constructed robots like these.

But this difference between food and tools poses an interesting problem. In our simple scenario, food has more value than tools and therefore the peasant robots have more exchange power than the artisan robots. In modern human societies we find the opposite. People who produce goods that increase the value of other goods tend to have more exchange power than people who produce food or other goods which can only be directly consumed. So what should we do so that tools have more value than food and the artisan robots have more exchange power than the peasant robots? The artisan robots might modify the tool tokens so that they make it possible to extract even more energy from the food tokens. In this way, the value of tools would increase indefinitely compared to the fixed value of food and the artisan robots would have more exchange power than the peasant robots. And, of course, the "artisan" robots could create new

goods – new things that the "peasant" robots want to have – and this would indefinitely increase their exchange power.

We have made our robots more realistic by not hardwiring their exchanges but by letting them autonomously decide if and what to exchange. But there is something which we still don't like in our robots: two robots can make an exchange independently from where they are in physical space. This clearly is a limitation because, to exchange food and other physical things, two robots must be near to one another in physical space. Human beings have solved this problem by selecting one particular place where people go to make exchanges: the marketplace. Marketplaces can be reproduced by placing a landmark in the environment of the robots that the robots can see from everywhere. When a robot wants to make an exchange, the robot goes to the landmark (the marketplace) where it finds other robots ready to make exchanges. We have not constructed these robots but we have constructed other robots that solve in another way the problem of being near to one another to make exchanges. All the exchanges we have described so far are based on the assumption that a robot asks a token from another robot because it needs the token to eat it (food) or to use it (tools). But human beings also make another type of exchanges. An individual obtains a good from another individual through an exchange but the individual is not interested in consuming the good or using it in some other way. He or she is only interested in exchanging the good with a third individual. Human beings who do this are called merchants. We describe merchant robots in the next section.

4. Merchant robots

To make an exchange two robots must be near to one another because food or tools are physical entities and they can be transferred from one robot to another robot only if the two robots are near to one another in physical space. So far we have ignored this requirement of spatial proximity. When we select two robots for an exchange, we assume that they are near to one another in physical space so that they can directly proceed to the exchange. Or they live sufficiently near to one another that they can meet other robots with whom to make exchanges in a marketplace. We don't want to make these assumptions any more. This creates a new problem which can be solved by constructing merchant robots.

The robots live in an environment which contains both black food tokens and white food tokens. The black tokens and the white tokens have the same energetic value but the two types of tokens are in two different zones separated by a large empty space. The robots that live in one zone collect one type of tokens and the robots that live in the other zone collect the other type of tokens. These are the producer robots. The producer robots can survive by eating only the food tokens which exist in their respective zones but they live better – have more survival and reproductive

chances – if they eat both types of food. The problem is that the two food zones are very distant and the possibility to go from one zone to the other zone is ruled out for the robots. This problem is solved by constructing another population of robots, the merchant robots. The merchant robots go back and forth from one zone to the other zone of the environment bringing with them in their external stores both black and white tokens (see Figure 15). When they are in the black zone, they obtain from the producer robots that live there black tokens in exchange for white tokens. When they are in the white zone, they obtain from the robots that live there white tokens in exchange for black tokens. The merchant robots do not directly produce any food but they produce another good which the producer robots want to have: the possibility to eat both types of food.

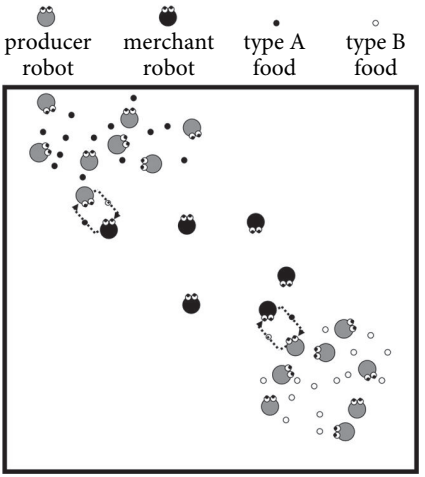


Figure 15. The producer robots live in two different zones of the environment which contain two different types of food. The merchant robots go back and forth from one to the other zone to exchange one type of food for the other type of food with the producer robots.

The producer robots are our basic robots. Their behaviour is controlled by a neural network with evolved connection weights that allows them to collect the food tokens existing in their respective zone. The merchant robots do not have a neural network for collecting food but they have a neural network for making exchanges with the producer robots. In each cycle one of the merchant robots, randomly chosen, visits either the black zone or the white zone and it tries to make exchanges with the producer robots that live there. But since the neural network of the merchant robots allows them to decide the terms of the exchange, what we find is that the merchant robots do not ask the producer robots one token of one type in exchange for one token of the other type. The merchant robots ask for *more than one* token of one type in exchange for *one* token of the other type. The reason is clear. Like the producer robots, the merchant robots

can survive only it they eat the food tokens and, therefore, they must necessarily ask for more than one food token of one type in exchange for one food token of the other type because, otherwise, they would have no gain and they would not be able to eat. The neural network that controls the behaviour of the merchant robots takes as input how many tokens of the two types are presently contained in the merchant's external store and it gives as output the exchange rate of that particular merchant robot: the number of tokens of one type which the merchant robot wants from a producer robot in exchange for one token of the other type (see Figure 16).

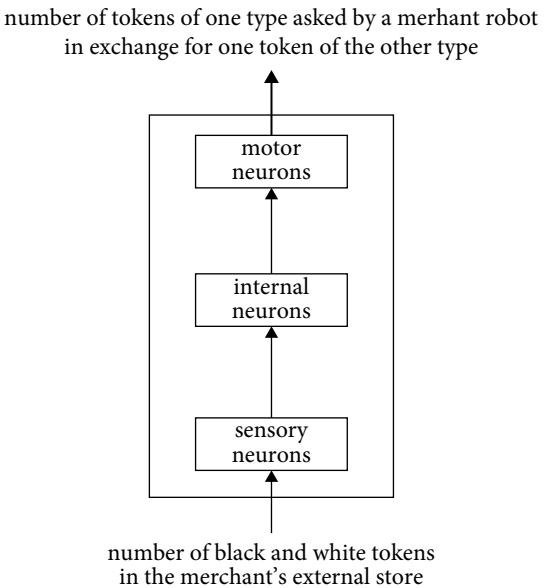


Figure 16. The input to the neural network of a merchant robot is how many black and white tokens are contained in its external store and the output is the number of tokens the merchant robot wants from a producer robot in exchange for one token of the other type.

While the exchange behaviour of the merchant robots is under the control of a neural network with evolved connections weights, the exchange behaviour of the producer robots is hardwired by us. We decide what is the maximum number of tokens that the producer robots are disposed to give in exchange for one token of the other type and, only if the exchange rate proposed by the merchant robot is lower than this number, the exchange takes place. Otherwise there is no exchange. At the beginning the merchant robots are not very good at proposing the appropriate exchange rate because their neural networks have random connection weights and they ask for too many tokens of one type in exchange for too few tokens of the other type – and, therefore, there are few exchanges. But, after a certain number of generations, they become able to propose

the appropriate exchange rates and this makes it possible for both the producer robots and the merchant robots to eat both types of tokens.

With this robotic economy which includes merchant robots, we find many interesting results. The first result is that at the end the exchange rate proposed by the merchant robots is just a little lower than the maximum exchange rate which is accepted by the producer robots – which, as we have said, is decided by us and is identical for all the producer robots. In other words, the merchant robots are able to obtain the tokens from the producer robots at the best possible conditions – for them, the merchant robots. Second, the average merchant robot has more "wealth" than the average producer robot, where "wealth" is the quantity of food tokens of both types contained in a robot's external store. Third, there is more inequality of wealth among the merchant robots than among the producer robots. There are a few very rich merchant robots and many poor merchant robots while the differences in wealth among the producer robots are not so large. Fourth, the best producer robots live longer and generate a greater number of offspring, while the best merchant robots tend to live less and to have fewer offspring. This is interesting if we remember the robots which have economic inheritance (see Chapter 7). Since the offspring of a robot inherit not only the genotype of their parent but also the food tokens contained in their parent's external store, the strategy of the merchant robots is to leave many tokens to few offspring while the strategy of the producer robots is to leave few tokens to a larger number of offspring. And this is linked to our last result. While initially there is the same number of merchant robots and producer robots and this number is decided by us, at the end the merchant robots are fewer in number than the producer robots. These are interesting results because, although they have been obtained with extremely simplified robots, they make us look at human societies and human economies in potentially useful ways.

5. The emergence of money

So far the exchange which takes place between two robots is barter. The robots have goods which they can exchange for other goods but which they can also directly eat or use. Human beings have had barter in the past but, today, barter has almost disappeared because they have invented a new type of good which cannot be directly used and consumed but can only be exchanged with other goods. This good is money. Our next robots develop money.

The robots described so far need only two types of goods to remain alive and live well, either two types of food or food and tools, and since they need only two goods, there is a high a priori probability that, when two robots come together to make an exchange, they will have complementary needs. One robot has more of one good and the other robot has more of the other good and, therefore, the two robots can proceed to making the exchange. But if the robots need or want 10, 100, or 1000 different types

of goods to live well, it becomes increasingly improbable that, when a robot offers to another robot the good which it has in the greatest quantity in its external store in exchange for the good which it has in the smallest quantity, the other robot will have exactly the complementary needs. Many exchange proposals will be rejected and the robots will not be able to exploit the advantages of making exchanges.

How can the problem be solved? The answer is money. Money is a good which facilitates the exchange of goods. Whatever good a robot is asking to another robot, the robot will offer in exchange always *one and the same* good. And the other robot will always be ready to accept this particular good in exchange. The good which is exchanged in all exchanges is money.

Money spontaneously emerges in our robots. We construct a new type of robots which, to live and have offspring, need not two but eight different types of goods. We do not reproduce the robots' behaviour of procuring these eight goods. At each time step, one unit of one of the eight goods, randomly chosen, is added to the external store of all robots to simulate the independent procurement of goods by the robots. And, at each time step, a robot consumes one unit of one of the eight goods, also randomly chosen, and if this is impossible because the robot's external store contains zero units of that particular good, the robot dies. In all exchanges the good which a robot is asking to the other robot is hardwired by us and it is always the good most needed by the robot in that moment. Therefore, the only thing that a robot must be able to do is decide which good to offer in exchange. The neural network of these robots is the same as the neural network of the robots that propose and accept or refuse an exchange of goods and it is made up of two modules (Figure 17).

The input neurons of both modules encode how many units of each of the eight different goods are currently contained in the robot's store. The output neurons of one module encode which good the robot is offering to another robot in exchange for the good which it is asking for – which, as we have said, is hardwired to be the good which is less abundant in the robot's external store. This is the neural module that controls the behaviour of the robot when the robot is making an exchange proposal to another robot. The neural module which controls the behaviour of the robot when the robot must decide whether to accept or refuse the exchange proposed by another robot has input neurons which, in addition to encoding the state of the robot's external store, also encode the exchange proposed by the other robot, and output neurons which encode the decision of the robot to accept or reject the exchange.

Initially, the connection weights of the robots' neural network are random and therefore the robots tend to offer any type of good in exchange for the good requested and, since the other robot will probably not need that particular good, there are very few successful exchanges. Then, the selective reproduction of the best robots leads to a different behaviour. The robots tend to always offer *one and the same* good in exchange for whatever good they are asking for, and they *always* accept this particular good in exchange for the good they are asked for (see Figure 18).

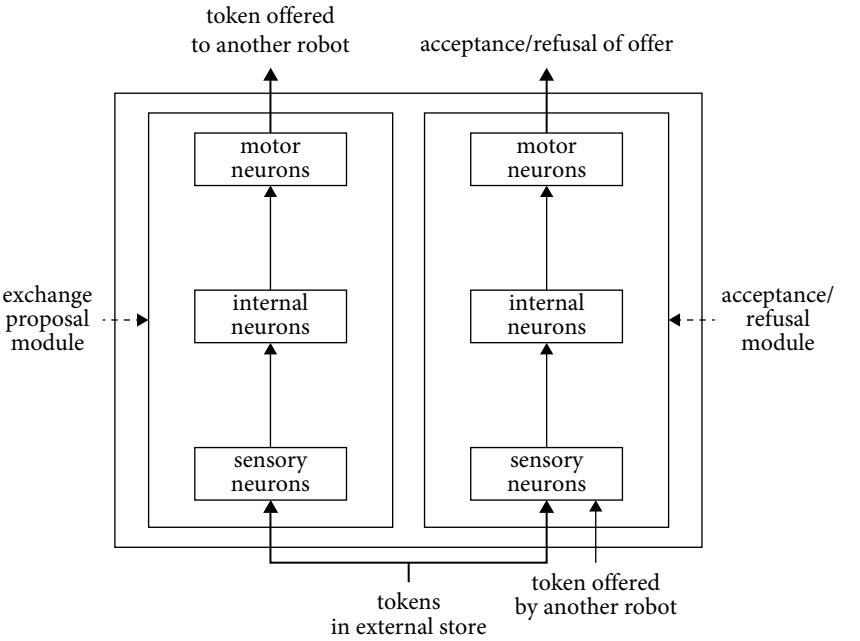


Figure 17. The neural network of the robots which end up exchanging food for money. The exchange proposal module decides which token to offer to another robot while the acceptance/refusal module decides to accept or refuse the proposal made by another robot.

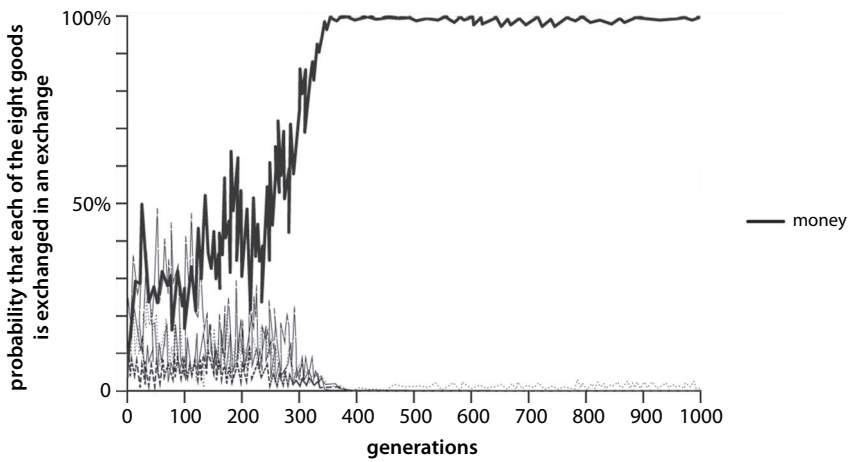


Figure 18. At the beginning any of the eight goods is offered in exchange for any other good but, after 200–300 generations, only one good (money) tends to be offered and accepted in all exchanges.

This good is money. Money is the good that the robots exchange in all exchanges. One would expect that, in exchange for the good which it is asking for, a robot will offer the good which it has in the greatest quantity in its external store and that the other robot will accept this good only if it does not have much of this good in its external store. But this is not what happens. Money is offered in exchange even if it is not the good which the robot has in the greatest quantity in its external store, and money is accepted in exchange even if the robot already has a large quantity of money in its external store. Since money is accepted in exchange for all other goods, every robot always needs money because the possession of money means that the robot is able to obtain any possible good from other robots. With money the external stores of the two robots that want to make an exchange – the buyer and the seller – do not need to be complementary for the exchange to succeed, and this is why money facilitates the exchange of goods. With the appearance of money the number of successful exchanges increases and the robots have more fitness – live longer (see Figure 19).

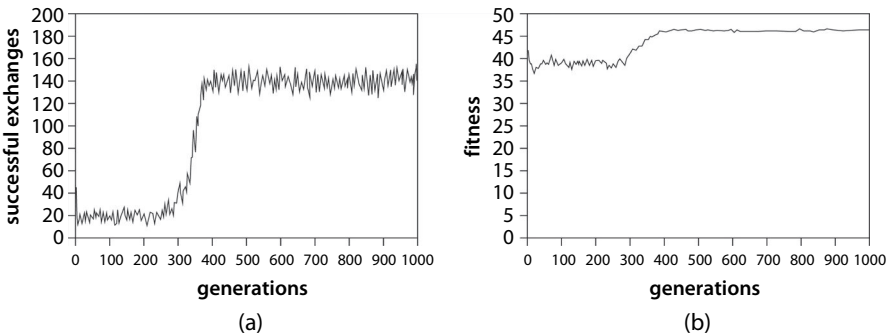


Figure 19. With the appearance of money the number of successful exchanges increases (a) and the robots have more fitness – live longer (b).

These robots tell us many interesting things. When the number of different goods which are exchanged among the robots becomes very large, money emerges spontaneously, without any explicit agreement among the robots. And money tends to emerge abruptly. For a certain number of generations there is no money in the sense that there is no single good which is exchanged in all exchanges and the number of successful exchanges is very small. Then, in a very few generations, one particular good is exchanged in most exchanges and the number of successful exchanges becomes much larger. This reveals the social nature of money. Something is money not because of its intrinsic properties but because it is accepted as money by all the robots. And this happens abruptly because one good can become money only if a *sufficient* number of robots accept that good as money, where "sufficient" implies a threshold which is either crossed or not crossed. A robot offers money to another robot in exchange for whatever good it is asking from the other robot because it "knows" that money will always be accepted

by the other robot. And the other robot always accepts money because it "knows" that with money it will be able to obtain any good in future exchanges with other robots.

The social nature of money also implies that money is arbitrary. Anything can be money, provided it is a good which is exchanged in all exchanges. And, in fact, which of the eight goods becomes money is different in the different replications of the simulation. In actual reality money tends to have specific physical properties – it does not occupy much space, it can be transported easily, it does not deteriorate – because these properties facilitate the exchanges of goods. We have ignored these physical properties of money because they are not relevant for our robots. But there is another property of money that we cannot ignore because it is the essence of money: money is made up of identical units (euros, dollars, pounds, yen) which can be counted, added, subtracted, multiplied, and divided. This property of money is important because, by being made of units that can be counted, money becomes a tool for measuring the *absolute*, not only the relative, value of a good. This is what we see in the next Section.

6. Money and the price of goods

The robots described in the preceding section pay one unit of money for one unit of whatever good they are buying and, therefore, all goods have the same value for them. But we can evolve robots that autonomously decide to exchange one unit of one good with more units of another good and, in fact, the robot described in Section 5 exchange one token of food A for two tokens of food B if food A is less abundant in their environment or if it is more important for the robots' survival compared with food B. As we have said, this gives us a measure of the value of goods. One good has twice the value of another good if one unit of the first good is exchanged for two units of the second good.

However, this is only a *relative* measure of the value of goods because the value of a good is determined only with respect to another specific good. Money is a measure of the *absolute* value of a good because the value of *all* goods is how many units of money a robot must give to another robot to obtain a good from the other robot. This is the price of the good, and the goods of our next robots have a price.

The new robots are divided into two categories, the seller robots and the buyer robots, and there are two categories of goods, apples and money. The seller robots give to the buyer robots apples in exchange for money and the buyer robots give to the seller robots money in exchange for apples. One unit of the good "apples" is one kilo of apples and one unit of the good "money" is one money-unit (say, one euro). In all exchanges only one kilo of apples is bought and sold but this kilo of apples can be exchanged for one or more money-units (one or more euros). Our question is: What determines how many money-units are given in exchange for one kilo of apples? In other words, what determines the price of apples?

As we have said, there are two separate populations of robots, the seller robots and the buyer robots. The genes of the seller robots encode how many money-units a seller robot asks to a buyer robot for one kilo of apples. The genes of the buyer robots encodes how many money-units the robot is ready to give to a seller robot in exchange for one kilo of apples – how much the buyer is disposed to pay for one kilo of apples. The behaviour of the robots is not controlled by a neural network but is directly controlled by their genes. The genes of a seller robot encode the number of money-units which the seller robot asks to buyer robots for one kilo of apples. The genes of a buyer robot encode the number of money-units that the buyer robot is disposed to pay for one kilo of apples. In each cycle a seller robot and a buyer robot are selected randomly from the two populations of sellers and buyers and we compare their genes. If the price of one kilo of apples which is encoded in the genes of the seller robot is not higher than the price encoded in the genes of the buyer robot, the two robots proceed to the exchange. If the seller robot is asking a higher price than the price that the buyer robot is disposed to pay, the exchange does not take place. Then we move to the next cycle and we select a new seller robot and a new buyer robot.

The evolutionary process takes place separately in the two populations of seller and buyer robots. The seller robots and the buyer robots reproduce selectively and with the constant addition of random changes to their inherited genes. For the seller robots, the criterion of selection for reproduction is how many kilos of apples a seller robot has sold during its life and the price at which it has sold those apples (which, as we have said, is encoded in its genes). Given two seller robots which have sold the same quantity of apples, the robot which has more offspring is the one which has a higher price encoded in its genes and therefore has gained more from selling the apples. The fitness of buyer robots is how many kilos of apples a buyer robot has been able to buy during its life and the price it has paid for those apples. Given two buyer robots which have bought the same quantity of apples, the robot which has paid less for one kilo of apples – the price encoded in its genotype – is more likely to reproduce than the robot which has paid more. Clearly, the price encoded in the genotype of both the seller and buyer robots must not be too high and it must not be too low. For sellers, if the price encoded in their genotype is too high, they will not sell their apples while, if it is too low, they will not earn much money from the apples they sell. For buyers, if the price encoded in their genotype is too high, they will spend too much money for buying the apples and, if it is too low, they will not be able to buy the apples.

The genes of the initial population of both sellers and buyers are random and this implies that both the sellers and the buyers are not very good at selling and buying apples. The sellers either do not sell many apples because they ask for too high a price or they sell many apples but they earn little money. The buyers buy few apples because the price that they are disposed to pay is too low and is not accepted by the sellers, or they buy many apples but they spend too much money because the price they are disposed to pay is very high. Then things begin to change. The genes of both sellers

and buyers gradually improve because of the selective reproduction of the best sellers and of the best buyers and the addition of random mutations to the inherited genes, until the population reaches a stable state. When the population reaches a stable state, we see that the number of successful exchanges is much greater than it was at the beginning, and this represents an advantage for both the seller and the buyer robots. Everything is as good as possible. The seller robots ask for a price which is just a little lower than the price the buyer robots are ready to pay for one kilo of apples, and this allows the seller robots to sell their apples and to earn as much money as possible. The buyer robots are ready to pay a kilo of apples just a little more than the price asked by the seller robots, and this allows them to buy the apples without spending too much. The number of money-units which are paid for one kilo of apples – the price of the apples – becomes stable and it is more or less the same for all seller and buyer robots.

What is this price? What determines the price? The answer is the number of seller and buyer robots. If there are more sellers than buyers, the price of one kilo of apples is lower while, if there are more buyers than sellers, the price of one kilo of apples is higher. In other words, the price of a good is a function of the relative number of sellers and buyers for that good. Our robots necessarily buy only one kilo of apples in each exchange and there are no limits to the quantity of apples that the sellers can sell and that the buyers want to buy. If we construct more realistic robots, we predict that what we would find is that the price of a good is a function of what economists call the demand and the offer of the good.

Our seller and buyer robots are in a symmetrical condition because the two categories of robots have the same "power" with respect to the other category. But this is not true for human sellers and buyers, and future robots should tell us why this is not true and in how many different ways it is not true. The seller robots compete with the other seller robots for the quantity of apples that each robot is able to sell and for the price at which the robot sells its apples. A seller robot cannot sell its apples at a price that is higher than the price practiced by another seller robot because, if it does this, the buyer robots will buy the apples of the other, less expensive, robot. But the seller robots could agree among themselves not to sell the apples at a price lower than a given price – and this would damage the buyer robots that cannot buy the apples at a lower price. This is why the law punishes these explicit agreements which economists call "cartels". Cartels are not possible among our robots because our robots do not know what is the price practiced by the other robots and it is only from one generation to the next that they can adjust the price at which they sell their apples. But if the seller robots could know the price practiced by the other seller robots, this might lead to the creation of spontaneous cartels which, although they damage the buyer robots, would not be punished by the law. All the robots would sell the apples at the same high price because they know – without any explicit agreement – that no other robot is selling the apples at a lower price.

Of course, the buyer robots could do the same. If they could know the price practiced by all seller robots, they would not buy the apples of a robot which sells its apples at a price higher than the price of other seller robots. But between sellers and buyers the situation is not a symmetrical one. First, in the real world the quality of goods varies and, while sellers know the quality of the goods that they sell, buyers do not really know the quality of the goods that they buy. Second, buyers can have an immediate need for some good and they may not have the time to visit many different sellers to see at what price they sell that good. Third – and this is the most important reason – sellers can increase the motivation of buyers to buy their goods by using all sorts of marketing techniques, while buyers cannot increase the motivation of sellers to sell their goods because the motivation of sellers to have money is already as strong as it can possibly be.

Measuring the value of a good in terms of the number of money-units which buyers are disposed to give in exchange for the good has three important consequences. First, measuring the value of a good on the basis of preferences (this good has more value than that good because it is preferred to that good; see Section 2) gives different results as a function of the particular individual which is expressing those preferences. A robot can prefer X to Y but another robot can prefer Y to X, and the same robot can have different preferences in different circumstances. (Remember what we have said on behaviour in the experimental laboratory and behaviour in real life.) Therefore, measuring the value of goods in terms of preferences is not a universal or "objective" measure of the value of goods. On the contrary, measuring the value of a good as the number of money-units which buyers are disposed to give to sellers to obtain the good, is independent of the particular individual which buys the good and is a universal or "objective" measure of the value of the good.

Second, the value of goods based on preferences is not very informative because preferences are not necessarily "transitive". If an individual prefers X to Y and he or she prefers Y to Z, it is not necessarily true that the individual prefers X to Z. Therefore, we need to compare all pairs of goods to establish the value of the goods. Furthermore, the value of goods in terms of preferences tells us that X is preferred to Y but it does not tell us how strong is the preference. In contrast, the value of a good expressed as the price of the good tells us not only that good X has more value than good Y but also *how much* the value of X is greater than the value of Y. And, here, transitivity is guaranteed. If the price of X is higher than the price of Y and the price of Y is higher than the price Z, the price of X is necessarily higher than the price of Z.

Third, the value of a good measured in terms of preferences tells us what is the value of the good compared to another specific good but not what is the value of the good compared to all other goods. Instead, the value of a good measured as its price tells us the "absolute" value of the good with respect to all other goods. And this is a crucial requirement for modern human economies.

At the beginning of this Chapter we have said that behaviour is intrinsically economic because it is concerned with the acquisition of goods. Should we conclude that every good has a value which can be measured as the number of money-units which an individual is disposed to give to another individual in exchange for the good? The answer is No because human beings do not obtain all their goods from others and, even when they obtain their goods from others, they do not obtain these goods in exchange for money. The science of economics is only interested in the goods which are obtained from others in exchange for money, and this is a serious limitation because – as we said at the beginning of this Chapter – to understand the economic behaviour of human beings, what is needed is a general science of goods and the value of goods. But the fact that the science of economics studies only the goods which are exchanged for money has another consequence. The science of economics influences how human economies are organized and how they function and, in fact, it has created modern human economies. But what is really important is that the science of economics pushes human beings to see all goods as goods which can be obtained in exchange for money. "We know what things costs but have no idea what they are worth", as Tony Judt – a political scientist, not an economist – has written. This may be bad for human beings and for their societies. To really understand human economies and to construct human economies which do not damage human beings, we need a general science of goods, and robots should provide us with a general science of goods which includes the science economics but is not restricted to the science of economics.

7. Entrepreneur robots and worker robots

So far, all goods are produced by individual robots. A robot explores the environment and it collects the food or tools which it is able to find in the environment. But in human economies many goods are produced not by single individuals but by organizations of individuals which are called private enterprises. The next step in the development of a robotic economy is to create organizations of robots in which many robots work together to produce goods that no individual robot would be able to produce by working alone. Private enterprises are uniquely human because they are based on the exchange of goods *within* the organization, and the exchange of goods only exists among human beings. What is exchanged within a private enterprise are two goods: work for money. Workers give their work to the owner of the enterprise and, in exchange, the owner of the enterprise gives some money to his or her workers.

Our next robots reproduce this type of exchange of goods. There are two separate populations of robots: worker robots and entrepreneur/owner robots. The worker robots produce goods (say, cars) for the entrepreneur robots and the entrepreneur robots pay a salary or wage to the worker robots. All robots have a certain number money-units in their external store and, at each time step, a robot consumes one of

these money-units to buy things and remain alive. Since the robots die if the number of money-units in their external store is zero, the robots must reintegrate the money-units they have in their external store. Worker robots reintegrate their money-units by selling their work to the entrepreneur robots. Entrepreneur robots reintegrate their money-units by selling the cars produced by the worker robots. Entrepreneur robots give a certain number of money-units to each of their worker robots in exchange for its work and, therefore, an entrepreneur robot consumes its money-units not only to buy things and remain alive but also to pay the salary or wage to its workers (see Figure 20).

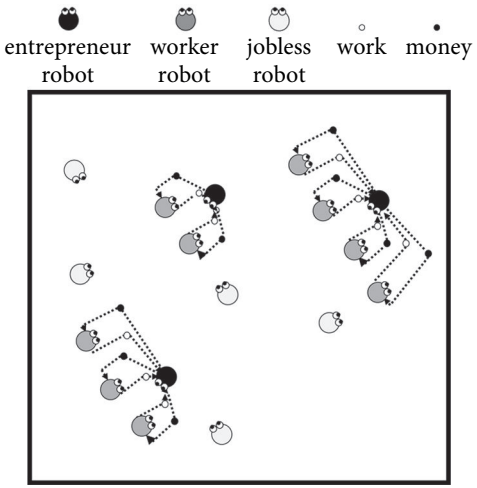


Figure 20. Entrepreneur robots exchange money for work with worker robots. Some worker robots are jobless.

What determines the reproductive chances (length of life, number of offspring) of the two categories of robots? The reproductive chances of a worker robot depend, first, on its ability to sell its work to an entrepreneur robot (to find a job) and, second, on the number of money-units that the worker robot is able to get from the entrepreneur robot (how high is its salary). The reproductive chances of an entrepreneur robot depend on its ability to employ as many worker robots as possible to produce and sell more cars and to pay a salary as low as possible to its worker robots. We do not reproduce the car market itself. The cars that are produced by the entrepreneur robots are all sold at the same price and all the cars which are produced are sold. Since the market has a fixed size, the success of an entrepreneur robot is measured by its market share – the proportion of all cars which are produced by that entrepreneur robot.

The fitness of both the entrepreneur and worker robots depends on how they behave in their exchanges with the other category of robots. Like the robots that sell and buy apples, the behaviour of entrepreneur robots and worker robots is encoded in

their genes. The genes of an entrepreneur robot specify the maximum salary that the entrepreneur robot is ready to pay to a worker robot for its work. The genes of a worker robot specify the minimum salary that the worker robot is disposed to accept from an entrepreneur robot in exchange for its work. But, since contacting an entrepreneur robot to find a job has a cost for the worker robots, a worker's genes also specifies the number of entrepreneur robots that the worker robot contacts when it is trying to find a job. When an unemployed worker robot contacts an entrepreneur robot (randomly chosen), the contact is successful if the maximum salary that the entrepreneur robot is ready to pay to the worker robot is higher than the minimum salary the worker robot is ready to accept. Otherwise, the contact is not successful and the worker robot must contact another entrepreneur robot. The worker robots work for an entrepreneur robot for a fixed number of cycles which is the same for all worker robots and then they become unemployed and must find another employer.

At the beginning the genes of all robots are random and, therefore, few worker robots find a job and few entrepreneur robots have workers that work for them. Both the worker robots and the entrepreneur robots have a short life and leave few offspring and, therefore, they both decrease in number. Then, in a succession of generations the worker robots and the entrepreneur robots become much better at exchanging work for money and their number increases until it reaches a stable value. At the beginning, the number of worker and entrepreneur robots is the same but, at the end, there are fewer entrepreneur than worker robots because one entrepreneur robot employs many worker robots. Since the entrepreneur robots which have more offspring are those which employ more worker robots (and, therefore, produce more cars), the maximum salary which these *successful* entrepreneur robots are disposed to pay to their worker robots gradually diffuses in the entire population of entrepreneur robots. And the same is true for the worker robots. The worker robots which have more offspring are those which tend to be employed rather than unemployed and, therefore, it is the minimum salary that these *successful* workers are ready to accept and the number of entrepreneurs that they contact when they are unemployed which are adopted by the entire population of workers.

What is the situation when everything stabilizes? The situation is not favourable to worker robots. The salary paid by the entrepreneur robots to the worker robots is only a little more than what is needed by the worker robots to remain alive. And another result goes in the same direction. As we have said, the success of an entrepreneur robot is measured in terms of its market share. If we increase the size of the market – more cars can be sold and bought – we find that the wealth of the entrepreneur robots, measured as the number of money units that they have in their external stores, increases while the wealth of the worker robots remains the same. Economic development increases the wealth of the entrepreneur robots but not the wealth of the worker robots.

But the situation is not so simple because in an economy based on private enterprises there is a complex interaction among the interests of the different categories of

robots. The entrepreneur robots compete with the worker robots for the salary they pay to the worker robots, which the entrepreneur robots want to be as low as possible and the worker robots want to be as high as possible, and the entrepreneur robots tend to win in the sense that they have more wealth than the worker robots. But there is also competition *within* each of the two categories of robots. The entrepreneur robots compete with the other entrepreneur robots for which of them employs more worker robots and in this way produces and sells more cars and has a larger share of the market. The worker robots compete among them for which of them finds a job by asking for the appropriate salary.

But what is more interesting is that there is also a *convergence of interests* between the two categories of robots. The entrepreneur robots need the worker robots to produce and sell cars and the worker robots need to be employed by the entrepreneur robots to get a salary. If there is a problem for one category of robots, there is also a problem for the other category of robots. This complex interaction of both divergent and convergent interests is typical of modern human economies.

Overall, our robotic scenario favours the entrepreneur robots against the worker robots. But there is something which the worker robots can do to improve their condition. As we have seen, if each worker robot individually negotiates with an entrepreneur robot the salary that the entrepreneur robot will pay to the worker robot, the entrepreneur robots pay to the worker robots a salary which is only a little more than what is necessary for the worker robots to remain alive. But the worker robots have one possibility open to them that may reduce their disadvantage: they can agree among themselves to ask for the *same* minimum salary when they contact the entrepreneur robots to find a job. In other words, the worker robots can create robotic unions. We construct another scenario in which the genes of all the worker robots specify the *same* minimum salary and we find that, with robotic unions, the salary paid by the entrepreneur robots to the worker robots is higher. The reason is clear. If a worker robot individually negotiates its salary with an entrepreneur robot, it does not make sense for the worker robot to ask for too high a salary because the entrepreneur robot can find another worker robot which accepts a lower salary. If the robots unionize, this changes. Entrepreneur robots *cannot* refuse to employ a worker robot which asks for a high salary because the other worker robots will also ask for the same high salary. Therefore, if the entrepreneur robots want to have worker robots for producing cars, they must accept the higher salary requested by the worker robots.

But there are two limits to this strategy of unionization of the worker robots. The first limit is that a sufficient number of worker robots must adhere to the agreement for the strategy to work. If this number is not reached and too few robots are members of the union, the entrepreneur robots will find other worker robots which ask for lower salaries and, in this way, they will be able to pay lower salaries. Figure 21 shows that the salary of worker robots increases with the number of workers that are unionized but it remains low if too few robots belong to the union.

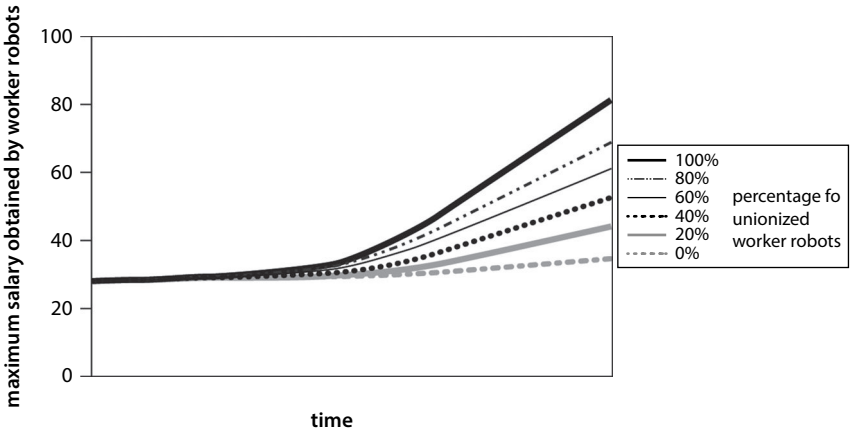


Figure 21. The salary that the worker robots obtain from the entrepreneur robots progressively increases only if more than 50% of the worker robots belong to a union.

The second, and more serious, limit is that the unionized worker robots cannot ask for a salary which is *too* high because, if they do so, the entrepreneur robots may not have enough money units to pay the salaries and they may prefer to close their enterprise leaving the worker robots jobless.

The complexity of an economy based on private enterprises emerges even more clearly if we look more widely at the society in which there are entrepreneur and worker robots. We have assumed that the cars produced by the entrepreneur robots are sold in a market but we haven't said anything on who buys the cars and on what is the price of the cars. Clearly, for an entrepreneur robot to sell its cars there must be robots that buy these cars. Who buys the cars which are sold by the entrepreneur robots? As we have seen, in an economy based on private enterprises most of the robots that make up the society belong to the category of worker robots. Therefore, most buyers of the cars are worker robots. In human societies there are many different types of goods, not only cars, and worker robots that produce cars can buy the goods produced by other worker robots, but the problem remains. This reveals another, more delicate, conflict of interests which exists in an economy based on private enterprises. There is not only a conflict of interest between the entrepreneur robots and the worker robots and among the robots that belong to the same category but there is also a conflict of interest *inside* the category of worker robots between their interests as workers and their interests as buyers. As workers, the robots want higher salaries, and this pushes the entrepreneur robots to sell their products at a higher price. But since the worker robots are also buyer robots, they want the prices of the goods to be lower, not higher.

And the situation is even more complex. As we have already said, between the category of workers and the category of entrepreneurs there are not only conflicts of

interests but also convergences of interests because the entrepreneur robots need the worker robots to produce their cars and the worker robots need the entrepreneur robots to get a salary. But the convergence of interests goes beyond that. In an economy based on private enterprises it is in the interest of both the entrepreneur robots and the worker robots that the quantity and variety of goods which are produced and sold increase as much as possible. If the quantity and variety of goods increase, the entrepreneur robots can make more profits and the worker robots can more easily find jobs and get a salary with which they can buy more and more goods. Another convergence of interests is that the entrepreneur robots are adversaries for the worker robots in the exchange of work for salary but they are adversaries which must be preserved and even helped because they are the producers and sellers of the goods that the worker robots as buyers want to buy. This mixture of conflict and convergence of interests explains why the an economy based on private enterprises is so complex and inherently full of conflicts and, at the same time, why it produce so many goods that human beings – and economists – cannot even imagine a different economy. Robotic economies should help us to understand the complexity of modern human economies and, since today the economy tends to be globalized and this increase its complexity, the complexity of a globalized economy.

Before we close this section on robotic private enterprises we want to mention another type of organization which produces goods: the state. In the preceding Chapters we have described states as central stores to which all the members of a community give some of their goods and which redistribute these goods to all the members of the community. But the state does not simply redistribute the goods that it receives from the members of the community but it uses these goods (taxes) to produce new goods: health systems, pension systems, educational systems, juridical systems, and their own administrative structure. So it is natural to ask how states and private enterprises compare as organizations for the production of goods. States are like private enterprise because they also pay salaries or wages to worker robots to produce the goods that they produce (public workers) but they are unlike private enterprises because they do not sell their products in a market.

We reproduce these phenomena by creating a population of *public* robotic enterprises which are like the private enterprises we already know with the only exception that the public enterprises do not compete among them on a market and they do not have the equivalent of an entrepreneur/owner robot which is personally interested in the economic performance of its enterprise and, therefore, works for its best performance. Another difference is that the number of public enterprises is fixed and they do not compete with each other for survival and reproduction.

What are the results? One result is that worker robots obtain higher salaries from public enterprises than from private enterprises because the robots that govern public enterprises do not pay the workers' salaries with their money and, therefore, are not interested in paying lower salaries. A second, related, result is that there are fewer

jobless workers with public enterprises than with private enterprises because public enterprises accept jobless workers whatever salary they ask for – within the limits of the state's budget – while this not true for private enterprises.

But not everything is better for the worker robots in a society without private enterprises. Worker robots are also buyer robots and the number of cars produced by private enterprises and that workers can buy is much larger than the number of cars produced by public enterprises (see Figure 22). The reason is that, while there is a pressure on the part of owner/entrepreneur robots for increasing the number of cars produced by their enterprise because it is this number that determines their personal wealth and their reproductive success, this pressure does not exist for public enterprises.

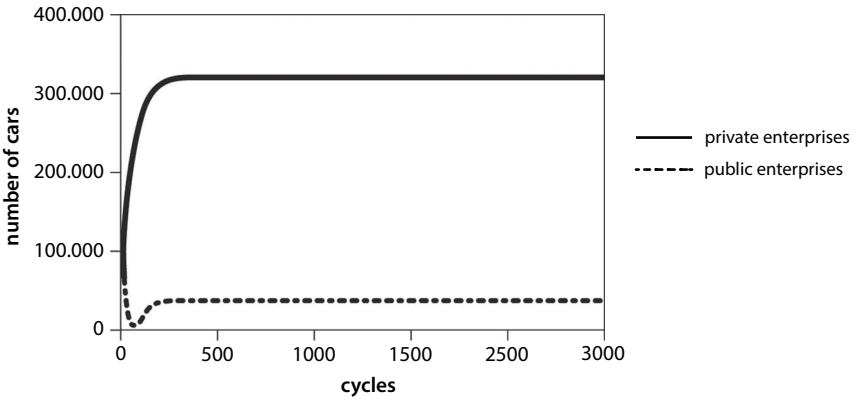


Figure 22. Number of cars produced by private and public enterprises in a period of 3000 cycles.

Another advantage of private enterprises is that they provide the means of living to a greater number of workers. At the beginning of the simulation, the number of workers is always 1000. At the end, this number is more than 3000 with private enterprises while it is only 300 with public enterprises. Private enterprises tend to employ more workers because they want to produce and sell more cars. Public enterprises are only interested in producing a given number of cars because they do not have owners that want to gain more by producing more cars and they do not compete with other, public or private, enterprises.

So, what our robots tell us is that, while private enterprises must necessarily produce their goods efficiently, this is not true for public enterprises and, more generally, for states. What is produced by a private enterprise must make more money when it is sold than the money needed to produce it. Otherwise, the enterprise closes. States are different. States may not perform very efficiently because they cannot close and because there is no one who is *personally* interested in the effectiveness of their performance. So the problem of their performance exists and robots should help us to understand this problem and to find new solutions.

8. The increase in the number of things that are goods for human beings

As we have said at the beginning of this Chapter, many more things are goods for human beings compared to nonhuman animals. Nonhuman animals want to have very few things, these things are strictly linked to their capacity to survive and reproduce, and they are more or less always the same. In contrast, the things which are goods for human beings are very indirectly linked to their survival and reproduction and, in the course of human history, their number have progressively increased to arrive to today's explosion. The robots described in this and other chapters reproduce some of the historical changes that have led to this increase in the number of things which are goods for human beings and, in this last section, we summarize these historical changes.

(1) Production of goods

The things which are goods for nonhuman animals are found in nature and, therefore, they are limited to what exists in nature. Almost everything which is a good for a human being does not exist in nature but it is produced by other human beings. And, since human beings produce new goods by using the already existing goods, this progressively increases the number of things which are goods for them. In Chapter 7 we have described robots that use tools to produce new food and in this Chapter robots that use tools to increase the quantity of energy extracted from food but, in both cases, the tools already exist and the robots must only discover their usefulness so that they become goods for them – things that they try to have. In Chapter 8 we have described robots that construct new vases by copying the existing vases. Therefore, vases are goods that exist only because the robots produce them. Vases are very ancient artefacts. Today, the constant progresses of technology lead to a constant increase in the number of goods that are produced by human beings.

(2) Specialization and exchange of goods

Nonhuman animals directly obtain their goods through their work and, therefore, the things which are goods for them are limited by their time and skills. Human beings obtain most of their goods from other human beings and this makes it possible to specialize in the production of one good and to obtain all the other goods from other human beings (Section 3). Therefore, the exchange of goods is another factor which greatly increases the number of things which are goods for human beings.

(3) Money

Barter, the exchange of any good for any other good, is only possible if the two individuals who are trying to make an exchange have complementary needs. If A needs a good which is owned by B, and B needs a good which is owned by A, the exchange can take place. The problem with barter is that, as the number of different goods increases, this condition of complementary needs is more difficult to satisfy and, therefore, the

exchanges are necessarily limited. Money solves this problem because money is automatically accepted in all exchanges (Section 5). Therefore, money is another very powerful mechanism that increases the quantity and variety of goods because it facilitates the exchange of goods.

(4) Private enterprises and states

Another factor which increases the number and variety of things which are goods for human beings is the manner in which goods are produced. Among nonhuman animals, each individual works alone to procure the goods it needs and, therefore, the goods of nonhuman animals are limited to the sum of the goods procured by each individual. Human beings create organizations that produce more goods than the sum of the goods that would be produced by the members of the organization if they were to work alone (Section 7). This is true for both private enterprises and the state. An entrepreneur gives money to his or her workers in exchange for their work and this allows the entrepreneur to produce a great number and variety of different goods. The members of a community give some of their goods (pay taxes) to a central store which uses these goods to produce new goods which are then distributed to all the members of the community.

(5) Financial economy

In a market economy everything can be obtained in exchange for money but, if one does not have the money which is necessary to buy X, he or she cannot have X. The problem is solved by creating another type of good: borrowed money. A buys borrowed money from B and it pays this borrowed money by giving some of its money to B (interest). Borrowing money is a good so important in modern human economies that these economies are almost completely based on borrowed money rather than on owned money. Private enterprises borrow money to buy what is necessary to produce the goods they produce. The wages or salaries that the entrepreneur robots described in Section 6 pay to worker robots are only one part of the total costs of production. Other costs are the costs of raw materials and production technologies, organizational costs, research costs, distribution costs, and marketing costs. If an entrepreneur robot does not have enough money to pay for all these costs, the entrepreneur robot has two options. In can either sell parts of its enterprise and share its gains with the other owners of the enterprise or it can borrow money from organizations of robots which are specialized in lending money – banks and other financial enterprises. This increases its capacity to produce goods and to produce always new goods.

Not only entrepreneur robots but also buyer robots and the central store itself – the state – can borrow money. If a robot wants to buy some good – say, a house – but it does not have the money necessary to buy the house, the robot can borrow money from a bank and buy the house. If taxes do not provide the state with sufficient money to

produce all the goods it wants to produce and distribute to its citizens, the state borrows money from individual robots or from financial enterprises. (The state can also "print" new money but this leads to the devaluation of money – a reduction of what one can buy with a given amount of money – and this can damage the entire economy.) Even banks and other financial enterprises borrow money but they borrow money not to produce real goods but to lend this money to private enterprises, to buyers, to the state, and to other banks. They are like the merchant robots described in Section 4 which exchange black food for white food and white food for black food. The only difference is that banks and other financial enterprises exchange money for money.

But the financial economy goes beyond the borrowing of money. The money which is borrowed must be returned to the lender at a time agreed upon between the lender and the borrower, and the borrower may be unable to give the money back to the lender at the agreed upon time. So, lending money implies the risk for the lender not to get its money back from the borrower. The financial economy solves this problem by creating another good: risk coverage. Lenders of money buy risk coverage from an insurance company and this means that, if the lender does not receive back the borrowed money from the borrower at the established time, it is the insurance company that gives the money back to the lender. Risk coverage is a good which exists independently from the borrowing of money. Many things can happen to human beings which are bad for them. Dying, being robed, having an accident or an illness, are all examples of bad things, and risk is the probability that one of these things actually happens. Risk coverage means that, if one of these bad things happens, the individual receives some money from an insurance company. But the individual must buy risk coverage from the insurance company. The individual gives some of its money to an insurance company and the insurance company gives risk coverage to the individual. Risk coverage exists independently from the borrowing of money but lending money inherently creates the risk that the money will not be given back at the established time and, therefore, the lender buys risk coverage from some insurance company. And, given the importance of borrowed money in modern human economies, risk coverage for borrowed money is another important good which is bought and sold in modern human economies.

But risk coverage is not one good among others but it is the very essence of the financial economy. The financial economy creates a new type of goods which, like money, can only be exchanged but not used or consumed: financial goods. A financial good is a document which attests that the owner of the document is the owner of one part of an enterprise or has lent money to an enterprise or to the state or has sold risk coverage to someone else. These financial goods are sold and bought like real goods. If A sells a financial good to B, B becomes the owner of that part of the enterprise or the lender of money to the enterprise or to the state or the seller of risk coverage to someone else. The financial economy mostly consists in selling and buying financial goods in what is called the financial market.

But financial goods are different from real goods. All goods which are sold and bought can change their value – measured in money units – and this value mainly depends on how many people sell and buy the goods. But the value of real goods – for example, the apples of Section 6 – does not change very much and does not change frequently because real goods satisfy real needs which have an intrinsic and rather constant strength. The value of financial goods changes much more and much more frequently because the number of people who want to sell and buy financial goods change – literally – from moment to moment. This is why financial markets are more unpredictable and "volatile" than economic markets and their unpredictability and "volatility" are increased by modern communication technologies which make it possible to immediately know if and how the value of a financial good has changed. The result is that the financial economy creates another financial good: coverage of risk in the case that a financial good loses its value. Coverage of risk for financial goods multiplies *ad infinitum* the number of financial goods, and most of the economy becomes an economy not of goods that are used and consumed but of financial goods that can only be sold and bought.

The science of economics tends to assimilate the exchange of financial goods to the exchange of real goods because it was born when real goods were more important than financial goods. But the financial economy is different from the real economy and it is much more difficult to understand and to control. Future robots should be able to create and exchange financial goods so as to help us to better understand how the financial economy is different from the real economy and what are the consequences of its increasing importance for human beings.

(6) Marketing

But what really explains the great number of things which are goods for today's human beings – according to some estimates, this number is 10 billion different goods – is another phenomenon which characterizes today's societies: marketing. Marketing is based on a mechanism of social life which we have already discussed in Chapter 6: if you want to make another individual behave in ways which satisfy your motivations, change the other individual's motivations. Marketing is an application of this fundamental mechanism of social life to an economy based on private enterprises and it transforms a *market* economy into a *marketing* economy. In a market economy sellers sell things which buyers already want to have and therefore are already goods for them. In a marketing economy sellers change the motivations of buyers so that buyers become motivated to buy the things the sellers want to sell. Since marketing relies on what is perhaps the basic mechanism of social life, marketing transforms a market economy into a marketing economy, and a marketing economy transforms the entire society into a marketing society: a society in which *everything* which is done, is done to be sold.

Marketing reveals an important aspect of the exchange of goods for money. Human beings do all sorts of things to increase the strength of the motivation of other individuals to have what they have to offer. If the exchange is of one real good for another real good (barter), the situation is symmetrical. Both participants in the exchange try to convince the other individual of the goodness of what they have to offer and both participants have the some power with respect to each other. But when goods are exchanged for money, the relation between sellers and buyers ceases to be symmetrical and sellers have more power than buyers. Sellers do all sorts of things to create or strengthen in buyers the motivation to buy what they sell. They spend a lot of money to invent always new goods, to inform buyers of the existence of these goods, to associate their goods with something that buyers already want to have, and to acquire all sorts of knowledge about buyers so as to know who they are and what they want. Buyers cannot do anything of this sort because they do not know very much about sellers and because the motivation of sellers to have money already exists, is always activated, and buyers cannot make it stronger. The only thing that buyers can do is not to buy what sellers offer but marketing makes this very difficult for them to do. It is this asymmetry between sellers and buyers which explains why today's human economies are marketing, not market, economies.

This poses an interesting question. Is finance or marketing the most important factor which explains the great increase in the number of things which human beings want to have – in economically developed societies and, in perspective, in all human societies? Borrowing money to produce more goods and always new goods is certainly important but the goods which are produced must be sold, and people buy so many things because marketing convinces them to buy all these things. So, the answer to our question is that marketing is more important than finance. Economists study both the real and the financial economy – although, as we have said, they understand the real economy better than the financial economy – but they completely ignore marketing, and the only people who know what is marketing are those that study marketing to become marketing professionals. Economists think that their social role is to advice the society on how to increase its gross domestic product – the total number of goods which are produced. The gross domestic product depends on the real economy but, in economically advanced societies, it increasingly depends on the financial economy and, even more, on marketing. No economist really wants to limit the financial economy but no one – economist or non-economist – even *thinks* of limiting marketing.

Economists – and everyone else – take marketing as a given because they know that marketing is a crucial factor for increasing the gross domestic product. But economists cannot understand marketing because marketing is manipulating the motivations of buyers, and economists do not know anything about motivations and how an individual's motivations change during his or her life as a function of the individual's experiences. Economists assume a fixed set of motivations, each with a fixed strength,

and they are only interested in what an individual does to satisfy "rationally" these motivations. But the behaviour of human beings is determined by their motivations more than by their cognitive skills, and this is why the science of economics cannot understand modern economies.

We have not constructed robots that sell and buy financial goods and do marketing to convince other robots to buy the goods they sell. But, given the importance of these phenomena in today's human societies, constructing robots which have a financial economy and do marketing is one of the most important tasks of a robotic science of human beings.

CHAPTER 12

Individually different robots and robots with pathologies

When ME looks at human beings, it sees that no two human beings are identical and it thinks that knowing how they are different is as important as knowing what they have in common. Therefore, to be like human beings ME's robots must all be different and they must be different from various points of view. They must be different in their body and in their brain, they must possess different capacities, they must possess different levels of the same capacity, they must be different in their motivations, and they must have different temperaments and personalities. And also the causes of these differences are important for ME. ME wants to construct human robots which are different because their genes are different and robots which are different because the experiences they have had during their life have made them different.

Another thing that ME discovers about human beings is that some of them, some of the time, have pathologies, which are states of their body or their mind that do not allow them to live a normal life, make them feel bad and have fears and, in some cases, may imperil their life. Therefore, to understand human beings, ME thinks that it must not only construct robots that are like human beings but these robots must also have the same pathologies of human beings because these pathologies can reveal much about human beings. ME is interested in pathologies of behaviour and mind, although it knows that human beings are bodies and, therefore, all their pathologies are pathologies of their body. ME also finds that human beings do not simply accept that they have pathologies but they develop all sorts of methods and technologies to cure these pathologies. Therefore, ME wants to construct robotic "doctors" that use these methods and technologies to cure ill robots.

1. Current robotics ignores inter-individual differences

When we look at real animals, we see that each individual is different and behaves differently from all the other members of the same species and this implies that, if we want to construct robots that are like real animals, no two robots can be identical and behave in the same way. This is not true for today's robotics. The goal of today's

robotics is to construct robots which do specific things which are useful to us. Once we have constructed a robot which exhibits the behaviour we are interested in, we are satisfied. Current robotics is an applied, engineering discipline. If robots must help us to reach our practical goals, as we do for all industrial artefacts we design a prototype robot and then we produce a number of identical copies of the prototype robot. But if robots must help us to explain the behaviour of real animals, this is not the approach to take because, in real animals, each individual is different from all other individuals of the same species and we can learn a lot about the behaviour of animals from studying how an individual is different from all other individuals. We must construct a multiplicity of robots which all belong to the same "species" but are each different from all others and we must ask: What are the different forms in which inter-individual variation manifests itself? What are its causes? What are its consequences? Why one particular characteristic of behaviour tends to be accompanied by another specific characteristic? Constructing individually different robots does not make much sense for robots as technology because individually different robots would be unpredictable and unreliable and no one would buy an unpredictable and unreliable car. But for robots as science not two robots can be identical and studying their differences is as important as studying what they have in common.

Evolutionary robotics works with populations of individually different robots and the differences among the robots are important because evolution depends on the selective reproduction of the best robots and the addition of random variation to the robots' offspring which makes the offspring robots different from both their parents and their siblings. But, although inter-individual variation is necessary for evolutionary robotics, evolutionary robotics is not really interested in inter-individual differences. For evolutionary robotics inter-individual variation is only a tool for constructing robots that possess desired behaviours. The differences among the robots are not studied as an intrinsically important property of the robots and as revealing important phenomena to be analyzed and explained. Evolutionary robotics is interested in constructing one *single* robot: the best individual of the last generation. What is the behaviour of the *other* robots of the same generation or of any other generation is not a topic of research for current evolutionary robotics.

Inter-individual differences are also ignored by those who construct robots that learn. A robot begins its learning with random connection weights in its neural network and, therefore, each robot learns in different ways and at the end each robot inevitably behaves differently. But this variation is considered as "random noise" and, to limit the effects of this "random noise", the researcher makes a certain number of robots with different initial connection weights in their neural network learn the same task. What the researcher is interested in is the average performance of the robots, not the different ways in which the robots learn and the different behaviours of the different robots at the end of learning.

Inter-individual differences are also ignored in collective robotics. Collective robotics constructs groups of robots which, by coordinating and communicating with each other, accomplish tasks that no single robot would be able to accomplish by acting alone. But the robots tend to be all identical and, if they behave differently, this is due to external circumstances, not to the different characteristics of the robots. In some cases, the researcher assigns different roles to the robots that make up the group but what we are talking about here are different roles spontaneously chosen by the robots and differences in how different robots play the same role.

So, the study of inter-individual differences in behaviour using robots is a task for the future. Researchers working with robots should learn from psychologists. For many psychologists – especially experimental psychologists – inter-individual differences are only "random noise" which they try to get rid of by computing statistical averages but there are many areas of psychology which are specifically interested in inter-individual differences. Some psychologists study inter-individual differences in ability – one individual possesses more of some ability than another individual – and difference in temperament or personality – one individual has a different personality than another individual. And both ability and personality are not single dimensions but they are multidimensional. Even intelligence is not considered by psychologists as a single dimension which can be measured by an intelligence test, but there are different types of intelligence – linguistic, spatial, mathematical, practical, social, emotional intelligence – and an individual can be better than another individual in one type of intelligence and less good in another type of intelligence. Temperament and personality are more difficult to define and identify but it is under the eyes of everybody that each individual has his or her own temperament and his or her personality.

Inter-individual differences are very important for social psychologists and this makes sense because for human beings the differences among other human beings are more important than their similarities. Social life is social interactions and, if an individual is free to choose with whom to interact, the individual will choose an individual with specific characteristics and the same will do the other individual. If two individuals need to interact together – for example, because they work together – they will be more or less happy to interact together because of their respective characters. People become friends because they like each other's character. For lovers beauty plays an important role because it is associated with sexual selection but there are different types of beauty and these different types of beauty signal different characters. The male and female robots described in Chapter 7 do not choose a mate because all male robots are identical and the same for the female robots. But real animals are all inter-individually different and, especially for human beings, character and personality play an important role in the choice of the partner and in the maintenance of a sexual relationship. Inter-individually different robots should help to understand this role and to predict when two robots will become friends or permanent sexual partners.

Another sub-discipline of psychology, clinical psychology, is specifically interested in individually unique human beings and it studies what makes each individual unique. The clinical approach is especially important for psychological disturbances and psychiatric illnesses and for their treatment. In psychiatry, as in all of medicine, each patient is, or should be, unique. (For robots which have pathologies, see later in this chapter.)

2. Fitness is not enough

If we want to study inter-individual differences in behaviour by constructing robots, we must construct not single robots but many individually different robots. This is what is done in evolutionary robotics. In evolutionary robotics one starts with a population of robots which have random genes and, since a robot's genes determine the robot's behaviour, each robot behaves differently from all other robots. The robots reproduce selectively based on some criterion of fitness and the offspring robots inherit the genes of their parents with the addition of random variations – and this guarantees that the robots remain inter-individually different in all successive generations. If we look at the robots of the last generation, we find that they have different fitnesses and, therefore, they are inter-individually different. Two robots can have different fitnesses not because they are different but because they live in different parts of environments, one with more food and the other with less food. But human beings change the environment in which they live and this means that, for human beings, differences in fitness (wealth) generally are differences in ability or character. A human robot which lives in one part of the environment with little food might develop agriculture and produce new food – if it has the necessary abilities and the necessary character.

But, in any case, to understand inter-individual differences in behaviour, differences in fitness are not enough. We must ask: Do robots which have the *same* fitness behave in the same way? Why a robot has more fitness than another robot? How is the behaviour of robots which have different fitnesses?

To answer these questions, we return to our basic robots that evolve in an environment with randomly distributed food tokens. The fitness criterion is the number of food tokens eaten during a life which has the same length for all robots. The robots of the first generation eat very few food tokens because they have random connection weights in their neural network and are unable to respond to the sight of a food token by approaching and reaching the food token. Then, in a succession of generations, their fitness progressively increases until it reaches a stable value. The robots have evolved the capacity to eat the food tokens. At this point we stop the evolutionary process and we examine the robots of the last generation. Each robot has a fitness which tends to be different from the fitness of all other robots. This clearly is an important dimension of inter-individual variation and, in fact, is the dimension of inter-individual variation

which makes evolution possible. But if we are interested in inter-individual differences in behaviour, fitness does not reveal much about the behaviour of the robots and we must go beyond fitness. Two robots can have the same fitness but they may not behave in the same way. This can already be seen if we look at how the robots behave in their natural environment but the differences in behaviour can be more precisely measured if we bring the robots into an experimental laboratory in which everything is controlled by us.

In its natural environment a robot can find itself in two different situations and it must respond appropriately to both situations. When it sees a food token, the robot must respond by approaching and reaching the food token. But in some occasions the food tokens are so distant that the robot cannot see them. In this situation the robot must equally behave appropriately: it must explore the environment in order to find the food tokens. Approaching and reaching the food tokens and exploring the environment are two different abilities that can be measured with precision in the experimental laboratory. To measure the first ability – the ability to reach the food tokens – we put a robot in an environment in which there is a single food token that is sufficiently near to the robot that the robot can see it and we count the number of cycles taken by the robot to reach the food token. To measure the second ability – the ability to explore the environment – we put the robot into an empty environment and we count the number of different parts of the environment – number of different pixels of the computer's screen – visited by the robot. The ability to approach and reach the food tokens and the ability to explore the environment are two different abilities. Two robots can have the same fitness but different abilities. One robot reaches a food token in little time when it sees the food token but it does not explore very much the environment when there are no food tokens in sight. The other robot takes more time to reach a food token when it sees it but the robot is better at exploring the environment when it does not see any food token. The two robots have the same fitness but their behaviour and their abilities are different.

This also applies to robots that have different fitnesses. One robot can have more fitness than another robot because it is better at reaching the food tokens or because it is better at exploring the environment or because it has more of both capacities. And the inter-individual differences can be more subtle. Two robots may take the same time to reach a food token but the two robots behave differently. One robot goes straight towards the food token but it is rather slow in its movements while the other robot follows a more indirect trajectory but it moves faster. The two robots reach the food token in the same time but their behaviours are different.

Another interesting question is if the particular environment in which the robots live influences the different abilities of the robots. To answer this question we do an ecological experiment by making two different populations of robots live in two different environments. In one environment, when a robot eats a food token, the food token disappears and another food token appears in a randomly chosen position so

that the total number of food tokens is always the same. The other environment is a seasonal environment in which, when a food token is eaten, the total number of food tokens present in the environment is reduced and the food tokens re-appear only at the beginning of the next season. In this second environment, while the ability to reach the food tokens continues to be important, the exploratory ability of the robots is more important because it is more probable that a robot does not to see any food token when, at the end of a season, there are very few food tokens in the environment.

These robots clearly show that fitness depends on different capacities and these different capacities can be isolated and quantitatively measured. But there may also be *qualitative* differences in behaviour – styles of behaviour – which are not reflected in fitness. For example, one robot can approach the food tokens from the right while another robot approaches the food tokens from the left. This, by itself, may not result in a difference in fitness – in the number of food tokens eaten – but it can equally be an interesting difference between the two robots. Another example concerns the wall that closes the environment. When a robot hits the wall, it bounces back in a randomly chosen direction or in a direction which depends on the direction from which it has hit the wall. If the robots have infrared neurons which tell them when they are near to the wall, two robots can have the same fitness but they have two different *styles* of behaviour. One robot tends to move in the environment without touching the wall while the other robot has a behavioural strategy which includes bumping against the wall and bouncing back. Notwithstanding these differences in behaviour, the two robots eat the same number of food tokens and therefore they have the same fitness. But if we only look at the robots' fitness, we miss something interesting concerning their behaviour.

Differences in behaviour which do not translate into differences in fitness may be important because they may play a role if the environment changes and the robots must adapt to a new environment. In the original environment touching the wall has no fitness consequences but in the new environment touching the wall reduces a robot's fitness because the environment has been invaded by dangerous insects that live on the wall. In the old environment the fitness of a robot depends only on the number of food tokens that the robot is able to eat during its life. In the new environment the fitness of a robot depends on two capacities: the capacity to eat and the capacity to avoid touching the wall. Two robots that in the old environment had the *same* fitness may *not* have the same fitness in the new environment. The robots that, in the old environment, had the behavioural strategy of never touching the wall will have more fitness than the robots with the behavioural strategy of bumping against the wall and bouncing back. This is an example of what biologists call pre-adaptation: the existence of one behaviour or capacity which in one environment has no role in determining the fitness of an individual but, when the environment changes, it acquires this role. The robots with the behavioural strategy of never touching the wall are pre-adapted to the new environment.

If the environment in which the robots live and their adaptive pattern are more complex – and both the environment and the adaptive pattern of even very simple animals are rather complex – fitness tells us less and less about behaviour. Remember the robots described in Chapter 2 which have two motivations to satisfy to remain alive and reproduce and which, therefore, must possess two capacities: they must be able to eat and to drink, to eat and to avoid being killed by the predator, to eat and to mate, to eat and to feed their offspring, to eat and not to move when their body is physically damaged. If we know that two robots have the same fitness, this tells us very little about how the two robots behave. Take the robots that must both eat and drink to remain alive. One robot is better at eating than at drinking and another robot is better at drinking than at eating, and this can be measured by bringing the two robots into an experimental laboratory, exposing them to a food token or a water token, and measuring the time taken by the two robots to reach the two different types of tokens. Provided the two robots eat and drink sufficiently to remain alive, they may have the same length of life and, therefore, the same fitness, but the two robots do not behave in the same way and do not possess the same abilities. And the same applies to the other robots. One robot can be good at eating but it is not very good at flying away from the predator while another robot does not eat much when the predator is absent but it is better at flying away from the predator when the predator arrives. One robot eats much and therefore has a long life but it is not very good at finding a mating partner, while another robot does not eat much and dies young but until it succeeds in remaining alive it mates often and has many offspring. And the same for the robots which have to both eat and feed their offspring. One robot can eat much and therefore it has a long life and generates many offspring but it does not feed its offspring which therefore die before reaching sexual maturity. Another robot eats less and therefore does not live much but it feeds its offspring and, therefore, the few offspring that the robot is able to generate during its short life will more likely reach the age in which they can reproduce. And, finally, for the robots that must stop moving when their body is physically damaged, one robot eats much but it continues to move when it feels pain, while another robot does the opposite: it is not very good at eating but it is very careful at not moving when it feels pain. In all these cases, knowing the fitness of a robot does not tell us much about the behaviour of the robot. We must actually observe and measure the robot's behaviour in its natural environment or in an experimental laboratory because only in this way we can know how the robot actually behaves and what are the differences in the behaviour of different robots.

In Chapter 7 we have seen that male and female robots behave differently. Male robots behave differently from female robots because males are more attracted by females than females by males. But we also find inter-individual differences among the male robots. Although all male robots must both eat and mate if they want to survive and leave offspring, one male robot is better at eating while another male robot is better at approaching and mating with female robots. When they are reproductive, all

females robots tend not to move much but some female robots move more than other female robots and these female robots are less attractive for males but they eat more.

Once we have distinguished among different capacities and different styles of behaviour, we can ask which one of these capacities or styles of behaviour is more important from the point of view of fitness. Are the robots which have more fitness those which are better at approaching and reaching the food tokens when they see them or those that are better at exploring the environment when they do not see any food token? The robots which are better at eating the food tokens or the robots which are better at not touching the wall with dangerous insects? The robots which are better at eating the food tokens or the robots which are better at flying away from the predator? The male robots which are much attracted by reproductive females but are not particularly interested in food or the male robots which are somewhat less attracted by female robots but eat more and live longer? How important is the speed of movement for a robot's fitness? A robot with little fitness may move very fast without being very much attracted by the food tokens while another robot with the same little fitness, although it is attracted by the food tokens, moves very slowly and therefore it also does not eat much.

In evolutionary robotics one decides a criterion of fitness – for example, the total number of food tokens eaten by a robot during a life which has the same length for all robots – and then the "curve of fitness" shows how the fitness of the robots increases in a succession of generations. But robots must tell us how behaviour, not only fitness, evolves. If the fitness of a robot depends on different capacities – for example, the capacity to reach a food token when the robot sees the food token and the capacity to explore the environment when the robot does not see any food token – we must look not only at how fitness evolves but also at how the robots' different capacities evolve. For example, given the same curve of fitness, the capacity to explore the environment may evolve earlier than the capacity to reach the food tokens – and this can be determined by measuring in an experimental laboratory the two capacities in the robots of successive generations.

To study how behaviour, not fitness, evolves, we construct robots that live in an environment with randomly distributed food tokens but the environment includes a zone in which lives a predator (see Figure 1). The robots know if they are inside or outside the predator zone because they have a smell neuron which is activated when they enter the predator zone.

The problem is that the food tokens which are inside the predator zone contain more energy than the food tokens which are outside the predator zone. Therefore, the robots are exposed to a dilemma. Should they enter the predator zone to eat the better food contained in the predator zone – and risk being killed by the predator – or should they ignore the predator zone and eat only the less energetic food which is outside the predator zone? The curve of fitness for these robots is the usual curve which goes up in the early generations and then it reaches a stable level. But this curve tells very little

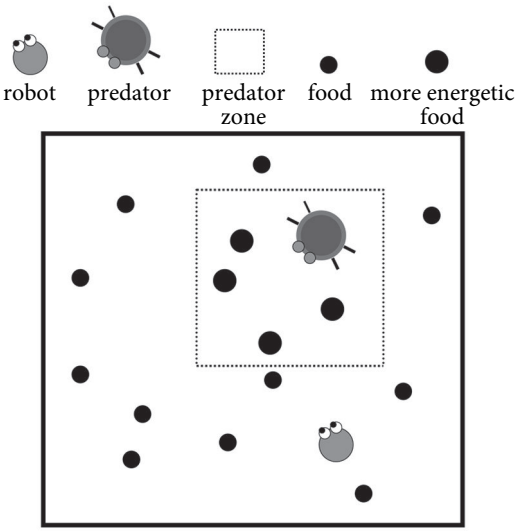


Figure 1. The robots live in an environment with randomly distributed food tokens and a zone in which the food tokens contain more energy but in that zone lives a predator that can reach and kill the robots which penetrate into the predator's zone.

about the behaviour of the robots. If we look at the robots at the end of evolution (after 1000 generations) we find that the robots have developed the more prudent strategy of avoiding the predator zone and eating only the less energetic food tokens that are outside the predator zone. But in the first 200 generations the behaviour of entering the predator zone is much more frequent and it disappears only later in evolution. This change in the behaviour of the robots is manifested in their ability to flee away from the predator. When we test the robots of the successive generations in an experimental laboratory in which there is no food but only the predator, we find the robots of the first 200 generations are much better than the robots of the later generations at fleeing away from the predator, and this is explained by the fact that the strategy of entering the predator zone creates a pressure on the robots of the first 200 generations to evolve the capacity to flee away from the predator. The robots of the later generations do not enter the predator zone and, therefore, this pressure does not exist any more for them and, as a consequence, in the experimental laboratory they are not very good at fleeing away from the predator. These robots clearly show that, to understand how behaviour evolves and why it evolves as it evolves, it is necessary to go beyond fitness and to examine the actual behaviour of the robots of the successive generations in their natural environment and in the experimental laboratory.

These are only some of the inter-individual differences that can exist among our robots. Consider the robots which are able to predict the consequences of their actions described in Chapter 5. One robot may be better than another robot at predicting the

consequences of its actions and at using these predictions to decide what to do while another robot is less good at make predictions but it is much better than the first robot at doing what it does. This also applies to imaginations and memories. One robot can have a richer mental life than another robot, with more imaginings and more memories of past events than another robot. Or, given the same experiences during their life, a robot can be happier than another robot because it tends to imagine good future events and to remember good past events while the less happy robot tends to imagine bad future events and to remember bad past events.

3. Motivations as a source of inter-individual differences

Motivations are a rich source of inter-individual differences. Most of the inter-individual differences we have discussed so far are differences in the capacities of the robots. Robot A and robot B have the same fitness but they are different because robot A is better than robot B at doing one thing and robot B is better than robot A at doing another thing. But, as we know from Chapter 2, a robot's behaviour is determined not only by its capacities but also by its motivations. Two robots can be equally good at fleeing away from a predator – and this can be demonstrated by testing the two robots in an experimental laboratory – but the motivation to eat is intrinsically stronger in one robot than in the other robot. Therefore, the first robot ceases to look for food and it flees away from the predator only when the predator is very near while the second robot immediately flees away from the predator as soon as the predator appears. This shows that inter-individual differences can be difference in motivations. Two robots can be motivationally different because they have different motivations or because they have the same motivations but the strength of their motivations is different. In both cases these differences will be reflected in their behaviour.

As we have said in Chapter 2, the presence or absence of a motivation is determined by looking at the robot's behaviour and seeing if the robot does something to satisfy the motivation. The motivations of our robots are all necessary to survive and reproduce and, therefore, they must be necessarily present in all individuals. But two robots which have the same motivations can be different because the intrinsic strength of their motivations is different. Motivations do not have an absolute strength but the strength of one motivation can be measured only with respect to the strength of the other motivations. We can measure the relative strength of two different motivations by offering to a robot a choice between two different actions which satisfy two different motivations, and by determining what is the action chosen by the robot. For example, a robot which has both the motivation to eat and the motivation to drink is put in front of a food token and a water token and we see if the robot approaches and reaches the food token or the water token. (On this, see Chapter 2.) If everything is equal – the distance of the two tokens from the robot and the current quantity of energy and water

in the robot's body of which the robot's brain in informed under the form of hunger and thirst – if the robot approaches the food token we can say that the intrinsic strength of the motivation to eat is greater in the robot than the strength of the motivation to drink. Another robot can do the opposite, showing that the motivation to drink is stronger than the motivation to eat in this robot. (The strength of a motivation is the same thing as the value of a good discussed in the chapter on robotic economies.)

The same applies to the male and female robots described in Chapter 7. Two male robots can both prefer mating to eating but, when they see a female robot, one male robot completely ignores food unless it is very hungry while the other robot's gives more value to food and it is more hesitant to change its behaviour. And something similar happens with the female robots. When they are reproductive, female robots tend not to move much and to wait for male robots to approach them and mate with them. But one reproductive female can more attracted by food while another reproductive female tends to completely ignore food and it almost completely stops moving so that it can be more easily reached by a male robot.

Inter-individual differences in motivations are differences in character or personality. But differences in character and personality can consist not only in the fact that for one robot one motivation is stronger than another motivation while the opposite is true for another robot. One robot can have *generally* stronger motivations than another robot. For one robot some things have much more value than other things while for another robot the value of all things is more or less the same. Or the character of a robot can be defined by the ensemble of its motivations, each with its specific strength.

Robots must reproduce both what psychologists call inter-individual differences in capacities and skills and what they call inter-individual differences in character or personality. Differences in capacities and skills mostly involve the cognitive half of the mind, while differences in character or personality mostly involve its motivational/emotional half. And this is another reason for constructing robots that have both halves of the mind.

4. Learning and inter-individual differences

Most of the robots described in this book have completely innate behaviours. A robot's inherited genes specify the connection weights of the robot's neural network and, since these connection weights do not change during the robot's life, a robot behaves in the same way during its entire life. But the behaviour of no animals is entirely innate. All animals learn and, through learning, not only their capacities but also their motivations change. Learning improves the capacities of an animal and adds new capacities but it also changes the strength of the animal's motivations and adds new motivations. Two robots which, without learning, would have the same capacities and the same motivations, with learning may become different in both their capacities and their

motivations. A robot can learn to approach a new type of food token while another robot does not learn this behaviour and, in this way, the two robots, which had the same capacities and were, from this point of view identical, become different. We have described robots that learn new capacities in Chapter 3 and we have mentioned robots that acquire new motivations in various places in this book. Here is a simple example of learning a new motivation. A robot discovers that one particular token which is not food can make the procurement of food easier and, as a consequence, a new motivation to have the token is acquired by the robot and it determines the robot's behaviour. (In human beings, this type of tokens – tools – are not "found" but they are "constructed".) The motivation to have the token is a learned motivation which is acquired because of the genetically inherited motivation to eat.

Learning increases inter-individual differences in that it creates a larger set of possibilities for the existence of inter-individual differences. If two robots possess only one capacity, they can only differ in the degree in which they possess that capacity. But if the robots learn new capacities, the possible inter-individual differences among the robots increase in number. And the same is true for motivations. If a robot acquires the motivation to look for tools, the robot possesses both the motivation to look for food and the motivation to look for tools, while another robot has only the motivation to look for food – or the second robot may acquire the motivation to obtain the tools from the first robot in exchange for food. (For robots that make exchanges, see Chapter 11 dedicated robotic economies.)

Learning new behaviours and new motivations is a rich source of inter-individual differences. Two robots can have the same innate behaviours and the same innate motivations but, because of the different experiences they have during their life, they develop different behaviours and different motivations as ways to satisfy their innate motivations.

Another important consequence of learning for inter-individual differences has to do with culture. Human beings learn most of their behaviours by imitating other human beings and two robots may behave differently because they live in different cultures and each robot imitates the robots of its culture. In this way, learning creates not only inter-individual but also intercultural differences in behaviour. (For robots that have culture, see Chapter 8.)

Inter-individual differences – both genetic and learned – play a critical role in the social life of human beings because the manner in which two individuals interact together – and even *if* they interact together – depends on their individual characteristics. Inter-individual differences can be differences in behaviour or differences in physical aspect. In Chapter 10 dedicated to robotic politics we have seen that one robot may be followed by the other robots with which it lives because it explores more widely the environment searching for preys. But inter-individual differences are not only differences in behaviour but also differences in physical aspect. One robot behaves differently from all the other robots but they all look identical because they all have

the same body. Future robots will need not only to behave differently but also to have individually different bodies so that one robot may "know" another robot "personally". Inter-individual and inter-cultural differences – and also age-related and sexual differences – are so important in the social life of human beings that robots which are physically and behaviourally different may tell us more about human beings than robots which are just robots.

5. Why construct robots that have pathologies?

An individual which has some pathology is different from a healthy individual – and this explains why we talk of robots with pathologies in the same chapter in which we talk of robots which are inter-individual different. Animals and human beings have all sorts of pathologies and, therefore, if our robots must be like real animals and real human beings, our robots should have all sorts of pathologies. Again, this follows from our principle "one robot/many phenomena". One and the same robot can be healthy or ill and its illness can have different characteristics and it may the consequence of different causes. But the manner in which one robot is ill depends on who is the particular robot.

Pathologies are conditions which, more or less directly and more or less seriously, impair an individual's survival and reproductive chances and reduce the individual's well-being. Pathologies can be pathologies of the body or pathologies of behaviour. Examples of pathologies of the body are bone fractures, heart diseases, and stomach ulcers. Examples of pathologies of behaviour are blindness, aphasia, unilateral neglect, major depression, schizophrenia, and the pathologies associated with an ageing brain. We are primarily interested in pathologies of behaviour but robots are theories of behaviour based on the premise that behaviour can only be explained by taking the body and the brain into consideration and, since bodies and brains are physical entities, pathologies of behaviour are necessarily pathologies of the body. We must construct robots that exhibit pathologies of behaviour but we must also reproduce the physical or chemical substratum of these pathologies. This is another application of the principle "one robot/many phenomena". Robots that reproduce both behavioural pathologies and their physical and chemical substratum are to be preferred to robots that reproduce only behavioural pathologies. If we "lesion" one part of a robot's neural network and we obtain one type of pathological behaviour, we should ask: Is there a correspondence between the locus of lesion in the robot's neural network and the locus of the lesion which in the real brain is associated with that pathological behaviour? Are the causes of neurological pathologies different from the causes of psychiatric pathologies? Psychiatric and psychological pathologies can be the result of specific experiences or they can have a genetic origin – and, in most cases, they are both the result of specific experiences and of the individual's genes. Can we identify the origin of psychiatric and psychological pathologies by manipulating a robot's genes and the robot's experiences?

If one constructs robots which have practical applications, a pathological robotics does not make sense. Why construct robots with behavioural pathologies that can only detract from their practical usefulness? A pathological robotics makes sense only if we construct robots because we want to understand the behaviour of animals and human beings as an end in itself. But robots with pathologies may also have practical interest. In human beings, pathological behaviours – or the propensity to exhibit pathological behaviours – may be linked to some of the highest manifestations of human creativity and, in the future, creative robots may have interesting applications. And pathological robotics may help us to better understand and, perhaps, to cure behavioural pathologies in human beings. (On this, see Section 7.)

6. Neurological pathologies and psychiatric pathologies

One of the first questions to which a pathological robotics must provide an answer concerns the distinction between neurological pathologies and psychiatric pathologies. Examples of neurological pathologies are the various forms of aphasia, problems with the visual recognition of objects or faces, attentional neglect, and motor disorders. Examples of psychiatric pathologies are pathological anxiety, the bipolar syndrome, obsessive-compulsive disorders, major depression, schizophrenia. What is the basis for this distinction? Neurological and psychiatric pathologies have different causes and they require different competencies and instruments for diagnosis and therapy. Can pathological robotics account for these differences?

As we have said, for robotics the distinction between neurological and psychiatric pathologies cannot be a distinction between pathologies which have a physical basis and pathologies which do not have a physical basis because robots are physical entities and there is nothing inside a robot or in the robot's interactions with the external environment which is not physical. So, how can we explain the distinction between neurological and psychiatric pathologies? One answer is that neurological pathologies are pathologies of the cognitive half of the mind while psychiatric pathologies are pathologies of its motivational/emotional half. Like the human mind, the mind of our robots functions at two levels, the motivational level and the cognitive level. Psychiatric pathologies are pathologies of the motivational level. They change the strength of the robot's motivations in such a way that the robot's behaviour is controlled by the wrong motivations – wrong from the point of view of the robot's survival and reproductive chances or, more subjectively, of the robot's well-being. Neurological pathologies are pathologies of the cognitive level. Whatever the motivation which the robot is trying to satisfy with its behaviour, the robot's capacity to satisfy the motivation is impaired.

If neurological pathologies are pathologies of the cognitive half of the mind and psychiatric pathologies are pathologies of the motivational half of the mind, this may explain why emotions are particularly involved in psychiatric pathologies. As we have

seen in Chapter 2, emotions are part of the motivational half of the mind. When a robot decides which motivation will control its behaviour at some particular time, emotions – the state of the emotional circuit which is part of the robots' neural network – help the robot to make more appropriate decisions. This is why psychiatric pathologies are emotional pathologies. They are linked to a malfunctioning of the emotional circuit of the robot's neural network. Neurological pathologies – like all pathologies – can have emotional consequences but they are not, by themselves, emotional pathologies.

Constructing robots that have neurological pathologies is easier than constructing robots that have psychiatric pathologies. We make a robot blind from birth or the robot becomes blind at a certain age and we see if a blind robot behaves like a blind human being. Or we construct robots which are deaf and see how deaf robots communicate with robots that are also deaf and with hearing robots. Or we lesion a robot's neural network in specific places and we see what are the consequences of these lesions for the robot's behaviour. For example, a robot becomes unable to speak but it understands what the other robots say (Broca's aphasia) or the robot speaks fluently – although with little meaning in what it says – but it does not understand what other robots say (Wernicke's aphasia). Or we progressively cancel the connections of a robot's neural network and we see if the robot exhibits the pathologies associated with an ageing brain. (In Chapter 4 dedicated to robots with language, we have used this method to reproduce the greater difficulty that old people have in producing words compared to understanding words.)

Constructing robots that have psychiatric pathologies is a more challenging task but some of the robots described in this book spontaneously exhibit behaviours which resemble psychiatric pathologies. Psychiatric pathologies can be caused by motivationally conflicting stimuli, and this is what happens to some of the robots that live in the environment with a zone in which the food tokens are more energetic but in that zone lives a predator. (See Section 2.) If the predator is very dangerous because it is very fast in its movements, the robots avoid the predator zone and they eat only the less energetic food tokens which is outside the predator zone. However, if the predator is not so dangerous, the best robots go into the predator zone to eat the more energetic food which is found there and they leave the zone before the predator reaches them. The less able robots are different. They either completely avoid the predator zone or they go into the predator zone but they are not fast enough at fleeing away and the predator kills them.

These are all "normal" behaviours. But there are some robots which, when they enter the predator zone and smell the odour which characterizes the predator zone, stop moving and, therefore, they are easily killed by the predator. Clearly, the odour of the predator zone is a motivationally conflicting stimulus. It is a positive stimulus because it is associated with more energetic food and it is a negative stimulus because it is associated with the arrival of the predator. Some robots react to this conflict with a pathological behaviour which may resemble a strong depressive state. What is malfunctioning is the emotional circuit of these robots. If we eliminate their emotional

circuit – which is like subjecting the robots to a pharmacological treatment – their behaviour becomes less pathological.

The role of conflicting motivations in pathological behaviour can also be demonstrated with the robots which live in the environment with food and water. When we test a robot in an experimental laboratory with both a food and a water token at the same distance from the robot and we vary the robot's level of hunger and thirst, what we find is that, if the robot is more hunger than thirsty, or vice versa, the robot goes to the token – food or water – which in that moment is more needed by its body. But what is interesting is how the robots respond when they have the same level of energy and water in their body. If they are neither hungry nor thirsty and moving has a cost, the robots tend not to move – and this behaviour is clearly adaptive. But what is interesting is what happens if the robots are both very hungry and very thirsty. Most robots go to either the food token or the water token, randomly chosen, but some robots simply do not move. They are paralyzed. Clearly, this behaviour is maladaptive and it resembles pathological behaviours in human beings which are caused by the conflict between two different but both very strong motivations.

Another difference between neurological and psychiatric pathologies has to do with the causes of these two types of pathologies. Both neurological and psychiatric pathologies may have genetic causes but, aside from these genetic bases, during a robot's life various things can happen that cause the robot to have neurological or psychiatric pathologies. However, the causes of these pathologies tend to be different. The neural network of a robot can incur some physical damage due to events in the external environment (for example, physical trauma that damages the brain), events inside the robot's body (for example, lesions in the brain caused by problems in blood circulation), or events inside the brain (for example, deterioration associated with age), and all these events can result in neurological pathologies. Therefore, if we want to construct robots with neurological pathologies, we must "lesion" the robots' neural network by eliminating some of its sensory, motor, or internal neurons or some connections between neurons, or by adding high level of noise to the activation level of the neurons. Psychiatric pathologies tend to have another type of causes. A robot has certain life experiences which change the robot's neural network in ways that cause the robot to have some psychiatric pathology. The robot's neural network cannot be said to be damaged in the sense that some of its neurons or connections are eliminated or noise is added to its functioning but the robot's life experiences cause the robot's neural network to acquire connection weights which produce psychiatric pathologies. Or what is damaged is the functioning of the neural network's emotional circuit, and this damage can be both cause and effect of the robot's life experiences.

Nonhuman animals have both psychiatric and neurological pathologies but in human beings psychiatric pathologies are more complex because human beings have a very rich mental life. Mental life is the self-generation of stimuli by the brain – memories, imaginations – and, like all stimuli, these self-generated stimuli influence the

strength of the different motivations. The process tends to be a circular one. The stimuli self-generated by the brain influence the strength of the individual's motivations which, in turn, evoke other self-generated stimuli. If the influence of the self-generated stimuli on the strength of the individual's motivations is not a healthy or adaptive one, the individual will have negative emotions and pathological behaviours.

A pathological robotics should help us to answer other research questions concerning neurological and psychiatric pathologies and we end this Section by listing some of these questions.

We have suggested two criteria for distinguishing between neurological and psychiatric pathologies. The first criterion is that neurological pathologies are pathologies of the cognitive half of the mind while psychiatric pathologies are pathologies of the motivational/emotional half. The second criterion is that neurological pathologies can be due to events that happen during an individual's life but they are not learned and are not associated with the individual's life experiences, while psychiatric pathologies are learned (aside from genetic predispositions) and they are associated with the individual's life experiences. What is the relation between these two different criteria? Why are cognitive pathologies largely independent from life experiences while motivational/emotional pathologies are linked to life experiences?

Another research question is how the nervous system is involved in the two types of behavioural pathologies. Most neurological pathologies are due to destruction of parts of the nervous system at the cellular level while psychiatric pathologies are not generally associated with lesions to the nervous system at the cellular level but to malfunctioning of the nervous system and of its interactions with the rest of the body at the molecular level. (On this distinction, see Chapter 2 dedicated to motivations and emotions.) Furthermore, while neurological pathologies tend to the associated to lesions to the cortex and to its prevalently cognitive functions, psychiatric pathologies tend to involve sub-cortical structures and their motivational and emotional functions. Can our robots with pathologies help us to answer these interesting but difficult questions?

We have made a distinction between pathologies of the body and pathologies of behaviour but pathologies of the body can cause pathologies of behaviour and pathologies of behaviour can cause pathologies of the body. Our robots should reproduce this two-way interaction between the body and the "mind" and help us to better understand why, when the body is ill, this may result in pathologies of the "mind", and when then the "mind" is ill, this may cause pathologies of the body. And the same applies to the distinction between neurological and psychiatric pathologies. The mind is made of two halves, the cognitive and the motivational/emotional half, but the two halves interact together, and this is also true for the pathologies of the two halves of the mind. When old people have difficulty remembering things, they may respond with a negative emotional reaction to this difficulty. This emotional reaction makes more difficult for them to remember those things because their brain is processing the emotional reaction and cannot remember them.

We have talked of neurological and psychiatric pathologies but there is another form of pathologies of behaviour: psychological disturbances. People with psychiatric pathologies go to a psychiatrist, people with psychological disturbances go to a psychotherapist. The distinction may be a blurred one and psychological disturbances may only be a less serious form of psychiatric pathologies but pathological robots should exhibit both forms of pathologies and, to understand behaviour and inter-individual differences in behaviour, robots with psychological disturbances may be as useful as robots with psychiatric pathologies. For psychoanalysts, psychological disturbances are due to "unconscious" causes. Human beings may be unable to remember past events because these past events would generate in them strongly negative emotional states. This can lead to psychological disturbances because the negative emotional states are not eliminated by not consciously remembering the past events that cause them. Psychoanalysts also think that bringing these past events to consciousness by talking about them eliminates or reduces psychological disturbances. Can robots help us to reproduce these phenomena?

This leads us to a final distinction between neurological and psychiatric or psychological pathologies, a distinction which concerns therapy. Neurological pathologies are resistant to drugs and they can be alleviated, but generally not eliminated, by using rehabilitation techniques. In contrast, psychiatric and psychological pathologies can be alleviated and sometimes eliminated with both pharmacological treatment and psychological – verbal and nonverbal – interactions between patient and therapist. Why? This is another question that our robots should help us to answer. We return to therapy in the next and final section of this Chapter.

7. Prediction, diagnosis, prognosis, and therapy

Nonhuman animals have various strategies with which they try to reduce the dangers associated with illnesses and to recover from them. (In Chapter 2 we have described robots that stop moving when their body is damaged.) Human beings have developed all sorts of methods and technologies – from the more ancient and primitive ones to the extremely sophisticated science-based technologies which exist today – for dealing with both pathologies of the body and pathologies of behaviour. Can robots help us to understand this human behaviour – medical behaviour? The task can be sub-divided into two steps. In the first step we, the researchers, act as doctors and robots with pathologies as patients. In the second step, we construct both robots that are doctors and robots that are patients.

Doctors do four things. First, they predict an individual's illness when the illness has not manifested itself yet (prediction). Second, based on symptoms, they classify the illness into some useful category (diagnosis). Third, doctors predict the course of the illness (prognosis). Fourth, they cure the illness (therapy). (Prevention should also be

added to this list.) One way to measure our success as constructors of human robots is if we, the researchers, are able to do these four things with respect to robots that have behavioural pathologies. We collect all sorts of information concerning an individual robot – information on the robot's genes, its brain, its life experiences, its current behaviour, and its mental life – and, on the basis of this information, we predict that the robot will incur some illness, we classify its illness into some diagnostic category, we predict the course of the illness, and we cure the illness. If we can do this for a variety of robotic behavioural pathologies, we will be more confident that the robots we have constructed are really like human beings.

The second step is to construct robotic doctors. Medical behaviour is one type of human behaviour, and we should be able to reproduce this behaviour with robots like all other human behaviours. The neural network of a robotic doctor receives as input various types of information concerning a robotic patient and, on the basis of this information, the robotic doctor decides to collect other information until it is able to generate a classification of the patient's illness, a prognosis of the future course of the illness, and a specification of the actions to be taken in order to cure the illness. These of course are very complex behaviours and we need to proceed very gradually. Today, there are increasingly sophisticated and often computer-based tools that help human doctors to diagnose illnesses but these tools lack the "clinical eye" of human doctors and tend to reduce personal interactions between doctors and patients. A robotic doctor should also directly see and talk with the robotic patient and a crucial proof that we have constructed human robots is that our robots possess the "clinical eye" of human doctors.

CHAPTER 13

Robots that have art, religion, philosophy, science, and history

When ME looks at human beings, it soon discovers that, beyond their many practical activities and practical interests, they also have art, religion, philosophy, and science. Art, religion, philosophy, and science are not practical activities – although science can also be practically useful – but they are so characteristically human that ME cannot claim to have constructed artificial human beings if these artificial human beings do not have art, religion, philosophy, and science. Human beings do not generally ask scientists to explain these human activities but they ask philosophers, theologians, and historians. ME is a scientist and it wants to understand art, religion, philosophy, and science in the same way as it understands all other human phenomena: by constructing robots that have art, religion, philosophy, and science. It wants to understand why human beings have art, religion, philosophy, and science, how these human activities are related to each other, what are their different forms, how and why they have changed in the history of human societies, and what will be their future. ME is not a philosopher and it does not want to establish what is the "essence" of art, religion, philosophy, or science – because for ME there are no "essences". ME wants to constructs robots that exhibit behaviours that English-speaking human beings would describe by using the words "art", "religion", "philosophy", and "science".

ME also knows that human societies have not always been how they are when it visits the Earth. They have a history, and ME wants to explain this history. ME is well aware that explaining the history of human societies is a difficult enterprise because human societies are very complex entities and each human society is different from all other human societies. But ME thinks that reproducing the past history of human societies with robots is essential for understanding both human beings and their societies.

1. Explaining this Chapter

This Chapter is dedicated to robots which have art, religion, philosophy, and science, and to robotic societies which have a history. Robots that have art, religion, philosophy, and science are as distant as possible from robots which have practical applications and this explains why constructing these robots seems to be nonsense. But a robotic science of human beings cannot ignore these phenomena because they are an important component of the human adaptive pattern. We have not actually constructed robots that have art, religion, philosophy, and science but we have some hypotheses on how to construct these robots and in this chapter we outline these hypotheses.

Another important component of the human adaptive pattern is that human beings live in societies which, unlike animal societies, change and have a history. Since how human beings behave depends on the societies in which they live, we must understand these changes and this history and, to understand them, we must reproduce them in a computer. We have already described some simulations of human historical phenomena in other chapters of this book and we will describe another simulation in this chapter. But we are especially interested in how doing history by simulating history changes the discipline called history.

2. Robots that have art

Art is such an important component of being human that, if our goal is to understand human beings by constructing robots that are like human beings, our robots must necessarily have art. When will we be justified to say that robots have art? We use a very practical and operational definition of art. Robots that have art are robots that spend a significant part of their time exposing themselves to things made by other robots – visual images, sounds, written texts – with no easily identifiable purpose on the part of both the robots that make these things (the artists) and the robots that expose themselves to them (the public). The robots described in this book have very practical motivations but, to be called human, a robot must also have the motivation to create or to expose itself to artefacts which do not have practical implications. Why should the robots have this motivation? What is the adaptive value of this motivation? Why should a robot subtract some of its precious time from its practical activities to create artistic artefacts or to expose itself to artistic artefacts made by others? To answer this question, we must begin with robots which, like nonhuman animals, do only practical things and, in a succession of generations, the robots must evolve – both genetically and culturally – the behaviour of creating artistic artefacts and exposing themselves to artistic artefacts.

Creating artistic artefacts requires manual and cognitive skills but art does not belong to the cognitive half of the mind but it belongs to its motivational/emotional

half. As we have said in Chapter 2, animals and human beings function at two levels: the motivational level and the cognitive level. Functioning well at the motivational level is more important than functioning well at the cognitive level because, if a robot does not choose appropriately the motivation to pursue at any given time with its behaviour, the robot's chances to remain alive and reproduce may be seriously compromised even if the robot is able to do what is necessary to satisfy its motivations. Functioning well at the motivational level is so important that evolution has created a special circuit in the brain that helps animals and human beings to make better motivational choices. The states of this neural circuit are called emotional states or emotions, and the robots described in Chapter 2 demonstrate the adaptive advantages of having an emotional circuit in one's brain. The sight of a potential mate evokes an emotional state which makes it more probable that the robot will pursue the motivation to mate rather than other motivations. The sight of a predator evokes an emotional state which increases the probability that the robot will run away from the predator instead of continuing to look for food. The sight of its offspring evokes an emotional state that causes the robot to feed and remain in proximity to its offspring instead of doing something else.

Emotional states are felt states because the emotional circuit influences the body's internal organs and system and, in turn, these internal organs and systems send sensory inputs to the organism's brain. But the emotional states of an organism also change the external appearance of its body and its behaviour and, in this way, they are not only subjectively felt but they are also made accessible to other organisms. In many animals the external "expression of emotions" is an inevitable by-product of the influence of the emotional circuit on the body, without functional or adaptive value. But for human beings, who are very social animals and live in an environment which is mainly made up of other human beings, the expression of emotions is an important mechanism of social interaction and social life. Expressing one's emotional states lets others know one's emotional states and this increases the probability that one will be helped by others to satisfy one's motivations. Furthermore, if we know the emotional states of another individual, we can predict the behaviour of the other individual and we can adapt our behaviour to the other individual – which are both important prerequisites for living together. Human beings "resonate" emotionally in the sense that, if one perceives the external expression of the emotional state of another individual, one tends to have the same emotional state of the other individual. This sharing of emotional states, that can involve many individuals, can help to coordinate one's behaviour with the behaviour of other individuals to reach some shared goal.

Artistic artefacts express emotions in the sense that they cause emotional states both in the artist who produces the artefacts and in the public. Our hypothesis on the adaptive value of art is that art helps human beings to train their emotions so that their emotions can better help them to make the proper motivational decisions during their life. This is the survival/reproductive value of art, the evolutionary origin of art, the reason why art has such an important and universal place in the life of human

beings. The emotional states of the artist become more articulated and more sophisticated if the artist expresses them in its artefacts. The emotional states of the public become more articulated and more sophisticated if they expose themselves to artistic artefacts. Furthermore, artistic artefacts allow human beings to have the emotional states they want to have and whenever they want to have them, to have emotional states without the dangers associated to some emotional states in everyday life, to make explicit what is implicit in their emotional life, to share their emotional states with others so that they can feel they are not alone. According to the Indian tradition of Tantra, aesthetic experience is "knowing independently of any association with I and My". Artistic artefacts allow human beings to have emotional states without the constant preoccupation for themselves and for what is "their" which accompanies emotional states in everyday life.

Some animals have behaviours which look like art, such as nest decoration in birds, but art is a specifically human phenomenon. Can the hypothesis that art is emotional training explain why only human beings have art? Human beings have a great number of different motivations and a very rich social life compared to nonhuman animals, and this inevitably makes their motivational choices more difficult than the motivational choices of nonhuman animals. But what is even more important is that, in addition to behaviour, human beings have a mental life (see Chapter 5). The stimuli that arrive to their brain do not only arrive from the external environment or from their body but they are self-generated by their brain under the form of images, memories, predictions, and reflections. Since the strength of the different motivations is influenced by the stimuli that arrive to the brain, the stimuli self-generated by the brain make their motivational choices more complex and more difficult. And, in fact, to help them to make their complex motivational choices, they have a very rich repertoire of emotional states. Art makes these emotional states more articulated and more sophisticated. And since mental life allows human beings to live not only in the real world but also in the possible worlds self-generated by their brain, artistic artefacts allow them to live and learn not only in possible *cognitive* worlds but also in possible *emotional* worlds.

Given these characteristics of human beings, the hypothesis that artistic artefacts are made to train their emotions explains why only human beings have art. If we construct robots which, like human beings, have a rich repertoire of motivations and emotions, a rich social life, a rich mental life, and the capacity to construct visual, acoustic, and verbal artefacts, we predict that the robots will autonomously develop a tendency to create artistic artefacts and to expose themselves to these artefacts because art makes their repertoire of emotional states more sophisticated and more articulated and, therefore, more useful.

The hypothesis that art is emotional training is a hypothesis on how to actually construct robots that have art. We start with robots which do not have art and then artistic artefacts gradually emerge among the robots and they evolve in a succession of generations. Artistic artefacts evolve culturally, not biologically, although it is because

of their genes that human beings have art while nonhuman animals do not have art. Some robots, the "artists", choose some of the artistic artefacts produced by the robots of their own generation or of past generations and they copy these artefacts. But they also change these copies so as to create new artefacts that express *their* emotional states. The other robots, the "public", expose themselves to the artistic artefacts of their own generation or of past generations but they also are selective: they choose the artefacts to which they expose themselves. And this is what makes some artistic artefacts "beautiful" and "eternal" and other artistic artefacts "ugly" and easily forgotten.

Art as emotional training must be distinguished from the idea that human beings like artistic artefacts because artistic artefacts make them have *positive* emotions – emotions they like to have. Many artistic artefacts induce negative emotions such as fear, anxiety, and sadness. Artistic beauty is different from non-artistic beauty. The portrait of an ugly woman or of an unhappy man can be beautiful pieces of art. The portrait of a beautiful woman or of a happy man can be bad art.

We do not have actually constructed robots that have art but it is possible to construct such robots and the proof that we have succeeded in constructing robots that have art is that we can answer a number of questions. Here is a list of such questions.

Robots that have art must also have a tendency to expose themselves to objects such as flowers and natural landscapes in much the same way in which they expose themselves to artistic artefacts – without practical implications. But artistic artefacts are different from flowers and natural landscapes as objects of aesthetic contemplation. Artistic artefacts are made by human beings – the artists – and they make possible a sharing of emotional states between the artist and his or her public. Flowers and natural landscapes are not made by anyone. Is art adaptive because it makes human beings emotionally more articulated and sophisticated – and this is something that also flowers and natural landscapes can do – or is it adaptive because it makes it possible to *share* one's emotional states with the artist? When human beings see flowers and natural landscapes as art, do they "think" that someone – God or nature – has made them because He/She wanted to share His/Her emotions with them?

Playing and seeing others play sports, cards, chess and other games are behaviours which resemble creating works of art and exposing themselves to works of art because they also lack practical value. But, again, we should ask: What makes playing different from art? Playing and seeing others play may involve only some specific emotional states – those associated with reaching goals and winning over others – while art makes it possible to practice all sorts of emotional states. Playing may increase one's ability to reach goals and to win over others, while this is not true for art, and the Tantric interpretation of aesthetic experience as "knowing independently of any association with I and My" does not apply to playing and seeing others play. Another difference is that playing may produce improvements in the cognitive half of the mind – better behaviours and better cognitive skills – while artistic artefacts do not produce improvements in the cognitive half of the mind but only in its motivational/emotional half.

Listening to music is not just hearing music but it implies an active role on the part of the listener. Robots that listen to music must react to the music they hear even if music has no practical implications. If we observe human beings when they listen to music, we see that they tend to do three things: (1) they move some part of their body (head, foot, hand) rhythmically, reproducing or anticipating the rhythm of the music they are listening; (2) they reproduce or anticipate the melody of the music by singing to themselves either aloud or, more often, subvocally; (3) they exhibit facial and postural expressions that are normally caused by the emotional states expressed by the music. It may not be so difficult to construct robots that do (1) and (2). Constructing robots that do (3) is a more difficult but not impossible task. Robots that do (1) and (2) may be exclusively "cognitive" robots. Robots that do (3) must be robots that have emotional states and that are able to react to the music they hear by moving the muscles of their face or of their entire body in ways that correspond to the emotional states induced in them by the music.

Artistic artefacts evoke a response which focuses on their "form" and ignores their possible uses and practical implications. Since they evoke a response which focuses on their "form", we may think that good artistic artefacts are artefacts that have a "good form". But what is "good form"? If "form" and "good form" are intended in a mathematical sense, art would belong to the cognitive half of the mind. But is this true? Do human beings have art because artistic artefacts have a "good form" in the sense in which mathematicians can describe a form as "good", or do they have art because artistic artefacts help them to train their emotions? Human beings like the things which have a "good form" because "good form" allows them to predict and control those things – and this may induce positive emotions in them. But many things which have "good form" are not art and many artistic artefacts do not have a "good form" in the mathematical sense of this expression. Art does not belong to the cognitive half of the mind and human beings have art to train their emotional states and not because art is "good form".

If art is exposing oneself to some thing while ignoring the possible uses of that thing and its practical implications, art resembles science. Scientists respond to phenomena by focusing on the phenomena and ignoring their practical implications. In this sense, both art and science are "speculative". But artistic artefacts are prevalently processed by the motivational/emotional half of the mind while facts and phenomena are processed by the cognitive half of the scientist's mind. Art requires emotional participation while science requires emotional detachment. And phenomena are processed by the scientist's mind for at least one practical purpose, the purpose of being able to predict them, while artistic artefacts do not even have this practical implication.

Another difference between art and science is that both artists and scientists, as all human beings, want to be socially recognized but artists do not expect anything from other artists and from their public beyond recognition and participation while

scientists expect from other scientists criticism and objections that allow them to improve their work. Producing art is an individual human activity, even when it is doing music together or dancing together. Science is a social activity.

But the most important difference between art and science is another one and it is a difference which may help us to better understand what reality is for human beings. The manner in which human beings react to things in their everyday life is not only determined by what things are according to science but also by what may be called the "halo" which things have around them. The "halo" of things is not only the emotional reactions that things cause in human beings. The "halo" of things is that things are not "outside" human beings but they are "inside" them. They are "with" them, they "are" them. For human beings things do not possess well-defined boundaries, are not the same across time even if they do not change, are not independent of context, have relations to other things as an intrinsic property. And it is the "halo" that things have around them which explains why for human beings "symbolism" is so important and why their language is full of metaphors and allegories and it may even be basically and originally metaphorical and allegoric.

Art makes the "halo" that things have around time more accessible and explicit while science destroys the "halo" which is around things. One might believe that things have a "halo" around them for the human mind but that *in actual reality* they do not have any "halo" around them. But this mis-assigns the role of "primary" and "derivate". What is primary is that things have a "halo" around them and only a very recent and very special human activity, science, causes them to cease to have a "halo" around them. Much contemporary art – from futurism, cubism, and abstraction to today's techno-art – is an attempt at capturing that "halo" which science and technology themselves have around them.

Art varies as a function of society, culture, epoch, and its relation to other human activities such as religion and political life, and the beginnings of art are interwoven together with the beginnings of all these human activities. Robots must reproduce not only art in general but also the different ways in which art manifests itself in different societies, culture, epochs, and in relation to other human activities. To construct robots that have art, our robots must reproduce not only the phenomena studied by biologists, neuroscientists, and psychologists but also the phenomena studied by pre-historians, historians, and social scientists.

Robots that have art might also tell us something interesting about contemporary art. A hypothesis that could be tested with robots is that to function as a vehicle for sharing emotions and to be appreciated by the public, artistic artefacts must be variations of culturally inherited schemes much in the same way as organisms are variations of inherited genes. Since today's societies tend to reject traditions and inherited schemes, we must ask: Can artistic artefacts be produced in today's societies? Much of today's art is marketing, that is, the production of artistic artefacts to "sell" (or, literally,

sell) them rather than to express something with them. If the artistic artefacts which are produced today are not variations of inherited expressive schemes and they are marketing rather than art, will they be appreciated by future generations?

There are other questions that we should be able to answer by constructing robots that have art, and many of these questions concern inter-individual differences among human beings. (On robots that are inter-individually different, see the preceding chapter.) Can we predict which robot will become an artist? Why some robots expose themselves to artistic artefacts and other robots don't? Why some robots like one form of art – say, music – while other robots like another form – say, visual art? Why some robots tend to read novels and other robots poems? Why poems tend to be read and re-read many times, while this is much less true for novels? Are poems more linked to religious predispositions and attitudes than novels? Does a poet feel inspired by a "voice", while this is less true for novelists? What is the difference between high art and popular art? Why some robots prefer high art and other robots prefer popular art?

Other interesting questions could be answered by constructing male and female robotic artists. (On male and female robots, see Chapter 7.) Why are there more male than female artists? Is this because males vary more than females on a scale of masculinity/femininity and male artists have more feminine characters than ordinary males? And the same applies to the public of art. Are men who like art more feminine than men who don't like art? Do the differences between men and women have a biological basis or are they due to how human societies and human cultures are organized so that, if human societies and human cultures change, these differences will disappear?

A final question is how human artistic artefacts are related to the specific characteristics of the human body, of its sensory and motor organs, and of the specific adaptive pattern of human beings. One advantage of constructing robots to understand the behaviour of organisms is that we can construct robots which are different from any existing organism. Will robots which have bodies and sensory/motor organs different from human bodies and human sensory/motor organs produce artistic artefacts different from human artistic artefacts? What types of artistic artefacts would be produced by robots that do not have human-like hands, robots with less sophisticated visual capacities and more sophisticated smelling capacities than humans, robots that like sounds which we don't like, robots that displace themselves in such media as air and water rather than on surfaces? And the same is true for the adaptive pattern of the robots. Will robots that live alone produce artistic artefacts? What type of artefacts would be produced by robots which reproduce nonsexually, do not need to take care of their offspring, do not remember their past and do not imagine their future? If these robots produce artistic artefacts which are unlike human artistic artefacts, will human beings enjoy the artistic artefacts of these robots? Can the artistic artefacts produced by robotic artists suggest new types of artistic artefacts to human artists?

3. Robots that have religion

Constructing robots that have religion is not easy but, given the importance of religion in the adaptive pattern of human beings, to call our robots human, the robots must have – or must have had – religion. Religion is to believe in the existence of what cannot be known and understood as we know and understand the world of our everyday life. Religion manifests itself in a variety of forms in different societies and at different times and it is strictly linked with magic, witchcraft, myths, social ceremonies, and social power. What we must do is start with robots which, like non-human animals, do not have religion and then, in a succession of both biological and cultural generations, these robots must develop religious beliefs and practices and, when the robotic societies become like "modern" human societies, many robots must cease to have religion. Like robots that have art, robots that have religion are future robots but we can make some hypotheses about religion that we might test by constructing robots.

One hypothesis is based on the fact that, compared to nonhuman animals, human beings incorporate in their brain a greater number of regularities existing in their environment because they do not simply act on the environment but they also predict and control the consequences of their actions. In this way, they become able to explain what is happening now and what has happened in the past and to predict what will happen in the future. But they soon discover that there are many limitations to their knowledge. Many things happen or can happen that they are unable to explain and to predict and they pose to themselves questions that it is impossible for them to answer. To overcome these limitations and feel more comfortable, they imagine that existence of entities which are not of this world and which know what they cannot know and explain what they cannot explain.

This hypothesis links religion to the cognitive half of the mind, to what human beings know and what they don't or cannot know. Another hypothesis links religion with the motivational/emotional half of the mind. Human beings desire many things that it is impossible for them to have – for example, a never-ending life for themselves and for their loved ones – they fear what can happen to them and to their loved ones and, more generally, they would like reality to be different from what it is. To have what they desire and to avoid what they fear, human beings imagine the existence of entities that can realize their desires and calm down their fears, and they ask these entities to make reality what they would like it to be.

Religion is, or has been, linked to art. They were born together and only in modern societies they have become separated. But religion and art have different roles in the life of human beings. Art is emotional training, and is emotional training for all sorts of emotions, both positive and negative. Religion is a set of beliefs and practices which allow human beings to avoid negative emotions. Religion makes people less unhappy.

Other hypotheses about religion are of a more social and political nature. Religion means a shared set of beliefs, values, and practices, and human beings like to live in

communities with a shared set of beliefs, values, and practices because in this way they can understand and predict the behaviour of the other human beings with whom they live and because religion keeps a human community together and this allows the community to better compete with other communities. Religion is also linked to the need of human communities to have "chiefs" that tell the members of the community what they must do, and "chiefs" use, or have used in the past, religion to be obeyed. (For "chief" robots, see Chapter 12 on political robotics.)

These hypotheses about religion can all be true given the variety of different religions, the variety of roles that religion has played and continues to play in human societies, and the historical changes in religions and human societies. But they remain hypotheses. To test them, we must construct robots that incorporate these hypotheses and see if these robots develop religion and what religion they develop given the different societies in which they live. Our robots should also reproduce how religions have changed during the history of human societies and they should help to answer this question: Does the number of religious people tends to decrease in modern societies because these societies have science – which claims to know all that can be known about reality – and technology – that gives human beings a greater control on their life and their future?

Robots should reproduce the variety of religions and they should allow us to understand why different societies have different religions. For example, Western religion is very anthropocentric. The son of God is a human being and God himself is like a human being, not like a cat or a monkey or nature or the whole of reality. God has created the world in the same way as human beings create the environment in which they live. God is omniscient and human beings want to be omniscient. God is omnipotent and human beings want to be omnipotent. God loves human beings as human beings would like to be loved by other human beings. Robots with religion should help us to answer this question: Why is Western religion so anthropocentric?

And if we construct robots that have religion, we can do the same thing that we can do with robots that have art. We can construct robots that are in one aspect or another different from human beings and see what is the religion of these robots.

In any case a robotic science of religion is especially important because religion almost inevitably elicits attitudes which are either for or against religion. Robots can make human scientists look at religion with the necessary detachment which is required by science.

4. Doing metaphysics with robots

Some human beings are philosophers and, therefore, some robots must be philosophers, and these robots should make us better understand how philosophy is different both from religion and science, why philosophy was born where it was born and why

it has been mostly Western philosophy, and how it has changed in the course of time. This is not easy to do. In this section we will not describe robots that do philosophy because we have not constructed these robots but we will try to show how robots can help human beings to do philosophy.

One thing that philosophers do is try to establish what are the basic constituents of reality – objects, properties, actions, events, states, processes, categories, parts, wholes, time, space, qualities, quantities, will, freedom of the will, etc. – and they call this enterprise metaphysics or "first philosophy". But philosophers are human beings and whatever conclusions they may reach concerning the basic constituents of reality, we should ask: Are these conclusions independent of the fact that philosophers are human beings? Imagine that philosophers were worm philosophers or dog philosophers or chimpanzee philosophers. Would the basic constituents of reality be the same for human philosophers and for worm philosophers, dog philosophers or chimpanzee philosophers? And, if not, are we entitled to say that human beings know what the basic constituents of reality are, while worms, dogs, and chimpanzees don't? Only human beings do metaphysics and, more generally, philosophy because metaphysics and philosophy are verbal activities and only human beings have language. But, perhaps, when philosophers do metaphysics what they actually do is not trying to identify the "true" constituents of reality but to describe the basic constituents of reality of one particular species of animals: Homo sapiens. This extends to the differences among human beings. Human beings are of different ages, they belong to different cultures, they speak different languages, and they have different personalities and pathologies. Does their metaphysics depend on, and varies with, these factors? Infants do not have a notion of "object" from birth but they gradually acquire the notion during the first months of their life. The notion of "will" and the idea of "time" vary across cultures. Chinese people may not distinguish between objects and their properties because the Chinese language not does not distinguish between nouns and adjectives. People with autism may not have the same notion of "other human being" as people without autism. The metaphysics of a congenitally blind person may no be the same as the metaphysics of a non-blind person.

This also applies to philosophers. Philosophers are not ordinary human beings. When they try to establish what the basic constituents of reality are, they act as specialists, not as ordinary human beings. And, perhaps, if we want to know what the basic constituents of reality are for human beings we should ask ordinary human beings, not philosophers. The problem is that ordinary human beings may not understand the question and, therefore, they may be unable to answer the question.

But the problem is more radical. Should the answer to the question "What are the basic constituents of reality?" necessarily be given in words and through an analysis of words (philosophers would say: concepts)? Perhaps, to know what the basic constituents of reality are for human beings, we should not use language but we should look at how human beings behave and how their brain functions. Only human beings

have language, and if we could only answer the metaphysical question of what the basic constituent of reality are by using language, it would be impossible to know what the basic constituents of reality are for animals because animals do not have language. If we want to know what the basic constituents of reality are for both animals and human beings, we should ignore language and examine their behaviour and their brain.

Doing philosophy is inherently linked to language, and philosophers would not be interested in metaphysical questions if they were prohibited from using language to answer these questions. Unlike philosophy, science is not so linked to language. To be considered as correct, any scientific theory of what are the basic constituents of reality for this or that animal should be able to generate predictions that must correspond to observable facts about the behaviour and the brain of the animal. An important advantage of this way of proceeding is that we can adopt a comparative approach, which is a very useful tool of science. We can ask the question "What are the basic constituents of reality?" not only for human beings but also for other animals, we can do this for different types of animals, and we can explain why for this or that animal reality has this or that basic constituent.

This comparative approach to metaphysics has two steps. The first step is to find out what are the basic constituents of reality for an animal by observing the behaviour and the brain of the animal – not by asking the animal because in the latter case we would get an answer only from human beings or, more precisely, from a small subset of human beings – philosophers. The second step is the critical step and it is doing metaphysics with robots. We should try to answer the question of what are the basic constituents of reality for one particular animal – for example, a human being – by constructing a robot that behaves like that animal. We can establish what is the metaphysics of the animal not by observing the behaviour and the brain of the animal but by observing the behaviour and the brain of a robot that reproduces the animal.

This has a number of advantages. We can better study the behaviour and the brain of a robot that reproduces one particular animal because the robot has been constructed by us and we can dissect and manipulate the robot in any way we want in order to better understand what are the basic constituents of reality for the robot – and for the animal which the robot reproduces. Another advantage is that, if we work with robots rather than with real animals, we can extend the comparative approach beyond existing animals. We can construct a robot which is not like any existing animal and predict that, given the particular body and brain of the robot, the basic constituents of reality for the robot will be such and such – and then see if our predictions are correct. And a final advantage of doing metaphysics by constructing robots is that, since our robots are not programmed by us to exhibit this or that behaviour but they evolve or learn the behaviour in a specific environment, this allows us to link the basic constituents of reality for each type of robot with the specific environment and the specific evolutionary history of the robot. And this is important because the metaphysics of any animal is the result of the overall adaptive pattern of the animal and of the evolutionary history of its species.

Here are some of the questions that we should be able to answer by doing metaphysics with robots. When something "exists" for a robot? When can a robot be said to live in a world of "objects"? Must "objects" have what psychologist Jean Piaget called "object permanence"? How "object permanence" develops during the life of a robot? Do animal robots have "object permanence"? When do objects have parts and properties for the robot? Does space and time exist for the robot? Does the robot distinguish between actions and events? Does it distinguish between qualities and quantities? Does the robot have will? Does it have freedom of the will? Do "I" and "My" exist for the robot? Is the metaphysics of a robot without language the same as the metaphysics of a robot with language?

Constructing robots as tools for doing metaphysics transforms metaphysics into science. But doing metaphysics as philosophers do it is, by itself, an important human activity and, to call our robots human, some of these robots – robotic philosophers – should themselves do metaphysics and they should do metaphysics like human philosophers do metaphysics – by using words. Furthermore, metaphysics is only one part of philosophy. Philosophy has other parts and other goals. Like religion, philosophers look at a reality as a whole and they try to discover what is the "meaning" of life and of reality as a whole. And another of their goals is to find out – and to tell others – what is the "good life" for human beings. We can construct robotic philosophers to understand why (some) human beings are philosophers but we must remember that no human activity remains the same in the course of time and no human activity is eternal. Philosophy was born from religion, then it became a separate human activity, and today is closer and closer to science. Constructing robotic philosophers may help us to predict the future of philosophy and, perhaps, to predict that philosophy has no future.

5. Robots that do science

To construct robots that do science, we must start from a preliminary and very general question: Why (most) animals have a brain? What are brains for? All animals need to know the environment in which they live in order to survive and reproduce in that environment. But what is to *know* one's environment? To know one's environment is to incorporate in one's brain the co-variations which exist in the environment and in one's interactions with the environment. An animal's environment contains many co-variations: when X is the case, Y also is the case, where X and Y can be all sorts of things. They can be events: when an event happens, another event also happens. They can be parts or properties of an object: when an object has one part or property, the object also has another part or property. They can be an action and the consequences of the action: when one specific action is executed, the action is followed by one specific event. And the co-variations can be more complex. When X varies in one specific way, Y also varies in a specific way. Or the co-variations can involve more than two things. An event can be followed by another event but only in certain conditions. An

action can be followed by an effect but only in certain conditions. And X or Y may not only be parts of the external environment but they can also be events or states in the animal's body. When something happens inside an animal's body, something else also happens. When the animal does something, something else happens not only in the external environment but also inside the animal's body.

The brain is the organ which is specialized for incorporating in its structure and manner of functioning the co-variations which exist in the animal's environment. All organisms, even plants and brainless animals like jellyfish, incorporate in their body the co-variations present in their environment because this is necessary to survive and reproduce in that environment. But some animals have an organ which is specialized for registering and keeping a trace of these co-variations, and this organ is the brain. This is why brains have evolved and why (most) animals have brains.

Let us give some examples. Some of these examples apply to all animals, some apply only to human beings but we will use the word "animal" to also include human beings. By incorporating in its brain the co-variations which exist in the environment, when an event happens and the animal perceives the event, the animal knows that another event will also happen. When an animal does something, it knows that what it has done will be followed by some specific event. If an animal knows that an object has some property, it also knows that the object will also have another specific property. When an animal sees that one part of an object moves in one direction, it knows that the other parts of the object will also move in the same direction. When an animal sees an object which moves in space at a certain speed, its brain knows where the object will be at some later time. When an animal sees a three-dimensional object, it only sees the object from one specific perspective and it only sees some parts of the object, but the animal also knows how the object will look like if it goes to the other side of the object or, in the case of human beings, if they rotate the object with their hands. When an animal sees an object, it also knows that it will feel a tactile sensory input if the distance between its body and the object becomes zero – the animal has touched the object. When an animal touches an object and a tactile sensory input arrives not only from the part of its body which has touched the object but also from the object, the animal knows that it has touched its body – and this may be the beginning of knowing that one has body.

What proof do we have for saying that the animal's brain incorporates the co-variations which exists in the environment and in the interactions of the animal with the environment? The first type of proof is behavioural. If we look at the animal's behaviour, we find that, when X happens, the animal prepares itself to Y. When an animal sees a predator, the animal knows that it may be killed by the predator and it flees away. If lightning occurs, some human beings put their hands over their ears. When X happens and Y does not happen, an animal – or, perhaps, only a human being – looks surprised. If human beings see an object from one perspective and then they move around the object, they look surprised if what they see is different from what they expected to see.

And this is true in other cases. One part of an object moves in one direction and the rest of the object does not move in the same direction. The distance between an object and one's body is zero and there is no tactile sensory input from the object.

The other type of proof is more direct because it is neural. If we examine the animal's brain, we find that, if in the animal's environment X co-varies with Y, what happens in the brain when the animal perceives X is different from what would happen if X did not co-vary with Y. This can be seen by constructing robots which evolve or learn their behaviour. If the robots live in an environment in which X co-varies with Y and the robots' behaviour evolves in a succession of generations, we find that what happens in the brain of the robots of the initial generation when they perceive X is different from what happens in the brain of the robots of the last generation because, during the course of evolution, the robots have incorporated in their brain the co-variation between X and Y. The same is true for robots that learn. If they live in an environment in which X co-varies with Y, what happens in their brain when they perceive X before learning is different from what happens at the end of learning. In both cases the explanation is the same. Only if their brain incorporates in its structure and manner of functioning the co-variation which exists between X and Y, the robots can survive, reproduce, and live well.

All animals are born with a brain which already incorporates some of the co-variations existing in their environment because these co-variations are encoded in their genes. But other co-variations are learned during life and they are due to the individual's specific experiences in the environment. As we have said many times in this book, this is true for all animals but it is especially true for human beings since the behaviour of human beings is mostly learned rather than genetically inherited. Compared to nonhuman animals, human beings incorporate in their brain a much greater number of co-variations existing in their environment and in their interactions with the environment, and this is why the brain of human beings is bigger – relative to the size of the body – than the brain of most animals.

Another difference between human beings and nonhuman animals is that human beings have language, and language allows them to incorporate in their brain the co-variations existing in their environment without experiencing them directly because other people tell them what these co-variations are. But language is important not only because human beings can know from others the co-variations that they know but also because, as we have seen in Chapter 4 dedicated to language, language allows human beings to have a richer and more articulated model of their environment in their brain and, therefore, to incorporate in their brain a richer and more articulated set of co-variations existing in their environment. By expressing the co-variations in words, human beings find that certain co-variations co-vary with other co-variations, they discover new co-variations by using already known co-variations, and they identify more general co-variations that explain more specific co-variations. This can be done by the single individual (reasoning) and it can be done by interacting verbally

with other individuals (discussing). In both cases, language makes the system of co-variations incorporated in the human brain larger and more powerful.

A final difference is that human beings have what psychologists call intrinsic motivation, which is the motivation to do things not to reach some practical result but to know the co-variations which exist in the environment and in one's interactions with the environment. Intrinsic motivation also exists in other animals but it is typically human because human beings have hands and with their hands they can do all sorts of things and see what happens as a consequence of what they do.

These differences between human beings and nonhuman animals explain why, unlike most nonhuman animals, human beings construct artefacts and develop technologies. As we have said, incorporating in one's brain the co-variations that exist in the environment and in one's interactions with the environment is necessary to behave appropriately in the environment – and this is true for all animals. But when the co-variations incorporated in the brain become sufficiently rich and complex – as they are in human beings – human beings can use them not only to directly interact with the environment but also to construct artefacts and technologies which help them to interact in better ways with the environments. And the desire to construct better artefacts and technologies is a powerful pressure for looking for and incorporating in one's brain the co-variations which exist in the environment.

This is the beginning of science. Science is a human activity specifically aimed at capturing the co-variations which exist in the environment. Scientists are not only interested in the co-variations that they happen to observe but they *look for* new co-variations. They ask: What happens if A happens? If A co-varies with B, what are the specific properties of A that make A co-vary with B? If A co-varies with B, what else also co-varies with B? If A generally co-varies with B but this is not true in one particular occasion, why A did not co-vary with B in that particular occasion? If A co-varies with B and B co-varies with C, does A also co-varies with C?

All these operations are facilitated by language, and they are facilitated by language in two ways. Language helps scientists to clearly identify what A, B, and C are and, more generally, it articulates the model of reality that they have in their mind (on this, see Chapter 4). And, with language, scientists are able to deduce more general co-variations from specific co-variations and to construct systems of co-variations. The other way in which science is facilitated by language is that language is a communication tools and by using language scientists can work with other scientists, can communicate the result of their research to other scientists, and can discuss their work with other scientists. Language makes the knowledge that human scientists have of reality a social knowledge.

But the two most important characteristics of science are, first, that science looks at the environment in a quantitative rather than qualitative way and, second, that it does not limit itself to finding the co-variations which exist in the environment but it constructs theories that explain these co-variations.

Numbers and quantities play an important role in the everyday life of all human beings – for example, when they measure temporal and spatial extensions or when they exchange things with other human beings – but numbers and quantities "are" science. Numbers and quantities make it possible to observe the environment in a socially shared – objective – way because everyone must necessarily agree on numbers and quantities. They make it possible to capture not only the co-variations between things but also the co-variations among quantities of things, to quantify the probability of a co-variation, to discover the co-variations that exist in very large numbers of observations, and to find new co-variations on the basis of already known co-variations by using the very powerful tools of mathematics which make it possible to find co-variations which abstract from what co-varies.

But science is not mathematics. Beyond accumulating very large sets of quantitative data and finding the co-variations and the regularities that exist in the data by doing mathematical operations on the data – which today, because of the computer, have become "big data" – science develops theories which *explain* these co-variations. Science is not only finding the co-variations which exist in reality. Science is also finding *why* they exist. Some theories are expressed by using mathematical symbols and other theories by using words. Today, scientific theories can also be expressed by constructing computer-based artefacts which behave like the phenomena that scientists want to explain – and this new way of expressing scientific theories can be the future of all scientific theories.

Another characteristic of science is that science is linked to technology and, more generally, to practical applications – and this link is a two-way link. Science is necessary to develop new technologies but technology is necessary for science because it makes it possible to construct scientific instruments that allow scientists to better capture the co-variations that exist in the environment. And technology and practical applications are also useful to science because one important criterion for judging if science actually explains reality is its capacity to construct useful technologies and to have practical applications. This also applies to a robotic science of human beings, although a robotic science of human beings should be useful to human beings not only because it leads to better practical applications but also – and especially – because it can help them to understand and, perhaps, to solve the difficult problems that afflict them today. (On this, see the last chapter of the book.)

Since (some) human beings do science, (some) human robots also must do science and, if we are able to construct robots that do science, these robotic scientists can become regular members of the community of human scientists and they can collaborate with human scientists to accelerate the progress of science.

But robots that do science will also help us to answer a very important question: Does science actually reveal to us what is reality? Is reality what scientists tells us is reality? Or the picture of reality which emerges from science depends on the fact that scientists are human beings, and human beings have a body, a brain, and sensory and

motor organs which are different from the body, brain, and sensory and motor organs of other animals? As in the case of metaphysics (see the preceding section), we should adopt a comparative approach. We construct different robots that do science, robots with different bodies, different brains, and different sensory and motor organs, and then we see if the picture of reality which these different robotic scientists give us is the same as the picture of reality that human scientists give us. So far, asking if science actually reveals to us what is reality has been a philosophical question. With robots, it becomes a scientific question.

Scientists are human beings and they necessarily look at reality from the perspective of human beings. If we ask "Why is the world mathematical?" – as some physicists and philosophers do – the answer is that human beings make it mathematical because human beings count and measure. Non-human animals do not count and measure and, for them, the world is *not* mathematical. More generally, we should ask: "Does science actually tell us what is reality or it only tell us what is reality for human beings?" Like all human beings, human scientists are anthropocentric. They search for other beings in the Universe which are more or less like them, which have their intelligence or, perhaps, more of their intelligence, which have the same technologies that they have or, perhaps, more developed technologies, and they even ask if these non-terrestrial beings are good or bad as human beings are good or bad. But we don't know how reality has changed in other parts of the Universe from its beginning 14 billion years ago, and we cannot assume that it has changed more or less in the same way everywhere in the Universe. We don't know if biological organisms have emerged everywhere from inanimate matter and if they have had the same evolutionary history culminating in the emergence of Homo sapiens or of some animal resembling Homo sapiens.

6. Robots that have history

Given our principle "If you want to understand X, you must know how X has become what it is", if we want to understand animals and human beings, we must know their past – the biological past of their species and their individual past. This applies to both animals and human beings but human beings live in complex societies which play an important role in determining how they behave and, since their societies change, if we want understand why human beings behave as they behave we must know the past history of their societies.

Societies and their history are very complex phenomena and they are unique phenomena, and this is why most historians believe that that the history of human societies can only be narrated but not explained like all other phenomena of reality. The discipline called history is the description of a temporal succession of events, wars, important people, and economic, political, and religious institutions. Historians "interpret" the past – whatever "interpreting the past" may mean – but the interpretations

given by different historians are different and it is difficult to know which interpretation is the right one. Historians are professionally equipped to collect and analyze the documents of the past and to describe and narrate the past, not to explain the past as science tries to explain reality. The discipline of history is intrinsically linked to language and, more specifically, to written language. Historians read past history books, examine and discuss written documents of the past, and express the products of their work by writing books. Historians identify history with written language to such an extent that for them history is the history of societies which had written language while the history of societies without written language is not history but pre-history or proto-history. Furthermore, history continues to play a role which has characterized the discipline since its beginnings. It is a "memory" of the past aimed at establishing and preserving the identity of a group, the sense of belonging to one particular ethnic, cultural, and political community. History tends to be "our history", distinct from the history of others. And history can easily be involved in ideological struggles because human beings use history as a tool to legitimize different readings of the present.

Studying the history of human societies by constructing robots that reproduce this history changes all of this. Robots must explain human history, not narrate human history. Explaining human history is an intrinsically difficult task because human societies are very complex entities and they are unique entities. What can a robotic science of human beings say with respect to this complexity and this uniqueness? With respect to complexity, our robots are scientific theories and all scientific theories simplify with respect to reality and they let us understand and explain reality only because they simplify with respect to reality. So the complexity of human societies is not a problem although, of course, scientific theories must make the appropriate simplifications, those that help us to understand and explain reality – in this case, human societies and their history. With respect to uniqueness, it is true that human societies and their history are unique entities but all our robots are individually unique and all our robotic societies are unique – and they can help us to understand why human societies are unique and the different ways in which they can be unique.

If we explain the history of human societies by constructing robotic societies, robotics societies will not only reproduce past events, institutions, and important people but they will also identify causes, mechanisms, and processes that explain why past human societies have been what they have been, why they have been different in different places and different times, and why they have changed in the manner in which they have changed across years, centuries, and millennia. A robotic history is no more expressed in (written) words but it is a non-verbal theory of the past of human societies. We literally "see" on the computer screen how robotic societies change, when they change, why they change, in which way they change. And robotic societies become virtual experimental laboratories in which we manipulate the conditions we suppose to have been the causes of historical phenomena and we see if our explanations are correct – like in a real experimental laboratory. Robotic societies can also be used to

do what is called "counterfactual history". We change one factor which we know has played a role in the history of one society, we predict how the counterfactual history of that society would have been, and we verify if our predictions are correct by looking at the computer screen – and taking quantitative measures.

A robotic history of human societies can be done in three different ways. We can construct *types* of robotic societies which do not correspond to any specific human society which has existed in the past. These typological robotic societies reproduce the emergence and the historical changes in human settlements, human technologies, economic exchanges, money, financial institutions, political institutions, art, religion, philosophy, and science. (Some of the robots described in this book try to reproduce, in a very simplified way, these phenomena.) Or we can compare different robotic societies to understand how these phenomena have emerged and how they have changed in different types of human societies. And, finally, we can construct robotic societies that reproduce unique historical phenomena such as the expansion of the Indo-European languages in Europe from the IX millennium to the V millennium BC described in Chapter 8 or the changes in the human settlements in Ancient Southern Etruria in the second millennium BC described in Chapter 10. A third example is the expansion of the Assyrian Empire in the near East beginning in the XIV millennium BC and its final collapse in the middle of the VII century BC. We briefly describe here this last simulation.

In the XIV century BC the Assyrians began to expand from their homeland on the upper Tigris river (present-day northern Irak) and, after seven centuries – at the beginning of the VII century BC – they controlled most of the Ancient Near-East. To simulate this process of expansion and the emergence of the Assyrian Empire we divide the entire geographical area of the Ancient Near East into a grid of cells of the same size and, using actual historical data, we assign to each cell three indices – three numbers that can go from 0 to 1. The first index is the physical penetrability of the cell – cells with high mountains or desert areas are less penetrable than cells with hills and cultivable plains. The second index is the cell's political penetrability – cells which are already occupied by other peoples are more difficult to penetrate. The third index is the quantity of natural resources which exist in the cell – cells with more natural resources are more attractive for the expanding Assyrians.

At the beginning of the simulation the Assyrians are located in a cell near to the upper Tigris river and then, in a succession of cycles, they expand to the adjacent cells which are not already occupied by them. Starting from the original cell, the Assyrians living in a cell ask themselves if there are nonoccupied cells adjacent to their cell and they expand to the nonoccupied cell which is more physically and politically penetrable and has more natural resources. The probability that the Assyrians of a given cell expand to a new cell also depends on a fourth factor: the expansive force of the Assyrians who live in the cell. This expansive force decreases with the physical distance of the cell from the original Assyrian cell on the Tigris river – because of communication

and transportation costs – and it is greater if the already occupied Assyrian cell has many natural resources.

Each cycle of the simulation covers a fixed number of years and in each cycle the computer updates the boundaries of the Assyrian Empire. In this way we obtain a succession of maps of the simulated Assyrian Empire and, when we compare the final simulated map with the final historical maps of the real Assyrian Empire, we find that the two maps are rather similar (see Figure 1). This implies that we have identified the main factors which have determined the expansion of the Assyrian Empire and which explain this historical event.

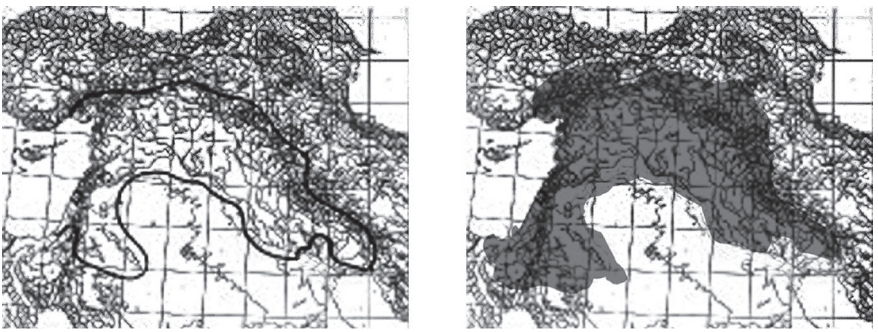


Figure 1. Historical (left) and simulated (right) maps of the Assyrian Empire in the middle VII century BC.

All our simulations – and all our robots – are open to changes and improvements. In the case of the expansion of the Assyrian Empire, we might change the value of the three indices associated with each cell, or add other variables based on historical data, and see if the simulation better approximate the historical events. For example, there is one historical phenomenon that our simulation should be able to reproduce but, in its current version, the simulation does not reproduce. In the second half of the VII century BC, the Assyrian Empire collapses with an extreme and very rapid contraction of its boundaries. Based on actual historical data we can formulate a number of hypotheses that explain this collapse and we can incorporate these hypotheses in the simulation. The collapse of the Assyrian Empire may have been caused by external causes such as the emergence and pressure of other political communities or a sudden change in climate. Or it may have been caused by internal causes such as political struggles, an incapacity to adapt to the very large size of the empire, or an excessive growth of the administrative and military apparatus. And, if we do not succeed in replicating the collapse of the Assyrian Empire by incorporating these hypotheses in the simulation, we can insert in the simulation – again, based on historical data – unpredictable and unique factors such as struggles for power among the Assyrians and the personal weakness of individual emperors.

But the discipline of history is not only the study of the historical changes that have occurred in human societies. Historians do not only study *past* societies but they also study how *past* historians have interpreted *past* societies because these interpretations are themselves interesting cultural phenomena. And then they provide their own interpretations which, of course, are modern interpretations and, therefore, tend to be different from ancient interpretations. Human robots must reproduce all these phenomena – past human societies, past interpretations of these societies, and the interpretations of today's historians – because they are all interesting *empirical* phenomena.

Human societies do not only have a past but they *want* to have a past, and this is why they keep all sorts of traces of their past – monuments, written documents, artefacts, and traditions. This is also true for individual human beings, who want to remember their individual past because their past creates and maintains their individual identity and because, when they become old, they know that their future is limited and what they have is only their past. Robots must reproduce all these phenomena but robotic societies that reproduce the history of human societies must not be memories of *our* societies and tools for preserving *our* cultural and social identity but they must be tools for explaining the history of *all* human societies as science understands all of reality: with complete detachment and absence of values. This is particularly important today because globalization is creating a single human society and a single human culture for all human beings.

A robotic history of human societies is a nondisciplinary history of human societies, and this solves another problem which inevitably confronts historians. Human societies are very different from one another and each society has many components – social, economic, political, religious, and cultural components. If history is done as it is traditionally done, with each historian working alone to collect data, interpret data, and discuss data with colleagues who study the same things which he or she studies, this inevitably leads to specialization, with some historians studying one society and other historians studying other societies, some historians studying one period of one society and other historian studying another period of the same society, some historians studying one component of a society and other historians studying other components – its economy, its political life, its culture. And history is only one of the many scientific disciplines that study the history of human societies. In addition to historians, there are pre-historians, proto-historians, archaeologists, epigraphists, paleontologists, and students of ancient climatic conditions. This is a problem because studying human history since its very beginnings, comparing the history of different societies, studying all types of past "documents", knowing the entire history of one society, and examining the interactions among the different components of a society, are all necessary to understand what human societies are and why they have changed as they have changed. Doing history by constructing robots solves this problem because robotic societies and their historical changes are computer-based artefacts, and computers can contain

enormous quantities of all types of information and they can find all sorts of relations among these different types of information. Doing history ceases to be the work of a single individual and it becomes a two-party work: on one part, many different students of the past working together and, on the other part, the computer.

But explaining the history of human societies by constructing robotic societies is non-disciplinary in a more extended sense. Human beings are the product of their past but their past is not only the past history of their societies but also their biological past. Human history begins at the end of another temporal process – biological evolution – and, if we want to understand human beings by knowing their past, we must recognize that their past does not begin when the history of their societies begins. Human societies are made of human beings, and human beings are not only social and cultural entities but they are also bodies, brains, and physical interactions of their body and brain with the physical environment. Therefore, if we are interested in the history of human societies, we must ignore the distinction between the social sciences and the biological sciences, and robots should help us to put together the biological and the socio-cultural history of human beings.

One phenomenon which is midway between biology and history is the biological and cultural differentiation of human beings. In the computer we can reproduce the biological evolutionary tree of a species by evolving a population of robots in one particular environment and then moving some of the robots of the population to a new environment while the other robots remain in the original environment. Because the two environments are different, the genes of the two sub-populations of become gradually more different until the robots are two different species, which means that a member of one species cannot have offspring by mating with a member of the other species. The process is repeated many times, with further sub-divisions of the original population of robots and, at the end, we can see the entire evolutionary tree of the different species of robots. We can find how two species of robots are genetically related, how much time has past from when they belonged to the same species, and how the evolutionary tree has acquired all its branches.

We can do the same for the cultural evolutionary tree of human societies. We start with a group of robots which have the same culture because they learn their behaviours by imitating the behaviours of other robots. Then some members of the group move to a new environment and become separated from the other members of the group and, since cultures and institutions change much in the same way as genes changes – on this, see Chapter 8 dedicated to cultural robotics and, in particular, the simulation of the historical differentiation of Indo-European languages – the culture and the institutions of this group of robots become different from the culture and institutions of the original group of robots.

Human beings belong to the same biological species and it is only their cultures which make them different – although some of the characteristics of their body are different because of biological adaptation to different environments in the last 50–100

thousands years. But studying both their biological past and their historical past can shed some light on what has made human beings human – having hands, leaving the forest, language, using language to talk to themselves, learning by imitating others and developing cultures, constructing artefacts, living in houses, storing goods, exchanging goods. Robots and robotics societies can help us to answer these questions and to identify both the biological and the social and historical factors that have made human beings human.

CHAPTER 14

Human robots are future robots

Human beings are very complex animals and, whatever way one studies them, they are difficult to know and to understand. An ancient Greek philosopher named Heraclitus said that nature "loves to hide itself" but, after five centuries of science, ME thinks that not nature but human beings "love to hide themselves". ME is convinced that only if science constructs robots which are like human beings, science can understand human beings but it also knows that human robots are future robots and that what it has already done are only the first steps towards a robotic science of human beings. Human robots should *really* be like human beings. Whatever human beings do, think, and feel, human robots should also do, think, and feel. They must be a pair of glasses through which human beings can see what they are in their entirety. To many people, both inside and outside science, human robots will seem to be science fiction but human robots are science fiction made science.

1. A new science of human beings and its problems

Human robots are a *new*, *unified*, and *truly scientific* science of human beings and human societies. New, because it is a science that expresses its theories not by using words or mathematical symbols but by constructing artefacts that reproduce human beings and human societies. Unified, because it is a nondisciplinary science of human beings and its artefacts must reproduce all the different human phenomena which, traditionally, are studied by separate scientific disciplines. Truly scientific, because the meaning of the words it uses is exhaustively defined by how its artefacts are made and how they function – and this makes the science of human beings similar to the science of nature which uses words whose meaning is defined by what scientists observe with their senses and by what they count and measure. And a robotic science of human beings is truly scientific in another sense. Science must look at what it studies from outside, and only a robotic science of human beings allows human beings to look at themselves from outside.

But a robotic science of human beings will need a considerable time to develop and to affirm itself. Today, almost all robotic research is dedicated to robots as practical applications and there is very little money for robots as science. This is a problem because, while the traditional disciplines that study human beings and human societies need

relatively little money for doing laboratory experiments and collecting empirical data, a robotic science of human beings needs a considerable amount of money to have access to the necessary computational expertise and the necessary physical tools. Another problem is that a robotic science of human beings will replace established scientific disciplines, cancel ancient traditions, and pose organizational problems such as how to train researchers who are able to construct human robots and how to make psychologists, anthropologists, sociologists, economists, political scientists, and historians work together – and this explains why both roboticists and traditional students of human beings and human societies may find this book a "strange" book.

Other reasons that will retard the development of a robotic science of human beings have a more psychological and social nature. Human robots will go where they are not expected to go and they are a direct proof that human beings are just complex physical entities – and this goes against a deeply felt human desire to be something more than physical entities. And human beings may not want that one constructs human robots.

These are all external reasons that explain why a robotic science of human beings will take time to develop. But there are also internal reasons. Human beings and human societies are very complex phenomena for any type of science and studying human beings and human societies by reproducing them with robots and robotic societies is still at its very beginning. In this book we have described robots that are part of this program of research but our robots reproduce only an extremely limited number of human phenomena and they make all sorts of simplifications in reproducing these phenomena.

For all these reasons, human – not humanoid – robots cannot but be future robots. But the program of research can and should be pursued and we conclude the book by listing some of the things that will need to be done.

2. What has still to be done

In constructing our robots we have tried to follow the principle "one robot/many phenomena" that we interpret as a special case of the general principle "one theory/many phenomena" which is at the basis of science. To make us understand reality, one and the same theory should account for as many different phenomena as possible. To make us understand human beings and human societies, one and the same robot or collections of robots should reproduce as many different human behaviours and aspects of human societies as possible. But most of the work is still to be done. In the course of the book we have listed some of the many phenomena that still remain to be reproduced and in this chapter we discuss the more general limitations of our robots.

Our robots reproduce a certain number of human phenomena but they reproduce these phenomena *separately*. What should be done is construct robots that reproduce

all these different phenomena *together*. The *same* robots should have many different motivations and emotions, they should evolve, develop, and learn, they should have language and a mental life, they should live together and interact socially, they should be male and female and live in families, they should learn from other robots and develop cultures, they should store their goods, they should specialize in producing one good and obtain the other goods from other robots, they should develop money and have economies, they should have a central store which keeps their community together and a "chief" which makes the central store function properly, they should have art and religion, they should do philosophy and science, and their societies should have a history. Human beings are or do all these things and they are what they are because they are or do all these things. Constructing robots and collection of robots that are or do all these things together is a formidable task but computers have sufficient capacities to let us accomplish the task.

All our "robots" should actually be robots. They should all have a body and a brain and they should live in a physical environment. In many cases our robots are constructed to reproduce the behaviour of single individuals while to reproduce social phenomena we use more abstract "agents" which do not have a body and a brain and do not live in a physical environment. Agent-based social simulations can be useful even if they ignore the body and the brain of the "agents" and, more generally, the biological bases of their behaviour. But human beings have a body and a brain and they live in a physical environment, and having a body and a brain and living in a physical environment do not cease to be important if what one is interested in are human social phenomena. Therefore, "agents" should progressively become robots. Constructing artificial societies which reproduce human societies – with their very complex social, economic, and political institutions – by using robots with a body and a brain will certainly not be an easy task. But, as we have already said, the computer has great memory and processing capacities and one day we will construct really robotic societies because they are necessary to understand human societies.

Comparing "agents" with "robots" raises an important question that needs to be discussed explicitly. A robotic science of human beings clearly is a materialistic science of human beings. Human beings are bodies and their societies are interactions among bodies. And our principle "if you want to understand X, you must reproduce how X has become what it is" requires that, to understand human beings, we reproduce how human beings have become what they are starting from nature and simple physical matter. But a robotic science of human beings is not a reductionist science of human beings. Reality is the progressive appearance of more complex entities – from atoms to molecules, from molecules to cells, from cells to multicellular organisms, from individual organisms to societies of organisms – and to understand one level of entities we must understand the preceding levels. But the entities of one level cannot be "reduced" to the entities of the preceding level because reality is the emergence of *new* entities, where "emergence" means that an entity is more than the sum of its parts

and that its properties cannot be inferred from the properties of its parts. And with the passage of time reality has become progressively more complex, which means that more recent phenomena can less and less be reduced to more ancient phenomena. As we have said, a robotic science of human beings is a materialistic science of human beings but "materialism" can mean two different things. One is that what physicists study is the *only* thing which exists. In this sense of "materialism", a robotics science of human beings *is not* materialistic because it recognizes that many other things exist beyond those that are studied by physicists and these other things cannot be understood by physicists. But a robotic science of human beings *is* materialistic if materialism is adopting the principle "if you want to understand X, you must understand how X has become what it is". Everything which exists has emerged from what, today, is studied by physicists and, if we want to understand anything which exists, we must go back to physical matter.

Another thing that will have to be done concerns the robots themselves. Our robots are not physically realized but they are only simulated in a computer. This may not be important because computers can simulate physical – and, in the case of organisms, also biological and chemical – reality in as much detail as we desire. But in many cases working with physically realized robots rather than with simulated robots can tell us something new and suggest new questions.

More serious is the objection that our robots – which, in most cases, are simulated e-puck robots – have a body which is very different from the body of human beings. A robotic approach to the study of human beings is based on the assumption that it is the possession of a human body which determines how human beings behave and how their "mind" functions. If we ask "Can a robot with a body which is not a human body behave like a human being and have a "mind" like the human "mind"?", the answer is No. Therefore, future human robots will have to have a body which is more similar to the body of human beings, with legs, arms and hands, and with human-like sensory organs and possibilities of movements of the body's different parts.

Our principle "if you want to understand X, you must not only reproduce X but also how X has become what it is" also applies to the body. We must start with a population of robots with a body which resembles the body of the last common ancestor of human beings and chimpanzees and, then, in a succession of generations the body of the robots becomes like the body of today's human beings. By varying the environment in which the robots live – because the environment changes for its own reasons, because some robots move to new environments, because the robots modify their physical environment and develop new technologies, because the robots live in a social environment which is always changing – we should be able to reproduce the different bodies of Australopithecines, Hominids, and modern humans and, in parallel, the changes in their behaviour and in their adaptive pattern.

But the body should not only be the external morphology of the body and its sensory and motor organs. The body should also be its internal organs and systems

because behaviour is not only the result of the interactions of the brain with the external environment but it is also the result of the interactions of the brain with the body's internal organs and system. An "internal robotics" is as necessary as the current "external robotics" (see Chapter 2) but the robots described in this book mostly ignore what is inside the body beyond the brain – although what is inside the body is especially important if the robots must have motivations and emotions, psychiatric or psychological pathologies, art and religion.

The brain is the only internal organ which our robots do have. But the brain of our robots is extremely simplified compared to the brain of human beings – and even of simpler animals – and the fact that the brain of our robots is so unlike the human brain goes against the principle "one robot/many phenomena" because human robots should reproduce not only the behaviour but also the brain of human beings.

Much work is dedicated today to studying the brain by simulating it in a computer, and this enterprise is called computational neuroscience. But there are two different approaches to constructing robots with an artificial brain which is like the real brain and what distinguishes the two different approaches is where one begins and where one arrives. Typically, one starts from the real brain, constructs an artificial neural network whose structure reproduces some of what neuroscientists know about the brain – its different parts, how these parts are connected together, what are the properties of the different parts and different neurons – and then one looks at the behavioural phenomena that can be reproduced by using robots endowed with this more realistic neural network (see Figure 1).

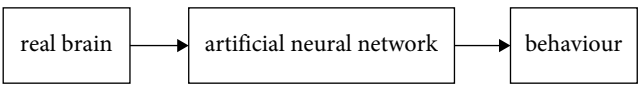


Figure 1. One starts from the real brain, constructs an artificial neural network that reproduces what neuroscientists know about the brain, and then tries to find what behaviours can be reproduced by this more realistic artificial neural network.

The other approach begins with robots which have a very simple neural network that reflects only the most basic and general facts that we know about the real brain – that the brain is made up of neurons linked by synapses. Then one gradually modifies the neural network by adding modules, links between modules, and properties of the neural network's different modules and different neurons that make it possible to enlarge the range of behavioural that the robots are able to reproduce. Once this is done, one looks at the real brain and sees if the artificial neural network's modules correspond to distinct parts of the real brain, if the modules are linked together in the same way as these different parts of the brain, and if the artificial neurons of the neural network have properties which are like the neurons of the real brain (see Figure 2).

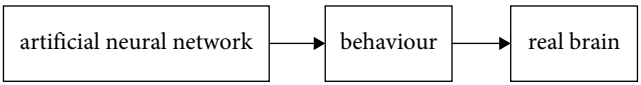


Figure 2. One starts from a very simple artificial neural network, progressively makes the artificial neural network more complex so that it can generate an increasing number of different behaviours, and then sees if the new artificial network resembles the real brain.

This second approach better *explains* why the brain has the structure it has, without simply taking its structure as given, and it is more consistent with the view of robots as theories. Neuro-robots are theories of both behaviour and brain and they must generate predictions on both behaviour and brain. Once we have constructed a robot which reproduces the behaviour of a given animal, we ask: Does the robot's neural network reflect the known empirical facts about the brain of the animal which exhibits this behaviour? How should we change the neural network of the robot so that it both reproduces the behaviour of the animal and matches its brain? Can the robot's neural network suggest new hypotheses about the real brain? Can it tell us what to look for in the real brain?

Today, with the availability of always more powerful computers, computational neuroscience is also done in another way. Computers can collect and process very large quantities of data on the brain and many research projects try to reproduce the entire human brain in a computer. But being able to reproduce the entire brain in a computer does not imply that we have understood the brain. To understand the brain, it is not sufficient to study the brain. Brains are biological entities and, if we want to understand biological entities, we must understand what is their function. The function of the brain is to produce behaviours, and brains evolve and become what they are because of the behaviours that they produce. Therefore, to understand the brain it is not sufficient to know how many neurons make up the brain, what different types of neurons exist, how they are connected together, what chemical and electrical processes take place in the brain, but it is necessary to study how this very complex object produces behaviour. To understand and explain reality, science needs both data and theories In our case a theory must generate predictions on how the brain's structure and manner of functioning produce behaviour. Neuro-robots are important because they are theories of *both* behaviour and brain.

Furthermore, future robots will have to have neural networks that reproduce not only the brain but the entire nervous systems. Current neural networks are models of the brain but behaviour is a product not only of the brain but of the entire nervous system which is made of a central part – the brain and the spinal cord – and a peripheral part – the autonomous nervous system which includes the sympathetic, parasympathetic, and enteric sub-systems. Future robots will need to have neural networks which reproduce the entire nervous system and not only the brain so that we will be able to

see what robots with these more realistic neural networks can do which robots with "classical" neural networks cannot do.

This greater realism should not only concern the different parts of the brain – especially the two asymmetrical hemispheres – and of the entire nervous system but also the fact that the brain and the nervous system are physical entities. Neural networks reproduce the brain but they mostly ignore that the brain is a physical entity existing in space and time. Neural networks – including the neural networks of our robots – tend to be purely *topological* structures. They are made of a certain number of neurons connected together in certain ways but where in physical space is a neuron, how distant is the neuron from another neuron, and other spatial properties of real brains as physical objects, tend to be ignored, while they may be important to understand how the brain functions and how it controls behaviour. The same applies to time. The activity of most neural networks – including the neural networks of our robots – is a succession of abstract input/output cycles but the time taken by a neural network to generate an output in response to an input – what psychologists call response time – is generally not taken into consideration and with most current robots it is impossible to distinguish between reflexes, behaviours, and actions which are thought of before being executed. Another crucial aspect of time in the brain is memory. Some neural networks have some form of memory because they learn and because what happens in one input/output cycle leaves a trace which influences how the network responds to the input of the next cycle. But memory has many different forms and most robots – including our robots – do not reproduce these different forms.

The human brain has not only an individual past but also an evolutionary past because it has evolved from the brain of the common ancestors of human beings and chimpanzees that lived around 7 millions years ago. Future human robots must reproduce the evolution of the human brain and, in particular, the great increase in its size. The brain is the body's organ specialized for incorporating in its structure and manner of functioning the co-variations which exist in an animal's environment so that the animal can behave in the environment in ways that guarantee its survival and reproductive chances. In comparison with other animals – even those which are very similar to them – human beings incorporate in their brain a much larger number of co-variations and much of their behaviour has the only purpose to discover these co-variations. What has been the cause of the increase in the size of the human brain: making tools, language, a more complex social life? Can we construct robots that reproduce the co-evolution of the human body, brain, and adaptive pattern? There is some work on evolving the architecture of neural networks (in Chapter 3 we have seen an example of this work) rather than arbitrarily deciding this architecture but most of the work remains to be done.

Another limitation of our robots has to do with genetics. The genes of our robots encode the connection weights and other parameters of their neural network and, in some cases, they directly encode behaviour as a number that expresses the probability that the robot will exhibit one particular behaviour. In real animals, the genotype is a very

complex chemical entity and the mapping from genotype to phenotype a very complex process. Furthermore, most animals – including human beings – reproduce sexually and the genes of one individual are a recombination of different parts of the genotype of its two parents. Almost all our robots are unisexual and even the offspring of the male and female robots described in Chapter 7 do not have a genotype which is recombination of the genotypes of their parents. Is genetic recombination only a stronger form of genetic mutations as a mechanism for introducing novelty in evolution? Or genetic recombination has a different role in evolution compared to genetic mutations? Future robots should have both genetic recombination and genetic mutations.

That our robots only partially obey the principle that to understand X we must understand how X has become what it is, is also true in other cases. Most of our robots evolve but they do not change during their life, and this does not make much sense if we want to construct human robots because almost all human behaviour is learned and not genetically inherited. (But, although most of human behaviour is learned and not genetically inherited, the biological evolution of Homo sapiens plays a decisive role in what human beings do and feel.) Some of our robots both evolve and learn and what is encoded in their genes influences what and how they learn but, in most cases, our robots *either* evolve *or* learn, or some connection weights of their neural network are evolved and other connection weights are learned. And in all cases our robots start their evolution from zero – with a random assignment of connection weights to the neural networks of the robots of the initial generation – while all evolutionary processes start from the results of past evolutionary processes. Only for some of our robots after a certain number of generations the environment changes and they must adapt to a new environment starting from the connection weights they had evolved in the old environment. In this way it is possible to reproduce important evolutionary phenomena such as behaviours which did not have adaptive value in the old environment but acquire adaptive value in the new environment or had one adaptive function in the old environment and acquire a new adaptive function in the new environment.

Another very important phenomenon we have almost entirely ignored is development. What is encoded in the genes of an animal is not only the initial state of the animal but it is also a program of changes which take place during the life of the animal, and the animal's behaviour changes during life not only because of learning but also because of this process of development. Development and its interaction with learning are very important phenomena and reproducing with robots how human beings change in the first months or years of their life is especially important if we want to understand human behaviour.

A comparative approach is important to understand all animals, including human beings, and therefore, if we want to understand human beings, we must construct not only human robots but also chimpanzee robots, dog robots, fish robots, and bird robots. And living organisms are not only animals but also plants. Plants are different from animals because they do not move from one place to another place but, like

animals, they are born, grow, and die, and they interact with the environment. So, to understand human beings and other animals, we should also construct plant robots.

If we construct robots that reproduce different types of animals – including non-existing animals – we can try to find an answer to many interesting questions. One of these questions concerns time. Every animal exists in time but only some animals have a "sense of time". What is to have a "sense of time"? Do only human beings have a "sense of time"? Do only they have a "mental past" and a "mental future"? These are difficult questions and they are also unclear questions. Robots should help us to make them clearer by translating them into behaviour and brain activity.

Evolution, learning, and development exist in all animals but human beings are also the result of other temporal processes – the history of their societies, the history of their cultures and technologies, the history of their physical environment which changes because human beings change it – and future human robots should reproduce all these histories. We have reproduced with our robots – in an extremely simplified form – some historical changes that have taken place in human societies in the last 10,000 years: food production, food storage, the creation of tools, the inheritance of goods, the exchange of goods, the emergence of "chiefs". But, as we said in Chapter 13, the history of human societies poses special problems because of the uniqueness of human societies and the variety of factors that determine why and how they change. We have simulated some of the historical changes which have taken place in some real human societies of the past (the changes in human settlements in ancient Southern Etruria and the expansion of the Assyrian empire) and in some real human languages (the Indo-European languages) but these simulations are extremely simplified with respect to what actually happened to these societies and to these languages and they reproduce human societies and human languages but not the individuals who made up those societies and who spoke those languages.

Today's human societies are made of many different parts and they have many different institutions: religion, morality, art, the political system, the judiciary system, the economic system. But in "primitive" human societies all these parts and systems were mixed together and, if it is true that to understand X, we must understand how X has become what it is, we must construct robotic societies that are like "primitive" human societies – like the Eskimo societies which were studied at the beginning of last century – and then reproduce how they have become made of separate parts. (And, perhaps, students at school should study these "primitive" societies and not only modern societies and their relatively recent history.)

Another limitation of our robots is that not only they are very simple but they live in very simple environments. Behaviour consists in the interactions between an animal and its environment and, therefore, we have been able to study only very simple behaviours. Future human robots will have to live in more complex environments which contain a variety of different things and in environments that they themselves have created such as buildings, cities, and cars. Furthermore, many of our robots live

alone and, even when they live with other robots, their interactions with the other robots are extremely limited. Human beings are extremely social animals and they live in very complex social environments. We have discussed what makes the social environment of human beings so different from their non-social environment at the end of Chapter 6 dedicated to social robots but human robots that actually live in a social environment are future robots. And what is important is that the environment of human robots must be an environment not only of individual robots but of complex organizations of robots – robotic institutions.

But the most serious limitation of the robots described in this book is that we have rarely compared the behaviour of our robots with the enormous quantity of empirical data which the traditional behavioural and social sciences have accumulated in the last two centuries on human behaviour and human societies. As we have said at the beginning of this book, the behaviour of a robot and the functioning of a robotic society are the empirical predictions which derive from the theory incorporated in the robot or in the robotic society. Therefore, a robotic science of human behaviour and human societies requires that we compare a robot's behaviour and a robotic society with the empirical data that exist on human behaviour and human societies. Only if we do so, we can claim that our robots and our robotic societies explain human beings and human societies and we can hope to establish a dialogue with the traditional disciplines that study human beings.

So, the robots described in this book are more an illustration of a program of research than the actual results of this program of research. Some – or many – of our results may be wrong in one sense or another but, if they are wrong, what we must do is not to cease to construct robots but to learn from our errors and construct better robots. Notwithstanding all the limitations of our robots, we hope we have demonstrated that a robotic science of human beings is possible and that it will provide us with a better understanding of why human beings behave as they behave and why their societies function as they function. A robotic science of human beings cannot ignore the traditional disciplines that study human beings and their societies. These disciplines are not only a rich source of empirical data but they also are a rich source of theories, and a robotic science of human beings must provide a clear empirical test for the theories proposed by the most significant living or dead psychologists, anthropologists, sociologists, economists, and political scientists. But we do believe that a robotic science of human beings is a "scientific revolution", a radical change in how science studies human beings, and it is a scientific revolution because it expresses its theories not by using words or mathematical symbols but by constructing artefacts which behave like human beings and function like human societies. The German poet Johann Wolfgang Goethe wrote that "All theory …. is grey, but green is life's glad golden tree". Theories are "grey" because they are abstract and "cold", while life is "golden" and "glad" because it is like a tree with its own autonomous life. The theories of a robotic science of human beings are green as trees, although they are artificial trees.

CHAPTER 15

How human robots can be useful to human beings

ME knows that human robots will look like monsters to the eyes of human beings because they inexorably "demonstrate" that human beings are only bodies. So ME is not sure that human beings will like what it is doing. But ME is a host of human beings and, although it is interested in constructing robots with the purely scientific goal to understand human beings, ME also wants to be useful to human beings. ME is not so much interested in new practical applications that its robots can inspire but in how its robots can help human beings to better understand the many difficult problems they face today and will face in the future – and, perhaps, to find solutions to these problems. And ME is convinced that, if its robots can help human beings to understand and to find solutions to their problems, this will be an indirect proof that its robots explain human beings.

1. Human robots should be practically useful to human beings

This might be the end of the book but we want to conclude the book by returning to our initial question: Why construct robots? In the first Chapter of the book we have distinguished between robots as science and robots as practical applications. Robots have – and will increasingly have – useful practical applications but in this book we have not been interested in robots as practical applications but in robots as scientific theories that should help us to understand human beings and human societies as science understands all other phenomena of reality. Robots as science and robots as practical applications should talk with each other because robots as science can suggest new practical applications and robots as practical applications can pose new problems to robots as science and can be a test of the goodness of robots as science. But robots as science must be kept distinct from robots as practical applications because they have different goals and their different goals dictate different research agendas and require different criteria of success.

Robots as practical applications are useful to human beings and this explains why we construct them. But can robots as science also be useful to human beings beyond the purely scientific goal of providing them with a better understanding of themselves?

Should researchers who construct robots as science live in an "ivory tower" and ignore what is outside their "ivory tower" and what are the problems of the society? We don't think so and we conclude the book by listing five ways in which robots as science can be useful to human beings beyond letting them better know and understand themselves. If robots can be useful to human beings in one or more of these ways, this will be one important proof that robots do actually make us understand human beings and their societies.

2. Robots make it possible for human beings to know themselves without being influenced by their desires and fears

As we have said in the first Chapter of the book, when human beings study themselves, they tend to be influenced by their desires and by their fears and to see themselves as they would like to be rather than as what they actually are – and this is an obstacle to their knowledge and understanding of themselves. This was also true for their knowledge of nature before the birth of a science of nature, when Galileo's *sensate esperienze* (sensory experiences) and *certe dimostrazioni* (mathematical predictions and proofs) replaced all sorts of religious and philosophical beliefs. A robotic science of human beings can do the same for the science of human beings. Robots as theories make it possible to study human beings without being influenced by religious and philosophical beliefs and by desires, fears, and values. If we construct robots that are like human beings, human beings are like these robots – and we can examine the robots, their brain, their behaviour, and their societies in a completely objective way, with the equivalents of Galileo's *sensate esperienze* and *certe dimostrazioni*. Here is a list of classical questions that human robots can help us to answer without being influenced by our pre-conceptions and values.

- Are human beings a product of an evolutionary process which has taken place in nonhuman primates?
- Are human beings what is encoded in their genes or are they the result of the particular social and cultural environment in which they live?
- Can human beings which have been genetically separated for a long period of time have not only different bodies but also different minds?
- Are men and women different because their genes are different or because they live in societies and cultures that make them different?
- What are the causes of the differences among individuals?
- Are human beings selfish or altruistic? Is selfishness encoded in the genes and altruism learned during life or they are both genetically inherited? Is altruism restricted to one's kin?
- In what conditions human beings do not compete but cooperate?

By constructing human robots which are the result of an evolutionary process, are born with inherited genes but learn most of their behaviours during their life, are males and females, are inter-individually different, live in specific societies, cultures, and economic systems, live in groups which can compete or collaborate with each other, we can find clear and detailed answers to these questions. In most cases the answer will not be a "yes or no" answer but we will know exactly from which point of view the answer is "yes" and from which point of view it is "no". In all cases the answer will ignore our desires, our fears, and our values because robots can really be like human beings only if they are constructed without any pre-conceptions about human beings.

Answering these questions without being influenced by our pre-conceptions and our values would be useful not only for scientists but for human beings in general. Knowing reality in the manner in which science knows reality does not lead necessarily to happiness but it may help human beings to find solutions to many of their problems. This is an example. If you ask science if human beings are all equal, science will answer "No". But if you *want* human beings to be all equal, science may suggest how to make them all equal.

3. Robots can help human beings to identify possible conflicts between biology and culture and the consequences of these conflicts for their happiness

Human beings are the product of two different historical processes: biological evolution and cultural evolution. What they do is the result of the genes which they inherit at birth and of what they learn during their life from other human beings. Both genes and culture change and they both change as a result of selective reproduction and the addition of genetic mutations, in the cases of genes, and selective transmission of cultural traits and the invention of new cultural traits, in the case of culture. From this point of view, the two historical processes of change are similar. But in other respects they are different and one of the most important differences is the speed with which genes and cultures change. Biological evolution is slow. For human beings, changes in their genes may take many thousands of years to become a permanent feature of their genotype. Cultural changes happen at the scale of years and, today, months or even days. Two factors explain this difference. One factor is that, while the genes of an individual are the result of an innovative recombination of the genes of only two other individuals – the individual's parents – the cultural traits of an individual are the result of the innovative recombination of the cultural traits of many individuals – today, all the human beings who live on the Earth. The other factor that explains the different rhythm of change of biological and cultural evolution is that one important class of cultural traits are the artefacts created by human beings. These artefacts increasingly constitute the environment in which human beings live and, today, they change with

great speed under the pressure of the economic system which is based on the creation of always new goods. Behaviour is an adaptation to the environment in which organisms live. But while the environment of nonhuman animals remains more or less the same from one generation to the next, the social, cultural, and technological environment of human beings changes from one generation to the next and even in the course of the life of the individual. This explains why, while most behaviours of nonhuman animals are genetically inherited, most behaviours of human beings are learned during life. The genes of nonhuman animals can contain all the information which is needed to live in a static environment. Human beings need to learn almost all their behaviours because they live in an environment which is always changing.

If the behaviour of human beings is a result of both their genes and their culture and if these two components of the human adaptive pattern change at different speeds, can the two components enter into conflict? Asking this question is important especially today because, as we have said, today the rhythm of cultural change has become very accelerated and the culture which is emerging – or which may emerge in the future – may have components that are in contrast with what is encoded in the human genes. (We are talking of the Western culture but we assume that the Western culture will become, sooner or later, the culture of all human beings who live on the Earth. On this, see Chapter 8.) Human beings who are born today have more or less the same genes of their ancestors of tens of thousands years ago but they live in an environment which is very different from the environment in which their ancestors lived. Can human genes enter into conflict with the manner in which human beings live today, with the behaviours, beliefs, feelings, and artefacts that their culture and the economic system impose on them? And, if such conflict exists, can it make human beings unhappy?

Here are some hypotheses on possible contrasts between what is encoded in human genes and the culture in which human beings live today.

- Most human beings may be born with a genetically inherited need to have offspring and to care for their offspring. Today, the number of human beings which have offspring is constantly decreasing.
- Most human beings may be born with a need to live permanently with an individual of the other sex, to mate together, to have offspring together, and to care for their shared offspring. Today, the number of human beings who live as singles is increasing and the length of time during which two individuals live together is decreasing.
- Human beings may be born with a need to be taken care of by others when they are very young and when they are old or ill without giving anything in exchange. Today, in many cases they can obtain this care only in exchange for money.
- Women can be genetically different from men but today's society pushes women to become men because today's society has been made by men and for men.

- Human beings may be born with a need to live in contact with nature, which is something which has a life like them and has not been made by them. Today, more than half human beings live in cities and this number is constantly increasing.
- Human beings may be born with a need not to decide, until they reach a certain age, what they should do but to be told by adults what they should do. Today, the authority of parents, teachers, and, more generally, adults tends to be rejected by an increasing number of children and adolescents and by increasingly younger children.
- Human beings may be born with a need not to feel to be entirely free. Today, the greatest freedom is perhaps the basic aspiration of human beings.
- Human beings may be born with a need to be with others, to help each other, to have a feeling of being members of a community. Today, living in cities and the economy push human beings to have a form of extreme individualism.
- Biological competition among groups may have created a genetically inherited tendency to have a group identity. Today, globalization and developments in social networking technologies may lead to the emergence of a single community which includes all human beings and can make group identity obsolete.
- Human beings may be born with a need to conserve a trace of their past and of the past of their group. Today's culture tends to cancel all types of past.
- Like nonhuman animals, human beings may be born with a tendency to be satisfied when they have what they need to have. Today, the economy pushes human beings to want always new things and to never be satisfied with what they already have.
- Human beings may be born with a tendency to believe in supernatural beings, in a supernatural world, and in an endless life for their soul. Today, science makes this type of beliefs impossible to entertain.
- Human beings may be born with a tendency to believe that they live in a world that they do not entirely know and that their life is partially immersed in mystery. Today, these beliefs are increasingly contradicted by science and by the collection of "big data" made possible by computers.
- Human beings may be born with a need to be with others. Today, Internet and the new digital technologies may give them the impression to be always with others but they may feel alone.
- Human beings may be born with a need to have a private life. Today, Internet and the new digital technologies may make it impossible to have a private life.

We don't know if these hypotheses are right or wrong but human robots can help us to test them. We let a population of robots evolve in an environment which is like the environment of our human ancestors and then the environment changes and it becomes like the environment in which human beings live today. The "ancient" robots live and evolve in an environment in which:

- they mate and have offspring
- they tend to remain with the same mate for a long period of time so that they can collaborate in taking care of their shared offspring
- they are taken care of by their mates, offspring, and other relatives when they are old or ill
- they live near not only to their mates and offspring but also to robots which are more distantly genetically related to them
- they live in close contact with nature
- adolescent robots are told what they should do by adult robots (parents, other adults) and they tend to do what they are told to do
- the robots want a limited number of things and they are satisfied when they have these things
- they tend to act altruistically towards other members of their community because this increases their chances of success when they compete with other communities
- they keep a trace of their past and of the past of their group
- they have an imaginary past and an imaginary future, they expect to live after their death, they believe in supernatural beings and they expect to be helped by these supernatural beings when in need or in danger and to be punished by them if they do not behave as they should behave
- they live in a world which they do not know and do not understand in the scientific sense of these words
- they live with others and are never alone
- they have a private life.

In this environment the robots evolve a number of genetically inherited needs which the environment of "modern" robots makes difficult to satisfy. The modern environment allows the robots to satisfy many important needs that were generally not satisfied in the old environment, like having more food, a longer life, a sophisticated medicine, and an increasing number of useful technologies, and the environment of "modern" robots must reproduce this aspect of today's human environment. But we must ask: What are the consequences of this change in the environment for the robots? Are the "modern" robots happy?

Again, what we are proposing here is only a hypothesis on possible contrasts between genetically inherited needs and the environment in which human beings live today – and will live in the future. We must abandon all our pre-conceptions and use the robots to test this hypothesis as objectively as possible. And we must find an appropriate balance between these two types of environment from the point of view of the robots' happiness. (Of course, this requires that we have a measure of the robots' happiness.) Our hypothesis must not be taken to imply that human beings must live "according to nature". Human beings have an adaptive pattern a crucial component of which is that they modify the environment in which they live. This artificial environment is not only a physical environment or a technological environment but it is

also an environment of culturally shaped beliefs, behaviours, and values. Human beings cannot "return to nature" because nature is *not* their "natural" environment but, perhaps, they can create an environment which makes them live a happier life – and human robots can help us to find how this environment should be.

4. Robots can help human beings to understand the impact of science on their life

Human beings have lived for the greater part of their existence without science. *Homo Sapiens* was born in East Africa more than 150 thousand years ago but science began in Greece and in other parts of Eurasia 2–3 thousand years ago and it has become "modern" science only in the last five hundred years in the Western world. Today, under the pressure of the economy and the desire to develop always new technologies, science makes always more rapid progress and it increasingly influences the life of human beings. The consequences of science for human life are positive because science gives a sense of knowing what reality really is and because science makes it possible to construct useful technologies that help human beings to live longer and better. But, aside from these obvious advantages, it makes sense to ask "What is the impact of science on human life?"

If we want to know the impact of science on human life, we must address the question with complete detachment, ignoring values, desires, and fears. We cannot be *for* or *against* science, we cannot a priori judge the impact of science on human life as *good* or *bad,* we cannot be enthusiastic about science or preoccupied with science. We must address the question as a purely scientific question. We propose a number of hypotheses concerning the consequences of science for the life of human beings but these hypotheses will need empirical data to know if they are true or false.

Human beings have all sorts of beliefs about the world and about themselves but science makes it difficult to entertain these beliefs if they go against what science knows about reality. This does not apply only to religious beliefs on supernatural entities and supernatural worlds but to all sorts of beliefs that human beings have about themselves and their societies. What science knows about reality is really known only to a minority of human beings – mostly scientists – but it transpires through the media, it is made accessible in principle to everybody by the extension of schooling and science education to all children, and it is demonstrated by the existence of a multitude of science-based technologies. Science appears to the eyes of everybody as very powerful and it is taken into great respect because it generally leads to a consensus among its practitioners – scientists – which is rarely seen in other aspects of social life, and because it makes continuous and cumulative progress, which also does not seem to be true for any other human activity. Therefore, human beings will continue to have all sorts of beliefs about the world and about themselves but, while in the past

they were free to choose what to believe, today they have to take into account what science tells them about reality.

It is not only a matter of beliefs. Human beings have the beliefs that they have about the world and about themselves because having those beliefs makes them happier. Therefore, if science obliges them to abandon their cherished beliefs about the world and themselves, they may be less happy. Science ignores – and it must ignore – if what it discovers about reality is good or bad for human beings. But if science tells human beings that they are only physical matter, that there is no life after death, that there are no supernatural beings and no supernatural worlds which go beyond the finitude of their life and their world, that they are one species of animals among others and have nonhuman ancestors, that their love towards their children and towards their companion is only the result of the tendency of their genes to replicate themselves, that they are selfish rather than altruist, that they are not the centre of world, human beings cannot ignore what science tells them and this can make them less happy. Science has become so powerful that human beings may develop a fear of science and of its capacity to know everything.

Science is very good at making predictions and its ability to make predictions which are confirmed by empirical data is the clearest proof of its strength. This, again, has an influence on the life of human beings. Generally, the ability of science to make predictions is useful to human beings but in some cases it can be a problem, for example, if science predicts that they will incur some illness or that some other type of bad event will occur in the future – especially if science itself cannot do much to avoid the occurrence of these future events. Human beings have always made "predictions" of future events but they were not based on science and, therefore, human beings were free to make all sorts of predictions and to believe or not believe the predictions made by others. Scientific predictions are different. Science cannot avoid making them, it cannot choose which predictions to make, and all human beings, as soon as they are informed of its predictions, must believe them.

Human beings have a rich mental life which includes living in imaginary worlds, remembering imaginary past worlds, and believing in imaginary future worlds. Science makes it more difficult to do this because science defines very clearly what is true and what is not true, what is reality and what is imagination. In the past human beings had a very limited knowledge of reality – or of what, for science, is reality – so that the separation between imaginary worlds and reality was a blurred one and it was easier to believe in imaginary words. Today, science knows always better what is reality, it *defines* what is reality, and for human beings it has become more difficult to believe in imaginary worlds.

In the past, philosophy and the humanities had an important role in human life. Philosophers tried to find what is the ultimate "truth" about reality, what is the "meaning" of life and reality, and what is the "good life". Historians put together narrations of the past history of human societies – especially "civilized" human societies – and

"interpreted" these histories, often using history to confirm the identity and, in many cases, the superiority of their ethnic, cultural, or political community. Intellectuals discussed and wrote about everything that concerns human beings and human societies, proposing and defending interpretations of social phenomena and solutions to the problems of the society. Students of art and literature analyzed, discussed, and interpreted works of art and literature and artistic traditions and movements. The books written by philosophers, historians, intellectuals, and students of art were not read by many people but they were taught in schools and universities and they had an indirect influence on the whole of society.

With the advancement of science the role and the place in society of philosophy and the humanities are changing. Science tends to make philosophy obsolete because, while scientific theories generate predictions which, if confirmed by the empirical data, are a proof of the correctness of the theory, what philosophy says about reality can always be debated and it is impossible to find an agreement on which of the many different philosophies is the right one. Another difference between science and philosophy is that science separates facts from values while in philosophy facts and values are so strictly linked together that it is almost impossible to distinguish between them. Since science shows that reality can only be known only if we separate facts from values, what reality is from what we would like it to be, this may imply the end of philosophy as an enterprise aimed at knowing reality. Philosophy could survive as an enterprise aimed at debating values but, to realize these values, we must *know* reality, and it is science, not philosophy, that knows reality.

Science has another consequence for philosophy. One important branch of philosophy takes science as its object of study. Philosophy of science is a philosophical reflection on what is science, what are the "foundations" of science, what science can know about reality, how science has changed in the past and continues to change today, what makes some scientific disciplines different from other scientific disciplines. Science makes it possible to have a science of science. It is science, not philosophy, that will tell us what is science as a human activity among others. And, what is even more important, science will also tell us what is philosophy. A science of philosophy can generate clear and unambiguous hypotheses about why human beings have philosophy and it can test these hypotheses with objective data on human evolution, the human body and brain, and the social and cultural environment. What will be the consequences for philosophy of having a science of philosophy is not clear but, if science tells us what is philosophy, it may become impossible to look at philosophy as we have looked at it in the past.

As for the rest of the humanities – historians, intellectuals, and students of the arts – science is increasingly invading their territories. Science is showing that the history of human societies can not only be narrated and interpreted – which is what historians do – but it can also be explained as science explains all other phenomena of reality. Historians will continue to collect and interpret written documents and other

artefacts of past human societies but it is science that will let us really understand and explain past human societies. Intellectuals will continue to study and interpret society and suggest how society should be organized, but science knows society better because it clearly distinguishes facts from values and it tells us what we should do to actually change human societies. As for students of art, their work will continue in so far as they themselves are "artists" and what they do is itself "art" but science, by using its tools, is increasingly understanding art and its roots in the human mind, brain, and society, and if science tells us why human beings create artistic artefacts and expose themselves to artistic artefacts, we may not be able to look at artistic artefacts as we did in the past.

Science may have an impact not only on how to study and understand art but also on art itself. As we have said in Chapter 11, science destroys the "halo" that things have around them while art reveals this "halo" and artistic artefacts are artefacts that make it possible to share this "halo" with others. By destroying the "halo" that things have around them, science can make more difficult to create artistic artefacts.

Philosophy and the humanities are particularly interested in their past while for science its past does not exist and science is only the most recent science. Science and technology make continuous progress while this cannot be said for philosophy and the humanities, and they are responsible, together with the marketing economy, for the exclusive interest of today's human beings for the present.

One of the great strengths of science is that science makes it possible to construct all sorts of technologies which have useful applications and economic value. In fact, the impact of science on human life is mainly an indirect impact: science creates always new technologies and these technologies are the new environment in which human beings live. But this has an important implication. Science needs freedom. Scientists must be free to study everything which they consider as important for understanding reality. Given the strict link between science and technology, this freedom is inherited by technology and this is why, today, technology is a "tiger without a lace": it goes wherever it wants to go and human beings have no power to direct its development or to block its development when necessary.

Science is the triumph of the cognitive half of the mind. The cognitive half of the mind is incorporating in one's brain the co-variations which exist in the environment and in one's interactions with the environment and, as we have said in Chapter 13, science is a human activity specifically dedicated to capturing and explaining these co-variations. The human mind will continue to be made by two halves, the cognitive half and the motivational/emotional half, but the motivational/emotional half of the mind will have to recognize the enormous increase in the importance of the cognitive half which is implicit in the constant progress of science.

Values belong to the motivational/emotional half of the mind but science is extraneous to values and it *must* be extraneous to values because it can only know and

understand reality if it ignores values, desires, and fears. But since science has an increasing impact on human life, its attitude towards values will be transferred to the "cultural climate" of human societies, and human beings will have to choose between ignoring science – for example, what science has to say about interindividual differences, sexual differences, racial differences, justice, violence and wars – or abandoning their values.

In the past one spoke of "two cultures": the culture of the natural sciences and the culture of the humanities. Then, a "third culture" was invented which is the culture of *natural* scientists who are interested in society and in its future. More recently, the "two cultures" have become the "three cultures": the natural sciences, the humanities and, in between, the behavioural and social sciences. But if our hypotheses on the impact of science on human life are correct, human beings and their societies will soon have only "one culture": the culture of science as the formulation of clear and unambiguous theories and the collection of facts that confirm the quantitative predictions derived from these theories.

A robotic science of human beings contributes to the emergence of this single culture. As we have tried to show in this book, a robotic science of human beings will make the science of human beings as strong as the science of nature. The weakness of the traditional behavioural and social sciences made it possible for human beings to look at themselves and at their societies without taking science into consideration. A robotic science of human beings will make this more difficult for them to do. The English poet Thomas Stearns Eliot wrote that "human kind cannot bear very much reality". Science tells human beings what reality really is. The poet's words are true for science in general but they will be especially true when robots will reveal to human beings what they really are.

Again, these are only hypotheses about the impact of science on human life and a robotic science of human beings should help us to test these hypotheses. We construct societies of human robots that do not have science and then, gradually, science emerges and it makes constant and increasingly accelerated progress. How is the life of robots that do not have science? How is the life of robots that have science? Are the hypotheses we have made on the impact of science on human life true or false? And, if we have a robotic measure of happiness so that we can say if a robot is happy or unhappy, we can also know if science makes human beings happy or unhappy.

5. Robotic societies as tools for voters

One crucial problem – *the* crucial problem – of today's democracies is that citizens vote without knowing and understanding most of the issues on which the people whom they elect will be called to take decisions and, therefore, they vote based on their ideologies, their personal and short-term interests, and the capacity of candidates to convince them. Today's societies are very complex societies and any decision on how they must function tends to have so many consequences that, inevitably, some of these consequences will be good and some bad. (On this, see Chapter 10 dedicated to political robotics.) Most voters do not have any idea of this complexity, and traditional tools for knowing and understanding how society functions such as books, newspapers, lectures, and television, use a communication tool, language, which is not motivating, is beyond the capacity to understand of most voters, and assigns a passive role to the reader or hearer. Today, the Internet makes it much easier to exchange information, to propose and discuss ideas with other people, and to organize political action. (On the role of communication in helping groups of robot to reach their goals, see Chapter 10.) This is called e-democracy. But e-democracy does not really "empower" citizens because citizens still do not understand how the society functions and do not know how it should be modified to function better. Therefore e-democracy, in its current version, cannot replace traditional representative democracy.

To really change political life, e-democracy needs "digital environments". Digital environments are computer simulations which – completely free from pre-conceptions and ideological commitments – allow citizens to know and understand how human societies function and what are the consequences of taking this or that decision. Digital environments are accessible to everyone by using any kind of devices and they are non-verbal, strongly interactive, and motivating as computer games. They function as experimental laboratories in which the user, alone or with others, does X and sees what are the consequences of doing X, finds out what must be done to obtain some desired result, and experiments with non-existing but possible societies. They are teacher-less learning environments which make it possible to learn and understand by doing and not by being told. For millennia human beings have learned and understood by doing but, then, the complexity of what they had to learn and understand has made verbal language necessary as a tool for learning. Today, computers make it possible to learn and to understand without language. Learners learn and understand not because they read or listen to words but because they do things and see the consequence of what they do. Digital environments can be very simplified or more realistic and they may exist in different versions for different types of users. They are the basis of a new type of learning, from schools to universities, from professional formation to training and re-training of all sorts of workers. And they can also be useful to policy makers, public administrators, and private enterprises. But their most important implication is for the political life of democratic societies.

By making available to voters digital environments which reproduce the economic system, the financial system, the fiscal system, the judiciary system, the political and administrative systems, how the state uses the money of the taxpayers, the consequences of human behaviour for the environment, globalization and its consequences, voters can interact with these digital environments and, through these interactions, they can know and understand what are the issues and what are the different options, they can ask political candidates to better explain what they will do, and they can propose what politicians should do and what problems they should address. And they can also know what politicians have done or not done after they have been elected. In this way, voters can vote as "real" citizens and they will be more motivated to vote. Digital environments change politics into policy. While science cannot be very useful to politics, it can be very useful to policy.

6. Difficult Problems

And, finally, robots can help human beings to better understand the many Difficult Problems they face today and will face in the future and, perhaps, to find solutions to these problems. Difficult Problems are problems which are difficult to recognize as problems, to understand, and to solve. They are difficult to *recognize* as problems because we tend not to think in such a radical way as is required by Difficult Problems and because we often believe we already have a solution for a Difficult Problem. Difficult Problems are difficult to *understand* because they are the result of a multiplicity of interacting causes and factors that seems to go beyond our cognitive capacities and that science with its disciplinary divisions is not prepared to address. Difficult Problems are difficult to *solve* because they require radical changes in the society, in the economic system, in our life, and in our minds. Difficult Problems are problems for all human beings who live on the Earth today. Problems that are specific to one particular country or one particular part of the world are not Difficult Problems, even if Difficult Problems can be more serious in some places and in some countries. Are Difficult Problems political? It depends on what is meant by "political". If "political" refers to the political system, the answer is No because the political system tends to ignore Difficult Problems and is not equipped to deal with them. But if "political" refers to the "polis" (in Greek, the "city"), that is, to all the members of a community – and, in fact, to all the members of the human community – the answer is Yes. They are the most important political problems.

Difficult Problems are the problems of the future. The future will necessarily be different from the present, whether we like it or not. Therefore, Difficult Problems require new ideas and new mental attitudes because, as Albert Einstein said, "the problems that exist in the world today cannot be solved by the level of thinking that created them". However, the ancient god that must guide our search is the Roman god Janus who had

two faces, one looking to the past and the other to the future. Difficult Problems are the problems of the future but they have a long past.

Here is a list of Difficult Problems:

- Is a market economy the best possible economy? Can we imagine an economic system for today's societies which is not a market economy?
- The financial economy is necessary to guarantee that the real economy has sufficient resources to produce all the goods it produces. Is there a way to guarantee that the real economy can make all the necessary investments without the excessive risks and misbehaviours of the financial economy that today are under the eyes of everybody?
- Banks and financial organizations determine most of what happens in today's world and they are becoming supranational entities that have more power than the states. What will be the consequences?
- Is modern economy becoming an economy in which information is the most important good? Is the science of economics prepared to deal with an information economy?
- Does the science of economics create modern economies? If this is true and if we want to change modern economies, how should the science of economics change?
- Is marketing the "first cause" of both the real economy and the financial economy?
- Marketing from an exclusively economic phenomenon has become an all-encompassing phenomenon which is transforming the entire society into a marketing society. With what consequences?
- How can we solve the conflict of interests which exists in many professions (economists, physicians, psychologists, teachers, journalists, media people, marketing people) that could help society to understand and solve many of its problems but prefer not to do so because this would go against their interests?
- The educational institutions as we know them are at the end of their long history. How can we construct a new system of education which is more at home with today's society? What is the role of the new digital technologies in this new system?
- Should we change not only how students learn but also what students learn?
- Are the new digital technologies creating a new "ecology of the mind" and a new social life? What are the consequences?
- What are the consequences of being able to construct artefacts endowed with a "mind" similar to the human mind and, therefore, with both a cognitive half and a motivational/emotional half? Should we construct such artefacts?
- What is the impact of computers on science and technology?
- Are science and technology making human societies too complex for human beings to understand them?
- Can the development of technology be guided?

- Today a process of globalization is underway in the economic, technological, and cultural areas but political power continues to be in the hands of the states that remain local entities, and this asymmetry creates many problems. What should be done? Should we – and can we – arrive at a global sovereign state?
- Is it true that Western culture is becoming the culture of all human beings who live on the Earth? What will be the consequences of this global Westernization?
- Is it possible for non-Western countries to have individual freedom, to eliminate poverty, to have a modern medicine and modern technologies and, at the same time, to preserve their traditional culture? Or the memory of their culture can only become something that can be "sold" like any other good or an illness of the mind?
- Is it necessary that humans behave selfishly? How can we make them behave less selfishly?
- Will wars become things of the past?
- What are the causes of ethnocentrism as the tendency to behave differently towards the members of one's group and towards the members of other group? How can we avoid the negative consequences of this tendency when globalization leads to an increase in the contacts between different groups?
- Behaviours that damage others play a very important role in human societies and, since human societies change always more rapidly, there are always new types of behaviours that damage others. How can we adapt our systems of laws so that they are able to deal with new forms of behaviours that damage others?
- How can we contain socially damaging behaviours without punishing them?
- The impact of human behaviour on the environment is increasing and it represents a danger for human beings. How can we reduce this impact and its consequences?
- Cities change the way in which human beings live. What are the consequences of these changes? What can be done to avoid these consequences if they are bad?
- Is it possible to design a society in which women do not have less power than men but are not forced to become men? Should women become soldiers or should women abolish wars?
- What should a society do with respect to the differences – all sorts of differences – among individuals?
- The Western culture is the culture of freedom. Are people living in Western-type societies really free?
- Freedom, understood as the complete absence of limitations on an individual, produces happiness for the individual?
- Adolescents today tend to reject all authority (of parents, teachers, adults generally, the past) and to grow up without any external "scaffolding" besides peers and the media. What are the consequences of this manner of growing up? How can we avoid its negative consequences?

- How can we preserve the advantages of democracy and at the same time avoid its defects, such as the fact that political chiefs tend to do what voters want them to do and not to pursue the "common good"? How political chiefs can have a longer temporal perspective than the period in which they are in charge?
- Can human beings live in large communities without a "chief" who has the power to make them do what they must do? Can e-democracy replace representative democracy? Do citizens need to better know and understand reality for e-democracy to succeed?
- Is defending one category of human beings against another – the poor versus the rich, the weak versus the strong, those who do not have freedom against those who negate them freedom – a thing of the past because, today, all human beings have the same problems and they must be mostly defended from themselves?
- How can "progress" be measured? Is science is the only thing that makes progress because each day we know more about reality – in the sense in which science understands reality – while this is not necessarily true for technology because progress in technology means that technology lets us live a better life, and it is not always true that technology let us live a better life?
- How can we solve the contrast between the fact that science and technology have an increasing impact on society and the fact that almost all the members of the society know little or nothing about science and technology?
- Those who govern human societies ignore most of the most serious problems that afflict human beings today – our Difficult Problems. How can this be changed?
- Which new institutions should be created to help who governs the society to make a greater and better use of science and technology? Should these institutions include both the natural sciences and the behavioural and social sciences?
- What will be the future of art, religion, and philosophy?
- What will be the future of human beings? If we want to influence their future, should we look at Western culture from outside as one culture among others and as a culture that, in the future, will necessarily become a different culture?
- Are human beings the culmination of a general process of evolution of reality or are they only one stage of this process? Can they predict what will be the next stage? Can they control what will be the next stage?

If these are the Difficult Problems, Difficult Problems necessarily require the contribution of the sciences of human behaviour and human societies, and even those that seem to be in the exclusive province of the natural sciences and their technologies, such as those that concern the environment, energy, medicine, transport, use of territory, use of natural resources, have important behavioural and social components. Therefore, we should ask a robotic science of human beings and human societies to help us to identify, understand and, possibly, solve these problems. We must construct imaginary robotic societies that allow us to experiment with solutions to Difficult Problems. Human robots are a better science of human behaviour and human societies but they

should also allow us to construct scientific utopias. Creating the future is always, in some sense, to imagine utopias. But the word "utopia" designates a society which does not exist and cannot exist and this is why utopias are considered as irrelevant and they are looked at ironically. Today, the meaning of the word "utopia" changes. In the past, we only had only *what exists* and *what does not exist*. Today, there is a third possibility: something can exist in a computer. Robotic societies exist in a computer. By constructing them, by analyzing them, and by doing experiments with them, we can see if they can be realized outside the computer, why they can't, what are their consequences, how we should change them so that they can be realized and have positive consequences.

Difficult Problems remain difficult problems. But, according to the Greek philosopher Heraclitus, "If you do not hope for what cannot be hoped for, you won't find it". Human robots can help us to hope for what cannot be hoped for.

References and additional readings

This is a list of references, grouped by topics, that provide more detailed information on some of the research described in the book. For some of the topics the list also includes a few additional references to relevant work by other authors.

Living in a "natural" environment

Parisi, D., Cecconi, F. & Nolfi, S. (1990). Econets: Neural networks that learn in an environment. *Network*, 1, 149–168. DOI: 10.1088/0954-898X/1/2/003

Nolfi, S. & Parisi, D. (1993). Self-selection of input stimuli for improving performance. In G. A. Bekey (Ed.), *Neural networks and robotics* (pp. 403–418). Berlin: Kluwer. DOI: 10.1007/978-1-4615-3180-7_23

Parisi, D. (1994). Are neural networks necessarily passive receivers of input? In F. Masulli, P. G. Morasso & A. Schenone (Eds.), *Neural networks in biomedicine* (pp. 113–124). Singapore: World Scientific.

Parisi, D. & Cecconi, F. (1995). Learning in the active mode. In F. Moràn, A. Moreno, J. J. Merelo & P. Chacòn (Eds.), *Advances in artificial life. Third European Conference on Artificial Life* (pp. 439–462). London: Springer.

Menczer, F. & Belew, R. K. (1996). From complex environments to complex behaviours. *Adaptive Behaviour*, 4, 317–363. DOI: 10.1177/105971239600400305

Parisi, D. (1997). Active sampling in evolving neural networks. *Human Development*, 40, 320–324. DOI: 10.1159/000278734

* * *

Duchon, A. P., Kaelbling, L. P. & Warren, W. H. (1998). Ecological robotics. *Adaptive Behaviour*, 6, 473–507. DOI: 10.1177/105971239800600306

Arkin, R. C., Cervantes-Perez, F. & Weitzenfeld, A. (1998). Ecological robotics: A schema-theoretic approach. In R. C. Bolles, H. Bunke & H. Noltemeier (Eds.), *Intelligent robots: Sensing, modelling and planning* (pp. 377–393). Singapore: World Scientific.

Ikegami, T. (2009). Rehabilitating biology as a natural history. *Adaptive Behaviour*, 17, 325–328. DOI: 10.1177/1059712309340855

The embodied and action-based nature of knowledge

Nolfi, S. & Parisi, D. (1999). Exploiting the power of sensory-motor coordination. In D. Floreano, J-D. Nicoud, & F. Mondada (Eds.), *Artificial Life 1* (pp. 173–182). London: Springer. DOI: 10.1007/3-540-48304-7_24

Schlesinger, M., Parisi, D. & Langer, J. (2000). Learning to reach by constraining the movement search space. *Developmental Science*, 3, 67–80. DOI: 10.1111/1467-7687.00101

Borghi, A. M., Di Ferdinando, A., & Parisi, D. (2002). The role of perception and action in object categorization. In J. A. Bullinaria & W. Lowe (Eds.), *Connectionist models of cognition and perception* (pp. 40–50). Singapore: World Scientific.

Di Ferdinando, A. & Parisi, D. (2005). Internal representations of sensory input reflect the motor output with which organisms respond to sensory input. In A. Carsetti (Ed.), *Seeing and thinking* (pp. 58–63). Berlin: Kluwer.

Caligiore, D., Borghi, A. M., Parisi, D. & Baldassarre, G. (2010). TRoPICALS: A computational embodied neuroscience model of experiments on compatibility effects. *Psychological Review*, 117, 1188–1228. DOI: 10.1037/a0020887

* * *

Steels, L. (1994). The artificial life roots of artificial intelligence. *Artificial Life*, 1, 75–110. DOI: 10.1162/artl.1993.1.1_2.75

Chiel, H. & Beer, R. (1997). The brain has a body: Adaptive behaviours emerge from interactions of brain, body, and environment. *Trends in Neurosciences*, 20, 553–557. DOI: 10.1016/S0166-2236(97)01149-1

Metta, G. & Fitzpatrick, P. (2003). Better vision through manipulation. *Adaptive Behaviour*, 11, 109–128. DOI: 10.1177/10597123030112004

Chemero, A. & Turvey, M. T. (2007). Gibsonian affordances for roboticists. *Adaptive Behaviour*, 15, 473–480. DOI: 10.1177/1059712307085098

Pfeifer, R. & Bongard, J. C. (2007). *How the body shapes the way we think*. Cambridge, MA: MIT Press.

Evolution and learning

Cecconi, F. & Parisi, D. (1991). Evolving organisms that can reach for objects. In J. A. Meyer & S. W. Wilson (Eds.), *From animals to animals 1* (pp. 391–399). Cambridge, MA: MIT Press.

Parisi, D., Nolfi, S. & Cecconi, F. (1992). Learning, behaviour, and evolution. In F. Varela & P. Bourgine (Eds.), *Toward a practice of autonomous systems* (pp. 207–216). Cambridge, MA: MIT Press.

Nolfi, S., Elman, J. L. & Parisi, D. (1994). Learning and evolution in neural networks. *Adaptive Behaviour*, 3, 5–28. DOI: 10.1177/105971239400300102

Lund, H. H. & Parisi, D. (1995). Pre-adaptation in populations of neural networks evolving in a changing environment. *Artificial Life*, 2, 179–197. DOI: 10.1162/artl.1995.2.2.179

Parisi, D. & Nolfi, S. (1996). The influence of learning on evolution. In R. K. Belew & M. Mitchell (Eds.), *Adaptive individuals in evolving populations* (pp. 419–428). Readings, MA: Addison-Wesley.

Miglino, O., Nolfi, S. & Parisi, D. (1996). Discontinuity in evolution: How different levels of organization imply pre-adaptation. In R. K. Belew & M. Mitchell (Eds.), *Adaptive individuals in evolving populations* (pp. 399–415). Readings, MA: Addison-Wesley.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996). *Rethinking innateness. A connectionist perspective on development*. Cambridge, MA: MIT Press.

Nolfi, S. & Parisi, D. (1997). Learning to adapt to changing environments in evolving neural networks. *Adaptive Behaviour*, 5, 75–98. DOI: 10.1177/105971239600500104

Nolfi, S. (2000). How learning and evolution interact: The case of a learning task which differs from the evolutionary task. *Adaptive Behaviour*, 7, 231–236. DOI: 10.1177/105971239900700205

Nolfi, S. & Floreano, D. (2000). *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*. Cambridge, MA: MIT Press.

Calabretta, R., Nolfi, S., Parisi, D. & Wagner, G. P. (2000). Duplication of modules facilitates the evolution of functional specialization. *Artificial Life*, 6, 69–84. DOI: 10.1162/106454600568320

Di Ferdinando, A., Calabretta, R., & Parisi, D. (2001). Evolving modular architectures for neural networks. In R. French & J. Sougné (Eds.), *Proceedings of the sixth neural computation and psychology workshop: Evolution, learning, and development* (pp. 253–262). London: Springer.

Parisi, D. (2003). Evolutionary psychology and Artificial Life. In S. J. Scher & F. Rauscher (Eds.), *Evolutionary psychology: Alternative approaches* (pp. 243–265). London: Springer. DOI: 10.1007/978-1-4615-0267-8_12

Calabretta, R. & Parisi, D. (2005). Evolutionary connectionism and mind/brain modularity. In W. Callabaut & D. Rasskin-Gutman (Eds.), *Modularity. Understanding the development and evolution of complex natural systems* (pp. 309–330). Cambridge, MA: MIT Press.

Floreano, D., Dürr, P. & Mattiussi, C. (2008). Neuroevolution: from architectures to learning. *Evolutionary Intelligence*, 1, 47–62. DOI: 10.1007/s12065-007-0002-4

* * *

Ackley, D. H. & Littman, M. L. (1992). Interactions between learning and evolution. In Langton, C., Farmer, J., Rasmussen, S. & Taylor, C. (Eds.), *Artificial Life 2* (pp. 487–510). Redwood City, CA: Addison-Wesley.

Cliff, D., Husbands, P. & Harvey, I. (1993). Explorations in evolutionary robotics. *Adaptive Behaviour*, 2, 73–110. DOI: 10.1177/105971239300200104

Harvey, I., Di Paolo, E., Wood, R., Quinn, M. & Tuci, E. (2004). Evolutionary robotics: A new scientific tool to study cognition. *Artificial Life*, 11, 79–98. DOI: 10.1162/1064546053278991

Development

Cangelosi, A., Parisi, D. & Nolfi, S. (1994). Cell division and migration in a 'genotype' for neural networks. *Network*, 5, 497–515.

Nolfi, S., Miglino, O. & Parisi, D. (1994). Phenotypic plasticity in evolving neural networks. In D. P. Gaussier & J-D. Nicoud (Eds.), *From perception to action* (146–157). Los Alamitos, CA: IEEE Computer Society Press.

Nolfi, S. & Parisi, D. (1995). Evolving artificial neural networks that develop in time. In F. Moran, A. Moreno, J. J. Merelo, & P. Chacòn (Eds.), *Advances in Artificial Life. Proceedings of the third European conference on Artificial Life* (pp. 353–367). London: Springer.

Parisi, D. (1996). Computational models of developmental mechanisms. In R. Gelman & T. K. Au (Eds.), *Perceptual and cognitive development* (pp. 373–412). San Diego, CA: Academic Press.

Parisi, D. & Nolfi, S. (2001). Development in neural networks. In J. P. Mukesh, V. Honavar & K. Balakrishan (Eds.), *Advances in evolutionary synthesis of neural networks* (pp. 215–246). Cambridge, MA: MIT Press.

Schlesinger, M. & Parisi, D. (2001). The agent-based approach: A new direction for computational models of development. *Developmental Review*, 21, 121–146.
Parisi, D. & Schlesinger, M. (2002). Artificial Life and Piaget. *Cognitive Development*, 17, 1301–1321.
Cangelosi, A., Nolfi, S. & Parisi, D. (2003). Artificial life models of neural development. In Kumar, S. & Bentley, P. J. (Eds.), *On Growth, form, and computers* (pp. 339–354). San Diego, CA: Academic Press.
Schlesinger, M. (2003). A lesson from robotics: Modeling infants as autonomous agents. *Adaptive Behavior*, 11, 97–107.
Schlesinger, M. (2004). Evolving agents as a metaphor for the developing child. *Developmental Science*, 7, 158–164.
Schlesinger, M., & McMurray, B. (2012). The past, present, and future of computational models of cognitive development. *Cognitive Development*, 27, 326–348.
Cangelosi, A. & Schlesinger, M. (2014). *Developmental robotics: From babies to robots*. Cambridge, MA: MIT Press.

Motivations and emotions

Cecconi, F. & Parisi, D. (1992). Neural networks with motivational units. In *From Animals to Animats 2* (pp. 167–181). Cambridge, MA: MIT Press.
Parisi, D. (1996). Motivation in artificial organisms. In G. Tascini, F. Esposito, V. Roberto & P. Zingaretti (Eds.), *Machine learning and perception* (pp. 3–19). Singapore: World Scientific.
Mirolli, M. & Parisi, D. (2003). Artificial organisms that sleep. In W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, & J. Ziegler (Eds.), *Proceedings of the seventh European conference on Artificial Life* (pp. 377–386). London: Springer.
Parisi, D. (2004). Internal robotics. *Connection Science*, 16, 325–338. DOI: 10.1080/09540090412331314768
Ruini, F., Petrosino, G., Saglimbeni, F., & Parisi, D. (2010). The strategic level and the tactical level of behaviour. In Gray, J. & Nefti-Meziani, S. (Eds.), *Advances in Cognitive Systems* (pp. 271–299). Herts, UK: IET Publisher. DOI: 10.1049/PBCE071E_ch10
Parisi, D. & Petrosino, G. (2010). Robots that *have* emotions. *Adaptive Behaviour*, 18, 453–469. DOI: 10.1177/1059712310388528
Saglimbeni, F., & Parisi, D. (2011). Input from the external environment and input from within the body. In G. Kampis, I. Karsai, & E. Szathmáry (Eds.), *Advances in Artificial Life. Darwin meets von Neumann* (pp. 148–155). London: Springer. DOI: 10.1007/978-3-642-21283-3_19
Parisi, D. (2011). The other half of the embodied mind. *Frontiers in Psychology*, 69, 1–8.
Petrosino, G., Parisi, D. & Nolfi, S. (2013). Selective attention enables action selection: Evidence from evolutionary robotics experiments. *Adaptive Behaviour*, 21, 356–370. DOI: 10.1177/1059712313487389

* * *

Arbib, M. A. & Fellous, J. M. (2004). Emotions: From brain to robot. *Trends in Cognitive Sciences*, 8, 554–561. DOI: 10.1016/j.tics.2004.10.004
Ziemke, T. (2008). The role of emotions in biological and robotic autonomy. *Biosystems*, 91, 401–408. DOI: 10.1016/j.biosystems.2007.05.015

Cos, I., Canamero, L. & Hayes, G. (2013). Learning affordances of consummatory behaviours: motivation-driven adaptation for motivated agents. *Adapative Behaviour*, 18, 285–314. DOI: 10.1177/1059712310375471

Cos, I., Canamero, L., Hayes, G. & Gillies, A. (2013). Hedonic value: enhancing adaptation for motivated agents. *Adaptive Behaviour*, 21, 465–483. DOI: 10.1177/1059712313486817

Language

Parisi, D. (1997). An artificial life approach to language. *Brain and Language*, 59, 121–146. DOI: 10.1006/brln.1997.1815

Cangelosi, A. & Parisi, D. (1998). The emergence of a 'language' in an evolving population of neural networks. *Connection Science*, 10, 83–97. DOI: 10.1080/095400998116512

Parisi, D. & Cangelosi, A. (2002). A unified simulation scenario for language development, evolution, and historical change. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 255–275). London: Springer. DOI: 10.1007/978-1-4471-0663-0_12

Cangelosi, A., & Parisi, D. (2004). The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. *Brain and Language*, 2, 401–408. DOI: 10.1016/S0093-934X(03)00353-5

Mirolli, M. & Parisi, D. (2004). Language, altruism, and docility: How cultural learning can favour language evolution. In J. B. Pollack, M. Bedau, P. Husbands, T. Ikegami & R. A. Watson (Eds.), *Artificial Life 9* (pp. 182–187). Cambridge, MA: MIT Press.

Mirolli, M. & Parisi, D. (2005). How can we explain the emergence of a language that benefits the hearer but not the speaker? *Connection Science*, 17, 307–324. DOI: 10.1080/09540090500177539

Mirolli, M. & Parisi, D. (2006). Talking to oneself as a selective pressure for the emergence of language. In A. Cangelosi, A. D. M. Smith & K. Smith (Eds.), *Proceedings of the sixth international conference on the evolution of language* (pp. 182–187). Singapore: World Scientific.

Mirolli, M., Cecconi, F., & Parisi, D. (2007). A neural network model for explaining the asymmetries between linguistic production and linguistic comprehension. In S. Vosniadou, D. Kayser, & A. Protopapas (Eds.), *Proceedings of the European Cognitive Science Conference 2007* (pp. 670–675). Hillsdale, NJ: Erlbaum.

Floreano, D., Mitri, S., Magnenat, S. & Keller, L. (2007). Evolutionary conditions for the emergence of communication in robots. *Current Biology*, 17, 514–519. DOI: 10.1016/j.cub.2007.01.058

Mirolli, M. & Parisi, D. (2008). How producer bias can favour the evolution of communication: an analysis of evolutionary dynamics. *Adaptive Behaviour*, 16, 27–52. DOI: 10.1177/1059712307087597

Uno, R., Marocco, D., Nolfi, S. & Ikegami, T. (2011). Emergence of proto-sentences in artificial communicating systems. *IEEE Transactions on Autonomous Mental Development*, 3, 146–153. DOI: 10.1109/TAMD.2011.2120608

* * *

Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews*, 8, 339–356. DOI: 10.1016/j.plrev.2011.10.014

Steels, L., & Loetzsch, M. (2012). The Grounded Naming Game. In L. Steels (Ed.), *Experiments in cultural language evolution*. Amsterdam: John Benjamins. DOI: 10.1075/ais.3

Yuruten, O., Sahin, E. & Kalkan, S. (2013). The learning of adjectives and nouns from affordances and appearance features. *Adaptive Behaviour*, 21, 437–451. DOI: 10.1177/1059712313497976

The influence of language on the representation of the world in the human mind

Cangelosi, A. & Harnad, S. (2000). The adaptive advantage of symbolic theft over sensorimotor toil: grounding language in perceptual categories. *Evolution of Communication*, 4, 117–142. DOI: 10.1075/eoc.4.1.07can

Cangelosi, A. & Parisi, D. (2001). How nouns and verbs differentially affect the behavior of artificial organisms. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd annual conference of the Cognitive Science Society* (pp. 170–175). Hillsdale, NJ: Erlbaum.

Mirolli, M. & Parisi, D. (2005). Language as an aid to categorization: A neural network model of early language acquisition. In A. Cangelosi, G. Bugmann & R. Borisyuk (Eds.), *Modelling language, cognition and action. Proceedings of the ninth neural computation and psychology workshop* (pp. 97–106). Singapore: World Scientific.

Mirolli, M. & Parisi, D. (2009). Language as a cognitive tool. *Minds and Machines*, 19, 517–528. DOI: 10.1007/s11023-009-9174-2

Massera, G., Tuci, E., Ferrauto, T. & Nolfi, S. (2010). The facilitatory role of linguistic instructions on developing manipulation skills. *IEEE Computational Intelligence Magazine*, 5, 33–42. DOI: 10.1109/MCI.2010.937321

Mirolli, M. & Parisi, D. (2011). Towards a Vygotskian cognitive robotics: The role of language as a cognitive tool. *New Ideas in Psychology*, 29, 298–311. DOI: 10.1016/j.newideapsych.2009.07.001

Mental life

Cecconi, F. & Parisi, D. (1990). Learning to predict the consequences of one's own actions. In R. Eckmiller, G. Hartmann & G. Hauske (Eds.), *Parallel processing in neural systems and computers* (pp. 237–240). Amsterdam: Elsevier.

Caligiore, D., Tria, M., & Parisi, D. (2006). Some adaptive advantages of the ability to make predictions. In *From Animals to Animats 9* (pp. 17–28). London: Springer. DOI: 10.1007/11840541_2

Parisi, D. (2007). Mental robotics. In Chella, A. & Manzotti, R. (Eds.) *Artificial Consciousness* (pp. 191–211). Exeter, UK: Imprint-Academic.

Sociality

Cecconi, F., Denaro, D., Parisi, D. & Piazzalunga, U. (1994). Social aggregations in evolving neural networks. In C. Castelfranchi & E. Werner (Eds.), *Artificial social systems* (pp. 41–54). London: Springer. DOI: 10.1007/3-540-58266-5_3

Baldassarre, G., Nolfi, S. & Parisi, D. (2003). Evolving mobile robots able to display collective behavior. *Artificial Life*, 9, 255–267. DOI: 10.1162/106454603322392460

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Science*, 101, 2658–2663. DOI: 10.1073/pnas.0400054101

Castellano, C., Cecconi, C., Loreto, V., Parisi, D. & Radicchi, F. (2004). Self-contained algorithms to detect communities in networks. *Eur. Phys. J. B.*, 38, 311–319. DOI: 10.1140/epjb/e2004-00123-0

Parisi, D. & Nolfi, S. (2006). Sociality in embodied neural agents. In R. Sun (Ed.), *Cognition and multi-agent interaction: From cognitive modeling to social simulation* (pp. 328–354). Cambridge: Cambridge University Press.

Baldassarre, G., Parisi, D. & Nolfi, S. (2006). Distributed coordination of simulated robots based on self-organisation. *Artificial Life*, 12, 289–311. DOI: 10.1162/artl.2006.12.3.289

Mitri, S., Wischmann, S., Floreano, D. & Keller, L. (2012). Using robots to understand social behavior. *Biological Reviews*, 88, 31–39. DOI: 10.1111/j.1469-185X.2012.00236.x

Nolfi, S. (2012). Co-evolving predator and prey robots. *Adaptive Behavior*, 20, 10–15. DOI: 10.1177/1059712311426912

Lettieri, N. & Parisi, D. (2013). Neminem laedere: An evolutionary agent-based model of the interplay between punishment and damaging behaviour. *Artificial Intelligence and Law*, 21, 425–453. DOI: 10.1007/s10506-013-9146-y

* * *

Bonabeau, E., Dorigo, M. & Theraulaz, G. (Eds.). (1999). *Swarm intelligence: From natural to artificial systems*. Oxford: Oxford University Press.

Fong, T., Noubakhsh, I. & Dautenhahn, K. (2003). A review of socially interactive robots. *Robotics and Autonomous Systems*, 42, 143–166. DOI: 10.1016/S0921-8890(02)00372-X

Epstein, J. K. & Axtell, R. (2006). *Generative social science. Social science from the bottom up*. Princeton, NJ: Princeton University Press.

Families

Menczer, F. & Parisi, D. (1992). A model for the emergence of sex in evolving networks: Adaptive advantage or random drift? In F. Varela & P. Bourgine (Eds.), *Towards a practice of autonomous systems* (pp. 337–345). Cambridge, MA: MIT Press.

Parisi, D., Cecconi, F. & Cerini, A. (1995). Kin-directed altruism and attachment behaviour in an evolving population of neural networks. In N. Gilbert, & R. Conte (Eds.) *Artificial societies: The computational simulation of social life* (pp. 238–251). London: UCL Press.

Pedone, R. & Parisi, D. (1997). In what kinds of social groups can "altruistic" behaviors evolve? In R. Conte, R. Hegselmann, & P. Terna (Eds.), *Simulating social phenomena* (pp. 195–201). London: Springer. DOI: 10.1007/978-3-662-03366-1_16

Floreano, D., Mitri, S., Perez-Uribe, A. & Keller, L. (2008). Evolution of altruistic robots. *Computational Intelligence: Research Frontiers*, 232–248, LNCS 5050.

Mitri, S., Floreano, D. & Keller, L. (2011). Relatedness influences signal reliability in evolving robots. *Proceedings of the Royal Society B*, 278, 378–383. DOI: 10.1098/rspb.2010.1407

Da Rold, F., Petrosino, G., & Parisi D. (2011). Male and female robots. *Adaptive Behaviour*, 19, 317–334. DOI: 10.1177/1059712311417737

* * *

Todd, P. M. & Miller, G. F. (1993). Parental guidance suggested: How parental imprinting evolves through sexual selection as an adaptive learning mechanism. *Adaptive Behaviour*, 2, 5–47. DOI: 10.1177/105971239300200102

Culture

Cecconi, F., Menczer, F. & Belew, R. (1995). Maturation and evolution of imitative learning in artificial organisms. *Adaptive Behavior*, 4, 29–50. DOI: 10.1177/105971239500400103

Denaro, D. & Parisi, D. (1996). Cultural evolution in a population of neural networks. In M. Marinaro & R. Tagliaferri (Eds.), *Neural nets* (pp. 100–111). London: Springer.

Parisi, D. (1997). Cultural evolution in neural networks. *IEEE Expert*, 12, 9–11. DOI: 10.1109/64.608170

Ugolini, M. & Parisi, D. (1999). Simulating the evolution of artifacts. In D. Floreano, J.-D. Nicoud, & F. Mondada (Eds.), *Advances in Artificial Life* (pp. 489–498). London: Springer. DOI: 10.1007/3-540-48304-7_67

Parisi, D., & Ugolini, M. (2002). Living in enclaves. *Complexity*, 7, 21–27. DOI: 10.1002/cplx.10010

Parisi, D., Cecconi, F., & Natale, F. (2003). Cultural change in spatial environments: The role of cultural assimilation and internal changes in cultures. *The Journal of Conflict Resolution*, 47, 163–179. DOI: 10.1177/0022002702251025

Acerbi, A. & Parisi, D. (2006). Cultural transmission between and within generations. *Journal of Artificial Societies and Social Systems*, 9 (1).

Cecconi, F., Antinucci, F., Parisi, D., & Natale, F. (2006). Simulating the expansion of farming and the differentiation of European languages. In B. Laks & D. Simeoni (Eds.), *Origins and Evolution of Language* (pp. 234–258). Oxford: Oxford University Press.

Acerbi, A., Ghirlanda, S. & Enquist, M. (2014). Regulatory traits: Cultural influences on cultural evolution. In S. Cagnoni, M. Mirolli & M. Villani (Eds.), *Evolution, complexity, and Artificial Life* (pp. 135–148). London: Springer. DOI: 10.1007/978-3-642-37577-4_9

* * *

Curran, D. & O'Riordan, C. (2006). Increasing population diversity through cultural learning. *Adaptive Behaviour*, 14, 315–338. DOI: 10.1177/1059712306072335

Nehaniv, C. L. & Dautenhahn, K. (2007). *Imitation and social learning in robots, humans, and animals*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511489808

Economic and political life

Parisi, D. (1997). What to do with a surplus. In R. Conte, R. Hegselmann, & P. Terna (Eds.), *Simulating social phenomena* (pp. 133–151). London: Springer. DOI: 10.1007/978-3-662-03366-1_10

Cecconi, F. & Parisi, D. (1998). Individual versus social survival strategies. *Journal of Artificial Societies and Social Simulation*, 1(2).

Parisi, D. (1998). A cellular automata model of the expansion of the Assyrian empire. In S. Bandini, R. Serra & F. S. Liverani (Eds.), *Cellular Automata* (pp. 194–200). London: Springer.

Delre, S. A. & Parisi, D. (2007). Information and cooperation in a simulated labour market: A computational model of the evolution of workers and firms. In M. Salzano & D. Colander (Eds.), *Complexity hints for economic policy* (pp. 181–200). London: Springer.

Gigliotta, O., Miglino, O., & Parisi, D. (2007). Groups of agents with a leader. *Journal of Artificial Societies and Social Simulation*, 10, 1–10.

Cecconi, F., di Gennaro, F., Parisi, D. & Schiappelli, A. (in press). Simulating the emergence of proto-urban centres in Ancient Southern Etruria. In J. A. Barcelò (Ed.), *Mathematics and Archaeology*. Enfield, NH: Science Publishers.

Index

A
adjectives 128–130, 136
adverbs 130, 136
art 428–436
attention 64–69

B
body 13–15, 23–49, 53–75, 422–424, 429–430, 440–441

C
category 131–140
chiefs 189–190, 210, 223, 338–348, 350–353, 355
classical conditioning 101, 106, 117
community 202–215, 264–267, 276–281, 317–320, 330–344, 353–355
counting 149, 153–154
cultural globalization 289, 296
culture 259–260, 282–291, 293–297

E
economic inheritance 322–324
economy 362, 369, 372–399
emotions 55–80, 420–455
environment 18–20
evolution 20–21, 80–90, 109–120, 126–136, 185–186
exchange 372–405

F
family 210–225, 234–235, 252–257, 276–278, 301–303, 313–330

H
history 444–450

I
imitation 180–181, 203–205
imprinting 107–109
instrumental learning 101–102, 104–106
inter-individual differences 407–418
internal robotics 38, 68–70, 72, 74–76, 79, 455

K
kin-relatedness 221, 231–235, 302, 318

L
language 121–128, 144–149, 153–159, 180–182, 295–299, 437–442
language production 142, 146–148
language understanding 123, 141–142, 147–148
learning 80–120, 144–152, 271–274, 277, 280–282
living together 189–193, 196–199, 211–213

M
males and females 235–239, 244–253
mental life 159–163, 184–186
motivation 34–55, 216–218, 364–366, 416–418, 420–421, 428–429

N
nouns 126–130, 135, 140, 149

P
pathological 62–63, 419–424
pathologies 419–425
philosophy 10–12, 427–428, 436–439, 468–470
political 338–341, 444–447, 472–473, 475–476
predicting 159, 165–169, 180–184

R
religion 427–428, 433–439
robotic metaphysics 436–439
robots as practical applications 13–15, 451–452
robots as science 13–15, 451–452, 461–462

S
social behaviour 182–193, 215–216
social environment 215–219
social life 187–189, 217–220, 418–419
sociality 188–193, 195–196, 220–222
society 339–344, 352–355, 404–405, 469–477
socially damaging behaviour 199, 202–210, 352–355
stores 253–257, 274–275, 281, 301–344, 373–396

V
verbs 126–130, 136, 155