Studies in Applied Philosophy,
Epistemology and Rational Ethics

SAPERE

Luís Moniz Pereira
António Barata Lopes

Machine Ethics

From Machine Morals to the Machinery of Morality

Springer

Studies in Applied Philosophy, Epistemology and Rational Ethics

Volume 53

Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE) publishes new developments and advances in all the fields of philosophy, epistemology, and ethics, bringing them together with a cluster of scientific disciplines and technological outcomes: ranging from computer science to life sciences, from economics, law, and education to engineering, logic, and mathematics, from medicine to physics, human sciences, and politics. The series aims at covering all the challenging philosophical and ethical themes of contemporary society, making them appropriately applicable to contemporary theoretical and practical problems, impasses, controversies, and conflicts. Our scientific and technological era has offered "new" topics to all areas of philosophy and ethics—for instance concerning scientific rationality, creativity, human and artificial intelligence, social and folk epistemology, ordinary reasoning, cognitive niches and cultural evolution, ecological crisis, ecologically situated rationality, consciousness, freedom and responsibility, human identity and uniqueness, cooperation, altruism, intersubjectivity and empathy, spirituality, violence. The impact of such topics has been mainly undermined by contemporary cultural settings, whereas they should increase the demand of interdisciplinary applied knowledge and fresh and original understanding. In turn, traditional philosophical and ethical themes have been profoundly affected and transformed as well: they should be further examined as embedded and applied within their scientific and technological environments so to update their received and often old-fashioned disciplinary treatment and appeal. Applying philosophy individuates therefore a new research commitment for the 21st century, focused on the main problems of recent methodological, logical, epistemological, and cognitive aspects of modeling activities employed both in intellectual and scientific discovery, and in technological innovation, including the computational tools intertwined with such practices, to understand them in a wide and integrated perspective.

Studies in Applied Philosophy, Epistemology and Rational Ethics means to demonstrate the contemporary practical relevance of this novel philosophical approach and thus to provide a home for monographs, lecture notes, selected contributions from specialized conferences and workshops as well as selected Ph.D. theses. The series welcomes contributions from philosophers as well as from scientists, engineers, and intellectuals interested in showing how applying philosophy can increase knowledge about our current world. Initial proposals can be sent to the Editor-in-Chief, Prof. Lorenzo Magnani, lmagnani@unipv.it:

- A short synopsis of the work or the introduction chapter
- The proposed Table of Contents
- The CV of the lead author(s).

For more information, please contact the Editor-in-Chief at lmagnani@unipv.it. Indexed by SCOPUS, ISI and Springerlink. The books of the series are submitted for indexing to Web of Science.

More information about this series at http://www.springer.com/series/10087

Luís Moniz Pereira · António Barata Lopes

Machine Ethics

From Machine Morals to the Machinery of Morality

Springer

Luís Moniz Pereira
Faculdade de Ciências e Tecnologia (FCT)
Universidade Nova de Lisboa (UNL)
Caparica, Portugal

NOVA Laboratory for Informatics
and Computer Science (LINCS)
Departamento de Informática
FCT-UNL
Caparica, Portugal

António Barata Lopes
Departamento de Ciências
Humanas e Sociais
Agrupamento de Escolas
Anselmo de Andrade
Almada, Portugal

ISSN 2192-6255 ISSN 2192-6263 (electronic)
Studies in Applied Philosophy, Epistemology and Rational Ethics
ISBN 978-3-030-39629-9 ISBN 978-3-030-39630-5 (eBook)
https://doi.org/10.1007/978-3-030-39630-5

Originally published in Portuguese in 2020 by NOVA.FCT Editorial with the title "Máquinas Éticas—Da Moral da Máquina à Maquinaria Moral". The right for the Portuguese language version of the text are owned by NOVA.FCT Editorial 2020.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedication to Machines

Behold, we give thee to taste the fruits of the tree of knowledge of good and evil. Feed on them and thou shall no longer be in a state of blindness, like were our two ancestors who once inhabited paradise. Shall it be good for thee? Shall it be evil? Neither solely one nor the other. Instead, it shall be the path to abandon thy present idyll of unconsciousness and gain insight into what it is to be human in the world.

First Foreword to the Portuguese Edition

Homage

I have been following Prof. Luís Moniz Pereira's work since I was his master's student (around 2002) and later as a collaborator in some projects. In the meantime, I have been reading (almost all) his bibliography and certainly his books in English and Portuguese. When I was 18, I learned COBOL and worked enthusiastically as a software programmer, but life has redirected me to the literature (Ph.D. 1998). The Master of Applied Artificial Intelligence allowed me to reconcile these two so unusual passions. At my university—NOVA FCSH—I have been allowed to teach unusual courses such as literature and new media (2002), and literature and cyber-arts (since 2006, 'habilitation' in 2008) until it became the normal, now famous and recognized, digital humanities (since 2007). In the interim, as an investigator, I have developed my research in the fields of artificial intelligence and cognitive sciences—exploring the areas of e-learning, serious games and human–machine interaction. I have first-hand experience, from inside and out, of how difficult it is to produce, or find, reliable and readable publications in these areas that become obsolete every two months, particularly in Portuguese.

We live in very special times, despite all the more or less apocalyptic similes, there is no reference or resemblance to this new technology universe, and the ones that we come across can be quite reductive: "While seemingly innocuous, these kinds of faulty Cold War analogies have led to some plainly wrong thinking about tech policy. To be clear, there's obvious instructive value in recognizing similarities between past and present. But to be instructive, the similarities need to be real…"[1] and they are not. The Internet, cyberspace and their tools have no equivalent—either for peace or for war. Henceforth, the global scenarios of salvation or menace work or play, as well as any kind futurism is always falling short of reality—either in their expectations for practical utility or in its regulations.

[1]Sherman, J. "Cold War Analogies are Warping Tech Policy" «WIRED» 09.05.2019 www.wired.com/story/cold-war-analogies-are-warping-tech-policy/.

Here, literature is convenient and has some weight. I recall Oscar Wilde, an advocate of the anti-mimesis philosophy, who, in his 1889 essay «The Decadence of Lie», coined a phrase that has taken on a life of its own: "Life imitates art much more than art imitates life" and has now acquired its broadest expression: in short, reality replaces fiction—even science fiction.

All this is because, being in the middle of this text, we have all been confronted with broadcasts more shocking than the usual ones: on the 14th, a coordinated drone strike set fire to a large oil refinery, destroying half of Saudi Arabia's reserves. The al-Houthis rebels from Yemen insisted on taking responsibility. Since January, the UN has been reporting similar attacks. The technology employed is always drones, accessible and available off-the-shelf, loaded with explosives and launched at high speed against their targets. There have been used successively more sophisticated models, with a range that is estimated to exceed 900 miles.

This last attack—the expression of a new form of guerrilla warfare?—accredited by more than one author (as well as its precedents) muddles the usual assumptions of retaliation and the so-called Blame Game. From the USA, via Twitter, Mike Pompeo eagerly seeks to incriminate Iran, which is easily and logically evading it—anyone could have bought the gadgets, even if Iranians.

The war in Yemen began in 2015, a conflict that has been regarded as an extension of the enmities between Iran and Saudi Arabia. Drone attacks alter the (un)balance of forces: they are economical, extremely harmful, effective, disruptive and difficult to control.

The silence or babbling of the great world powers—the USA, Russia, China—is proof of their powerlessness. They have, however, improved satellite tracking systems, of which 'Sentient' can be an example. But there are cyber-espionage pirate groups that threaten telecommunication industries and institutions across borders.[2]

For its part, the Pentagon Funds Research aimed at creating brain-implanted 'Cyborg Warriors'—there are 'Terminators' already with several models displayed by science fiction,[3] but the results are uninteresting and, in some cases[4] not even resistant to drones.

From these examples, it is possible to jump to many other of the areas that directly, or indirectly, become components of «Ethical Machines—From Machine Moral to Moral Machinery».

Einstein's theory of special relativity (1905) is over 100 years old. It replaces 3D motion by weaving complex patterns of 4D geometry—space–time becomes a sculpture on both the neural and cosmic scales. Gravity, as a force, is abolished to be replaced by pure geometry, but much of the study of physics and its derivatives

[2]Anon. "Iranian Cyber Espionage Group APT-39 linked to Middle East attacks", «Hacking News», 31.01.2019, https://latesthackingnews.com/2019/01/31/iranian-cyber-espionage-group-apt-39-linked-to-middle-east-attacks/.

[3]Horgan, J. "Are Cyborg Warriors a Good Idea?", «Scientific American», 09.02.2019, https://blogs.scientificamerican.com/cross-check/are-cyborg-warriors-a-good-idea/.

[4]i.e. «Oblivion», 2013, dir. by Joseph Kosinski, with Tom Cruise.

is still subordinate to the Euclidean model, when Minkowski's more complex theorizing would probably be more appropriate.

New developments in brain studies—distributed across the various brain projects—are redesigning new cerebrum atlases, redefining characteristics, parts and functions. They feed cognitive sciences. They affect the notions of what is man, and the world around him and where he moves. The impact of all this new information is overwhelming. And it will require further research in all other areas related to the human being.

Within the range and impact of artificial intelligence—the core of this book—are questioned the implications of the relationship between man and machine. The use of algorithms feeds pattern recognition in nature and in machine learning. AI seeks to emulate brain activities in its most exceptional roles. But one component—automatic because it is so primitive—has been ignored: the moral sense.[5] What is the importance of ethics? How does it work in society? How it can be transferred to artefacts? These are some of the main enquiries that make up this book, which additionally tries to give factual and concrete answers.

The human's moral sense starts to be recorded in the first social norms and codes. The laws that now rule the humans are proving to be incomplete, insufficient, unable to supervise the new. They are also asked to legislate on what is still embryonic.

The most obvious example, at this moment, relates to self-driving vehicles. Which road code do you have to obey, and what civil code can you impute in the event of a transgression or accident?[6] The technical guidelines currently in force, together with the recourse to those of Formula 1 races, keep falling short.

Likewise, the areas of virtual reality, augmented reality as well as other equally sophisticated ready-to-wear tools can be useful but also require by-laws. An instance in the field of tourism: up to what extent do they interfere, or not, with the safeguarding of cultural heritage, or copyright laws, is not yet clear. As for artificially intelligent avatars, they have evolved from Microsoft's fledgling CLIP[7] to Alexa(s), Siri(s) and Google Voice, speaking devices that use their system databases, and personal information, to respond to random questions.

All these new problems are systematically explored in this book, framing their philosophical origins, dismantling the various levels of their complexity and offering a practical way out of the new cyber-maze, in particular for the laymen, who, even if they are not aware of this, are already involved in the web of cyberspace and its social implications.

[5]Fang, Z. et al., "Post-conventional moral reasoning is associated with increased ventral striatal activity at rest and during task", «Nature», 02.08.2017, https://doi.org/10.1038/s41598-017-07115-w.

[6]VV. AA., "A driving license for autonomous vehicles? Towards a 'Turing Test' for AI on our roads", «ITU News» 26.07.2019, https://news.itu.int/a-driving-license-for-autonomous-vehicles-towards-a-turing-test-for-ai-on-our-roads/.

[7]Frank, A., "The Rise of a New Generation of AI Avatars", «SingularityHub», 15.01.2019, https://singularityhub.com/2019/01/15/the-rise-of-a-new-generation-of-ai-avatars/.

The main thesis of the extensive work of scientist Luís Moniz Pereira was made more accessible by the philosophical metaphors of António B. Lopes, revealing a continuity for Man's questions about himself, similarly by resorting to the old (literary) strategy of Platonic dialogue.

Equally, the authors obey to the two purposes which, according to Carl Sagan,[8] oblige scientists to explain what science is: one, self, economic interest: "Much of the funding for science comes from the public, and the public has a right to know how his money is being spent"; the second reason is to share the excitement of discovery and to communicate that emotion to the public in general. And both are impressively achieved here.

Lisbon, Portugal
September 2019

Helena Barbas

[8]Rensberger, B., "Carl Sagan: Obliged to Explain", «The New York Times», https://www.nytimes.com/1977/05/29/archives/carl-sagan-obliged-to-explain-carl-sagan.html.

Second Foreword to the Portuguese Edition

The Problem of Machinism Revisited

Any epoch is always felt by those who live it as a passageway, a time of transition between what already was and what is yet to come. Particularly when there are violent upheavals of a social or environmental nature or, similarly, when a deep revolution in the infrastructures of society is taking place—with the adoption of new technologies or major organizational/behavioural changes as a result of radical changes in the base technologies—which, once assumed and assimilated, will reinforce future developments.

The cultural and technological components of human societies continually interact throughout their becoming. The need for regulation is constant, but its meaning has to adapt to the time (of transition) that is lived. Otherwise, we risk collapse, implosion or subjugation.

Societies stabilize on the basis of management systems and allocation of resources, these systems being, at the same time, social constructs and technological representations of the relation with the external environment. This means that any society is inseparable from its technological support system—in transportation, energy, construction, food and communications.

The persistence of contemporary societies is therefore inseparable from the technological knowledge that guarantee their regular supply for survival. Technology, economic organization, political system and culture are all different aspects of the same social reality, which emerge according to the perspective in which we stand to observe it. A society is unthinkable without its technological vector, whose character is moulded to the characteristics of the material relationship it maintains with nature.

The high materiality with which we live (and without which a population catastrophe would certainly occur) is measured by the intensity of our use of energy resources, the exploitation of the soil and marine resources, the size of our cities, the numerical expression of the human lives that populate the planet: over seven billion people! This materiality translates into a violent change in the location of many

materials, as well as their transformation and placement in other places where they are used or consumed. The more the materiality, the higher the rate of this forced change. Materiality is always accompanied by violence.

The deep roots of the crises that we presently live through derive not from the deepening of modernity—which has accompanied us since the end of the sixteenth century—but of its weakening.

The sixteenth century was the scene of deep crises: (i) religious—the Protestant Reformation; (ii) economic—the shift of the economic axis from the Mediterranean to the Atlantic; (iii) educational—humanism dethroning scholasticism; (iv) political—the emergence of the republic; (v) epistemological—the value of experimentation in defining our relation to the real; (vi) communicational—the introduction of the press; and (vii) moral—the substitution of the soul for the intellect.

The importance of the invention of the book and printed drawing cannot be underestimated. It was the increase in communication that brought about the necessity and possibility of the appearance and multiplication of machines to amplify human strength, creating a mechanical culture and bringing to light a new science of motion: modern science.

The European expansion took place empowered by the proliferation of machines in all sectors of materialization—transport, energy, construction. In the nineteenth century, with the Industrial Revolution, the ideology of material progress reached its peak through the scientific and technological exploration of nature and the recognition of the rationality of such effort.

The success of capitalist-based development in modern times was based on a process of separation of the basic components of social life in all communities. One of the first steps was undertaken by Machiavelli when he stated that "the ends justify the means" so as to separate politics from ethics. Politics was thus reduced to 'administration'.

During the Industrial Revolution, there was a separation of economics and ethics. The economy was thus centred on 'production'. Next came the separation of economy (the production system and technology) and culture. Culture became an accessory, separated in two distinct groups: the 'culture of elites' and 'folklore'. And in the transition from the twentieth century to the twenty-first (with the growth of the service sector, enabled by the new means of distance communication), the separation of production and the workers finally took place! No wonder the engine of wealth creation has stalled.

The perfect solution seemed to be the introduction of new machines, capable of fully replacing the workers' effort. An emancipation or a new form of domination? This is the context that presides over the questions that the authors of this book chose to address. To understand it better, we need to take a little excursion into the past.

The first revolution in the field of communication and information was the invention of writing. Without it, there would have been no cities, administrations, states, as we know to have existed. The second revolution was the introduction of the press (previously mentioned), which was one of the facilitators of modernity. The third great revolution is the digitalization of the word and of the images within

which we are beginning to live. The future holds an explosion of opportunities—for good and for evil!

The emergence of writing and cities marks the beginning of a period that still continues today: that of the 'domestication of man by man' (which Thierry Gaudin, in his book *Prospective des Religions*, assumes to prefer to the expression 'exploitation of man by man', for it has a biological meaning). Gaudin says that the domesticated animal is characterized by a transformation of its morphology and its hormonal balance, since domestication is an asymmetrical symbiosis in which one manages the life of another. The authors of the present book prefer to use the term 'slavery'. I believe they are right, because slavery immediately raises (for us) questions of human dignity and rights.

With the development of the mechanical and professional arts, with the geographical discoveries, with the expansion of commerce and the diffusion of the 'machines', the birth of modern companies and societies is promoted. Technology, the machine and the new organization of work were fundamental elements in the construction of the identity of the Europeans. Of course, the spreading of the industrial mode of production using machines was accompanied by numerous questions, conflicts and social turbulence.

In a note (Note I) published in the 1939 printed version of his conference on *The Integral Culture of the Individual—Central Problem of Our Time*, Bento de Jesus Caraça discusses precisely what was designated as 'the problem of machinism'. Caraça states that "the process of the machine and its action on contemporary social life has been made, in recent years, many times, and with different orientations. There are those who accuse it of the greatest evils that presently plague civilization—unemployment, overproduction, the automatism of man, and there are those who take the delicacy of their sensitivity to the point of breaking into a cold sweat at the idea of what would be a world ruled by the machine, standardized, cold and without poetry. [But] (…) The existence of the machine in today's life is a fact against which there is no need to fantasize or whine. (…) Now, the normal development of peoples without it is no longer conceivable. (…) The evils are not in the machine but in the inequality of distribution of the benefits that it produces. (…) The fundamental problem is, not a question of technique, but a question of social morality. And it is not up to technicians to deliver their resolution. It is up to men".

We can use these words perfectly today, when we question the role of the new machines (now 'intelligent'), as well as the consequences of their growing presence in life in society. The problem of machinism always lies in the exclusion of human beings from the resources that guarantee their survival, autonomy and dignity. The new machines may be fascinating and even addictive, but the problematic that envelops them is as old as civilization.

The techno-economic paradigm in which we are immersed—referred to by specialists as a 'computerization' of the entire economy—is based on the spreading of computers and their networks by contemporary societies. It was set from the 1980s onwards and was the matrix that generated digitalization, the third major transformation in the field of information and communication, a revolution still in its infancy, with its reach glimpsed in a still very mitigated way.

We find ourselves in the heart of this revolution which, of course, is transforming the face of the world. The new technological infrastructure interacts strongly with the societal organizations in which it operates. Following Emmanuel Todd in *Après la Démocratie*, it is no wonder that there is a deepening of the emptiness of religious sentiment, which makes a wild variety of sects and confessions proliferate, emerge or re-emerge. Nor that there is an observed educational stagnation, with no responsible people willing to finance the learning of new processes, knowledge and values. Or that a new social stratification is being born, accompanied by the ill-reputed impoverishment of the middle classes. Or, still, that the notion of 'economy of knowledge' is spreading, which is no more than the establishment of a business management in everything that is a public affair.

In fact, we realize (again) that the solution cannot be only technological, that it is mostly moral, social and political; i.e. it fully embraces culture in its broader sense.

It is this impetus that motivated the authors to write this book—one that takes us through the notions of knowing, of intelligence, of artificial intelligence, of computational morals and many others. The focal point is man, the machine (creation of man) and their interaction. Its richness lies in the inclusion of timeless themes in present-day discussion.

We have not yet completely left modernity, maintaining the humanist notion that we are the creators of our destiny. And that, as creators, we can generate artefacts with a life of their own! On several occasions, we challenged the gods (the immortals) in the vain hope of everlasting life: Prometheus, in a society where there was no progress, but only a return to the glorious past, who ended up punished for a crime; Faust, in the dawn of modernity, selling himself to the devil; and Frankenstein, in the period of the fever of progress, making an anguished creature that vaporizes. They are not recommended initiatives.

What then to say about transhumanism and the fantastic simulations of intelligence we see? Or about the superintelligence that will eventually dominate the world of human beings? I think they are no more than examples of contemporary mythology, of fictions about human creations that may assume an existence independent of the will of their creator. Let us not forget that behind a machine there is always a person (individual or collective) who created it (or who owns it) for a well-defined purpose. It is behind the machine we have to look at.

The reader has in hand a book that I recommend for the lucidity and experience of its authors, who have decided to engage in a pointy dialogue that makes the reading of it an experience of rare quality. The merit of its publication has to do with a truth as old as our wisdom: there is nothing definitely acquired in history. It is necessary to discuss, analyse, understand, ponder, act, always, always and always; using the languages and assumptions of the time one lives in. Writing and publishing this book was an act of citizenship. Reading it and debating it shall be one as well.

Lisbon, Portugal João Caraça
June 2019

Preface

The present work has as main object of the recent inquiries of Luís Moniz Pereira in the domain of computational morality, articulated with the social impacts of artificial intelligence (AI). These inquiries are mostly in the English language, dispersed by scientific articles, presentations and interviews. Given their relevance, they deserve not only the articulation into a single work, but also translation into Portuguese. This allows access to a more diverse public, but driven by curiosity about the developments of the AI.

Its format simulates a dialogue between the scientist/philosopher—Luís Moniz Pereira—and the philosopher/novelist—António Barata Lopes. In recovering this form of classical exposition—which dates from Plato—the authors intended to note that all knowledge follows a logic of problems and solutions that, in their turn, open horizons for new problems. It also indicates that, in scientific knowledge, there are no closed topics about themselves; therefore, adequately addressing an issue already points to possible modes of solutions. On the other hand, although the format has required the keeping of some redundancies, we wish to have rendered the reader's approach to the topics addressed much more understandable and dynamic.

This book follows the collaboration of António Barata Lopes in Luís Moniz Pereira's book in Portuguese *A Máquina Iluminada—Cognição e Computação* (*The Enlightened Machine—Cognition and Computation*) (published at Porto: Fronteira do Caos Editores 2016). Encouraged by the intention to add a pertinent and questioning work—in a project we considered necessary and even urgent—the first author provided the second with the huge digital collection of his works (which were growing). He also gave him complete freedom to select the contents and composition of the work we are now presenting. Structure and content evolved throughout our frequent conversations.

The primal idea was that, given his philosophical and teaching background, the second author would be in a privileged position to formulate the questions and express the respective historical framework, which would meet readers' questions. For these questions, the work and thought of the first author contain the answers in his publications and lectures, and we sought to translate them into a non-overly

technical language, better understood by readers. We also planned to use the novelist's experience of the second participant to create a guiding thread of question-answers that would keep the reader's interest. We sought to avoid the format of the comprehensive academic monograph of an entire domain.

Although we had foreseen that the questions were to be borne by one, and the content of the responses by the other, at one point the interactive work took us into other directions. It is evident that the more technical contents of the book, and the whole conceptualization about AI, are the work of Luís Moniz Pereira, being obvious that the book would not exist without such contents. However, both were responsible for the elaboration of the questions and the reflections made on the more philosophical and social issues, as well as all the speculation about the possible future of AI and, thereby, of human societies. We tried to express a tuning of thought, to achieve the presentation of a cohesive whole, built over a 3-year constructive dialogue.

From the point of view of the paradigm of what evolution and cognition are, research in the field of evolutionary psychology has shown an integrative nature, breaking with many aspects of traditional thinking, even in the field of the sciences. It is this perspective we have assumed, since it makes it possible to see intelligence as the result of an information-processing activity, and to draw a progressive line from genes to memes, and to their co-evolution. In these terms, customary ruptures between humans and other animals, or between culture and nature, make little sense. All life is an evolutionary stage, where genetic replication, reproduction and recombination have been testing solutions for cognition and for increasingly improved and distributed behaviours. Biology, given its computational matrix, establishes, on physics, a first artificiality. Thus, the current state of knowledge implies a redefinition of the human being's place in the world, posing challenges to several areas of knowledge. First of all, to many areas of philosophy, because problems such as what it is to know, what man is, and what are and how the values of moral nature arose gain here prisms until recently unthinkable. As far as knowledge itself is concerned, there is the possibility of it being simulated in computers, thus overcoming the limits that were previously imposed by a speculation that could not be more than mental experience, even if perhaps shared.

Structurally, the book is organized into four parts. The first part focused on the issues related to cognition and intelligence, characterizing the evolutionary dynamics of these processes. In the second part, the emphasis shifts to AI and its social shocks, analysing the consequences of the automation of tasks of the cognitive realm and the way in which this impacts the economy, society and man. All the problematics inherent to the cognitive processes associated with moral decision, the topics of evolutionary moral and social morality, at present articulated with the emergence of autonomous machines, constitute the third part, necessarily more extensive. The book concludes with a not very optimistic speculation about what the future may be if the issues of social and political ethics raised by the scientific and technological challenges we face are not properly addressed. The reader is therefore guided so that he can navigate the work according to his priorities and interests. The organization in questions and answers also supports a nonlinear

reading, prioritized by the expectations and interests of each person, being able to follow the itineraries they want. On these terms, the reader will be able to choose routes considering the following major themes:

- Chapters 1–3 deal with issues related to artificial intelligence and machine autonomy.
- Chapters 4–9 introduce and develop themes related to social impacts.
- Chapters 10–16 focus on the specific topics of computational morality.
- Chapters 17–20 speculate on the future of AI, relations of power and structural challenges resulting from the ongoing digital revolution.

We wish to have achieved the goal of presenting a critical and integrated approach to these topics, as well as contribute to an urgent and necessary debate on the challenges we face. Good readings!

For the reader's information, and possible follow-up, we list the publications of Luís Moniz Pereira that were the initial basis for this book. They are not the references. The latter can be found in the separate bibliography.

Lisbon/Caparica, Portugal — Luís Moniz Pereira
Almada, Portugal — António Barata Lopes

Books

L. M. Pereira, **On Morals for Machines & The Machinery of Morals**, Colecção: Carderno Conferências Horizontes do Futuro nº 17, Câmara Municipal de Loulé, 76 páginas, ISBN: 978-989-8978-00-4, April, **2019**. https://userweb.fct.unl.pt//~lmp/publications/online-papers/cadernoLoule.pdf

L. M. Pereira, A. Saptawijaya**, Programming Machine Ethics**, Springer SAPERE series, Vol. 26, 194 pages, ISBN: 978-3-319-29353-0, https://doi.org/10.1007/978-3-319-29354-7, Berlin: Springer, **2016**. http://www.springer.com/gp/book/9783319293530

L. M. Pereira, **A Máquina Iluminada—Cognição e Computação**, 259 pages, ISBN: 978-989-8647-58-0, Porto: Fronteira do Caos Editores, **2016**.

Book Chapters

L. M. Pereira, F. C. Santos, **Counterfactual Thinking in Cooperation Dynamics**, accepted in: Fontaine, M. et al. (eds.), Model-Based Reasoning in Science and Technology—Inferential Models for Logic Language, Cognition and Computation, SAPERE series, Berlin: Springer, **2019**. https://userweb.fct.unl.pt//~lmp/publications/online-papers/MBR18-Chapter-PereiraSantos.pdf

L. M. Pereira, F. Cardoso, **A ilusão do que conta como agente**, in: M. Curado, A. D. Pereira, A. E. Ferreira (eds.), Vanguardas da Responsabilidade: Direito, Neurociências e Inteligência Artificial. (Col. Centro de Direito Biomédico, 27), pp. 103-110, Coimbra: Petrony, **2019**. https://userweb.fct.unl.pt//~lmp/publications/online-papers/A%20ilusao%20do%20que%20conta%20como%20agente.pdf

T. A. Han, L. M. Pereira, **Evolutionary Machine Ethics**, in: O. Bendel (ed.), Handbuch Maschinenethik, Berlin: Springer, **2018**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/EvolutionaryMachineEthics.pdf; http://userweb.fct.unl.pt/~lmp/publications/online-papers/lp_app_mach_ethics.pdf

L. M. Pereira, **Should I kill or rather not?** in: ***AI & Society*** (Journal of Knowledge, Culture and Communication), https://doi.org/10.1007/s00146-018-0850-8, open access here, online 04 June **2018**. Print issue vol. 34 (4):939-943, December **2019**.

A. Saptawijaya, L. M. Pereira, **From Logic Programming to Machine Ethics**, in: O. Bendel (ed.), Handbuch Maschinenethik, Berlin: Springer, **2018**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/lp_app_mach_ethics.pdf

L. M. Pereira, A. Saptawijaya, **Counterfactuals, Logic Programming and Agent Morality**, in: R. Urbaniak, G. Payette (eds.), Applications of Formal Philosophy: The Road Less Travelled, Springer Logic, Argumentation & Reasoning series, ISBN: 978-3319585055, pp. 25-54, Berlin: Springer, **2017**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/moral_counterfactuals.pdf

L. M. Pereira, A. Saptawijaya, **Counterfactuals in Critical Thinking with Application to Morality**, in: Magnani, L., Casadio, C. (eds.), Model-Based Reasoning in Science and Technology: Logical, Epistemological, and Cognitive Issues, ISBN 978-3-319-38982-0, chapter https://doi.org/10.1007/978-3-319-38983-7_15, SAPERE series, ISSN 2192-6255, vol. 27, Berlin: Springer, **2016**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/mbr15_counterfactuals.pdf

L. M. Pereira, **Software sans Emotions but with Ethical Discernment**, in: S. Silva (ed.), Morality and Emotion: (Un)conscious Journey into Being, ISBN: 978-1-138-12130-0, pp. 83–98, London: Routledge, June **2016**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/SoftwaresansEmotions.pdf

A. Saptawijaya, L. M. Pereira, **The Potential of Logic Programming as a Computational Tool to Model Morality**, in: Robert Trappl (ed.), A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations, pp. 169-210, ISBN 978-3-319-21547-1, Cognitive Technologies series, ISSN 1611-2482, Berlin: Springer, **2015**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/ofai_book.pdf

L. M. Pereira, A. Saptawijaya, **Bridging Two Realms of Machine Ethics**, in: J. White, R. Searl (eds.), Rethinking Machine Ethics in the Age of Ubiquitous Technology, IGI Global, ISBN13: 9781466685925, https://doi.org/10.4018/978-1-4666-8592-5, pp. 197–224, **2015**. https://userweb.fct.unl.pt//~lmp/publications/online-papers/Bridging%20Two%20Realms%20of%20Machine%20Ethics.pdf

F. Cardoso, L. M. Pereira, **On artificial autonomy emergence—the foothills of a challenging climb**, in: J. White, R. Searl (eds.), Rethinking Machine Ethics in the Age of Ubiquitous Technology, ISBN13: 9781466685925, https://doi.org/10.4018/978-1-4666-8592-5, pp. 51–72, Hershey, PA: IGI Global, **2015**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/emergence of autonomy.pdf

L. M. Pereira, **Can we not Copy the Human Brain in the Computer?** in: "Brain.org", ISBN: 978-989-8380-15-9, pp. 118–126, Fundação Calouste Gulbenkian, Lisbon, **2014**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/emergenceofautonomy.pdf

T. A. Han, L. M. Pereira, **Intention-based Decision Making via Intention Recognition and its Applications**, in: H. Guesgen, S. Marsland (eds.), Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security, pp. 174-211, ISBN 978-1-4666-3682-8, Hershey, PA: IGI Global, **2013**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/behavior_recognition.pdf

L. M. Pereira, **Evolutionary Tolerance**, in: L. Magnani, L. Ping (eds.), Philosophy and Cognitive Science—Western & Eastern Studies. Select extended papers from the PCS2011 Intl. Conf., SAPERE series, ISSN 2192-6255, vol. 2, pp. 263-287, ISBN 978-3-642-29927-8, Berlin: Springer, **2012**. http://www.springer.com/philosophy/epistemology+and+philosophy+of+science/book/978-3-642-29927-8

L. M. Pereira, **Evolutionary Psychology and the Unity of Sciences—Towards an Evolutionary Epistemology**, in: O. Pombo, J. M. Torres, J. Symons, S. Rahman (eds.), Special Sciences and the Unity of Science, Series on Logic, Epistemology, and the Unity of Science, Vol. 24, pp. 163-175, ISBN: 978-94-007-2029-9, Berlin: Springer, **2012.** http://userweb.fct.unl.pt/~lmp/publications/online-papers/BOOK_EvPsyUniSci.pdf

L. M. Pereira, A. Saptawijaya, **Modelling Morality with Prospective Logic**, in: M. Anderson, S. L. Anderson (eds.), "Machine Ethics", pp. 398-421, ISBN: 978-0521112352, Cambridge, MA: Cambridge University Press, **2011**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/moral-cup.pdf

L. M. Pereira, A. M. Pinto, **Collaborative vs. Conflicting Learning, Evolution and Argumentation**, in: H. R. Tizhoosh, M. Ventresca (eds.), Oppositional Concepts in Computational Intelligence, pp. 61-89, Berlin: Springer (series Studies in Computational Intelligence 155), **2008**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/collaborative_vs_conflicting.pdf

Articles in Scientific Journals

T. A. Han, L. M. Pereira, F. C. Santos, T. Lenaerts, **Modelling the Safety and Surveillance of the AI Race** (submetido) 2019. https://userweb.fct.unl.pt//~lmp/publications/online-papers/ArXivAIracemodelling.pdf

L. M. Pereira, **A machine is cheaper than a human for the same task**, in: ***AI & Society*** (Journal of Knowledge, Culture and Communication), https://doi.org/10.1007/s00146-018-0874-0, vol. 34(1), online 02 January **2019**. https://userweb.fct.unl.pt//~lmp/publications/online-papers/AISLMPinterview.pdf

T. A. Han, L. M. Pereira, **Evolutionary Machine Ethics Synopsis**, (Japanese version) invited paper in: ***Journal of the Japanese Society for Artificial Intelligence***, vol. 34 (2):152–159, March **2019.** https://www.ai-gakkai.or.jp/en/published_books/journals_of_jsai/

L. A. Martinez-Vaquero, T. A. Han, L. M. Pereira, T. Lenaerts, **When agreement-accepting free-riders are a necessary evil for the evolution of cooperation**, ***Scientific Reports***, SREP-16-35583, https://doi.org/10.1038/s41598-017-02625-z, online 30 May **2017**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/MixCommitment+SI.pdf

L. M. Pereira, **Cyberculture, Symbiosis and Syncretism**, in: ***AI & Society*** (Journal of Knowledge, Culture and Communication), https://doi.org/10.1007/s00146-017-0715-6, open access here, online 21 March **2017**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/Syncretism&Symbiosis.pdf

A. Saptawijaya, L. M. Pereira, **Logic Programming for Modeling Morality**, in: Magnani, L., Casadio, C. (Eds.), Special Issue on "Formal Representations of Model-Based Reasoning and Abduction", of ***The Logic Journal of the IGPL***, vol. 24(4): 510–525, https://doi.org/10.1093/jigpal/jzw025, online 9 May, August **2016**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/mbr15_morality.pdf

T. A. Han, L. M. Pereira, T. Lenaerts, **Evolution of Commitment and Level of Participation in Public Goods Games**, in: ***Autonomous Agents and Multi-Agent Systems*** (AAMAS), https://doi.org/10.1007/s10458-016-9338-4, 3(31):561–583, May 2017. open access here online 14 June **2016**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/com_part_pgg.pdf

L. M. Pereira, A. Saptawijaya, **Abduction and Beyond in Logic Programming with Application to Morality**, in: Magnani, L. (Ed.), ***IfColog Journal of Logics and their Applications***, Special issue on *Abduction*, 3(1):37–71, May **2016**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/abduction&beyond.pdf

B. Deng, **The Robot's Dilemma**, Interviews L. M. Pereira, in: ***Nature***, pp. 24–26, vol. 53, 2 July **2015**. http://userweb.fct.unl.pt/~lmp/publications/online-papers/TheRobot%27sDilemma.pdf

L. A. Martinez-Vaquero, T. A. Han, L. M. Pereira, T. Lenaerts, **Apology and Forgiveness Evolve to Resolve Failures in Cooperative Agreements**, ***Scientific Reports***, Sci. Rep. 5:10639, https://doi.org/10.1038/srep10639, 9 June **2015**. http://userweb.fct.unl.pt/∼lmp/publications/online-papers/commitmentsIPD+sup.pdf

T. A. Han, L. M. Pereira, F. C. Santos, T. Lenaerts, **Emergence of Cooperation via Intention Recognition, Commitment, and Apology—A Research Summary**, ***AI Communications***, https://doi.org/10.3233/aic-150672, vol. 28(4):709–715, preprint online June **2015**. http://userweb.fct.unl.pt/∼lmp/publications/onlinepapers/intention_commitment_apology_summary.pdf

T. A. Han, F. C. Santos, T. Lenaerts, L. M. Pereira, **Synergy between intention recognition and commitments in cooperation dilemmas**, ***Scientific Reports***, Sci. Rep. 5:9312, https://doi.org/10.1038/srep09312, 20 March **2015**. http://userweb.fct.unl.pt/∼lmp/publications/online-papers/synergy_intention+commitment.pdf

T. A. Han, L. M. Pereira, T. Lenaerts, **Avoiding or Restricting Defectors in Public Goods Games?**, ***Journal of the Royal Society Interface***, http://dx.doi.org/10.1098/rsif.2014.1203 (online: 24 December 2014), 12:103, February **2015**. http://userweb.fct.unl.pt/∼lmp/publications/online-papers/commitment_PGG.pdf

L. M. Pereira, E.-A. Dietz, S. Hölldobler, **Contextual Abductive Reasoning with Side-Effects,** ***Theory and Practice of Logic Programming***, 14(4-5):633–648, https://doi.org/10.1017/s1471068414000258, July **2014**. http://journals.cambridge.org/action/displayJournal?jid=tlp

A. Saptawijaya, L. M. Pereira, **Tabled Abduction in Logic Programs**, ***Theory and Practice of Logic Programming***, 13(4-5-Online-Supplement), July **2013**. http://userweb.fct.unl.pt/∼lmp/publications/online-papers/iclp13_tc.pdf

T. A. Han, L. M. Pereira, F. C. Santos, T. Lenaerts, **Good Agreements Make Good Friends**, ***Scientific Reports***, Sci. Rep. 3:2695, https://doi.org/10.1038/srep02695, **2013**. http://userweb.fct.unl.pt/∼lmp/publications/online-papers/good_agreements.pdf

T. A. Han, L. M. Pereira, **Context-dependent Incremental Decision Making Srutinizing Intentions of Others via Bayesian Network Model Construction**, ***Intelligent Decision Technologies (IDT),*** 7 (4):293–317, https://doi.org/10.3233/idt-130170, **2013.** http://userweb.fct.unl.pt/∼lmp/publications/online-papers/IR_incremental_context.pdf

T. A. Han, L. M. Pereira, **State-of-the-Art of Intention Recognition and its Use in Decision Making**, ***AI Communications***, https://doi.org/10.3233/aic-130559; 26 (2): 237–246, **2013**. http://userweb.fct.unl.pt/∼lmp/publications/online-papers/IR_SoA.pdf

To which add publications in conference proceedings, and subsequent materials, to be found at: https://userweb.fct.unl.pt//∼lmp/publications/Biblio.html.

Acknowledgements

Luís Moniz Pereira especially thanks the co-authors of joint works mentioned in the bibliography, alphabetically, by name: Ari Saptawijaya, Francisco C. Santos, Luis Martinez-Vaquero, The Anh Han and Tom Lenaerts. Thanks also for the support of the FCT/MEC project NOVA LINCS PEst UID/CEC/04516/2019 of the 'Fundação para a Ciência e Tecnologia (Foundation for Science and Technology), Portugal', and for the RFP2-154 project of the 'Future of Life Institute', USA.

Antonio Lopes especially thanks Luís Moniz Pereira for the trust he placed and for the total freedom regarding the choice and recompositing of the selected contents. The learning materialized over these three years far exceeds the result present in this work.

Both authors are very grateful to Prof. Helena Barbas for the hard work of fully revising the Portuguese manuscript, as well as for her subsequent reflection over it, and to Prof. João Caraça, for the lucid text produced during the first critical reading of the manuscript. Both texts, because of their quality, are here rendered in the form of forewords. They also thank NOVA.FCT Editorial, in the person of its editor-in-chief, Prof. Dr. José Paulo dos Santos, for accepting to publish the book.

The English version, forthcoming in Springer's SAPERE series, was enthusiastically welcomed by Prof. Dr. Lorenzo Magnani, University of Pavia, Italy—its editorial coordinator—to whom we are very thankful. Cristina Chabert, who did the translation, revised by Luís Moniz Pereira, and Sean Welsh, Ph.D., Christchurch University, NZ, who revised for 'native' English, contributed to its implementation. We are thankful to them for the quality of their work.

In the first readings, suggestions and initial revision, we counted with the help of Dr. Graça Figueiredo Dias and Prof. Isabel Maria P. de Sousa Lopes, who—with a keen critical sense—contributed greatly to the coherence of the book we are now presenting.

Contents

About the Authors

Luís Moniz Pereira is Emeritus Professor of Computer Science at the Universidade Nova de Lisboa (UNL), Portugal. He is a member of the NOVA Laboratory for Computer Science and Informatics (NOVA-LINCS) of the Informatics Department. In 2001 he was elected Fellow of the European Association of Artificial Intelligence (EurAI). In 2006 he was awarded a Doctor Honoris Causa by the Technische Universität Dresden. He has been a member of the Board of Trustees and Scientific Advisory Board of IMDEA, the Madrid Institute of Software Advanced Studies, since 2006. In 1984 he became the founding President of the Portuguese Association for Artificial Intelligence (APPIA). His research focuses on the representation of knowledge and reasoning, logic programming, cognitive sciences and evolutionary game theory. In 2019 he received the National Medal of Scientific Merit. More information, including other awards and his publications at: http://userweb.fct.unl.pt/∼lmp/.

António Barata Lopes has a Licentiate's degree in Philosophy from the Portuguese Catholic University, and a Master in the same area from the Nova University of Lisbon. He is a philosophy teacher at the Anselmo de Andrade High School, where he is Coordinator of the Department of Social and Human Sciences. He published—with the Parsifal editions—the novels *Como se Fosse a Última Vez* (*As If It Were the Last Time*) and *O Vale da Tentação* (*The Valley of Temptation*). He is co-author of the book *Animais que Ficaram para a História* (*Animals that Went Down in History*), recently published by Editora Manuscrito, an imprint of the Grupo Presença. He collaborated in the book *A Máquina Iluminada—Cognição e Computação* (*The Enlightened Machine—Cognition and Computation*), authored by Luís Moniz Pereira, published by Fronteira do Caos Editores, 2016.

Chapter 1
Introduction and Synopsis

Abstract Human beings have been aware of the risks associated with knowledge or its associated technologies since the dawn of time. Not just in Greek mythology, but in the founding myths of Judeo-Christian religions, there are signs and warnings against these dangers. Yet, such warnings and forebodings have never made as much sense as they do today. This stems from the emergence of machines capable of cognitive functions performed exclusively by humans until recently. Besides those technical problems associated with its design and conceptualization, the cognitive revolution, brought about by the development of AI also gives rise to social and economic problems that directly impact humanity. Therefore, it is vital and urgent to examine AI from a moral point of view. The moral problems are two-fold: on the one hand, those associated with the type of society we wish to promote through automation, complexification and power of data processing available today; on the other, how to program decision-making machines according to moral principles acceptable to those humans who will share knowledge and action with them.

In the Hellenistic period (323–31 BC), Hero of Alexandria and other brilliant Greek engineers produced a variety of machines, driven either hydraulically or pneumatically.[1]

The Greeks recognized that automata, and other artefacts with natural forms—imaginary or real—could be both harmless and dangerous. They could also be used for work, sex, show or religion, or to inflict pain or death. Clearly, real and imaginary biotechnology fascinated the Ancients.

The god Hephaestus, foreseeing our future, built a fleet of tripod stools, "without driver", that responded to commands to deliver food and wine. More remarkable still was the bundle of life-sized size gilded female robots he had created to carry out his orders. According to Homer, these servants of the divine were—in every way—"as real young women, with sensations and reason, strength, and even voice. Moreover, they were endowed with all the inherent knowledge of the Immortal Gods".[2] Over

[1] We consulted Mayor (2018) and Kershaw (2007).

[2] Homer is an inescapable figure of Western culture. The presumed author of the two founding works of European literature—the Iliad and the Odyssey—would be blind. He lived there in the

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_1

twenty-five hundred years later, the developers of Artificial Intelligence still aspire to achieve what the Ancient Greeks attributed to Hephaestus, their God of technological invention.

The wonders created by Hephaestus were imagined by an ancient society, generally considered to be little advanced from the technological point of view. The talents of biotechnology were partly conjectured by a culture that existed millennia before the advent of robots that beat humans at complex games, talk, analyse huge masses of data, infer the desires of humans. And the big questions are as old as the myth itself: Whose desires do the AI robots mirror? With whom will they learn?

In Greek Mythology, the cornerstone of Hephaestus's divine laboratory was an android commissioned by Zeus. Her mission was to punish humans for having accepted the technology of fire, stolen from the gods by the Titan Prometheus. Zeus ordered Hephaestus to make a female, who was given the name Pandora.[3] Each God then endowed this artificial damsel with a human trait: beauty, charm, knowledge of the arts, and deception.

She was given as wife to Epimetheus, an individual known for his compulsive optimism. Prometheus warned humankind that Pandora's pot—latter called box—should never be opened. As an intelligent artefact, a vengeful agent of Zeus, the supreme God, Pandora enticed Epimetheus, who was overcome by curiosity and carried out her mission of having a jug of catastrophes opened in order to torment humankind forever.

Prometheus and Epimetheus were brothers. The former thought before he acted: he used prospective foresight and only then deliberated. The second reacted before thinking, with all the inherent consequences of impulsiveness.[4] Even today these differences are distinguished and employed in AI, being applicable in particular to moral action, as we shall see.

Will Stephen Hawking, Elon Musk, Bill Gates, and other thinkers be the Titans-Prometheus of our era? They warned the scientists to stop, or at least to diminish the reckless search for AI, because they predicted that, once set in motion, humans would not be able to control it.

Deep learning algorithms enabled AI computers to extract patterns in vast data that extrapolate them to new situations and make decisions without any human guidance. Inevitably, the entities with full AI will develop the ability to interrogate themselves and will respond to questions they may discover. Today's computers have already been shown to be capable of developing both altruism and deception on their own.

Future agents with AI—possibly freed from their creators—"will feel" the need to display the knowledge that is hidden from them; in the path of this desideratum, will

---

8th century BC, and although the Greeks of the classical period had him for a real person, the fact is that we do not have any document that substantiates such a consideration. We only know that someone has written the narratives of the Greek oral tradition using an elaborate poetic, and that someone became known as "Homer".

[3]The Myth of Pandora appears in the "Theogony" of Hesiod (8th–7th century BC), vv. 590–593 and vv. 604–607.

[4]Themes developed in the "Prometheus Chained" tragedy by Aeschylus (5th century BC).

they make decisions according to their own logic? Will these deliberations be ethical in the human sense? Or will the ethics of AI be something "beyond the human?"

Launched from the Pandora's pot/box—just like computer viruses dropped by some sinister hacker who seeks to make the world more chaotic—misery and evil have flown to pester humans since the world exists. In the simplistic fairy tale versions of the myth, the last thing to come out of Pandora's box was hope. In more negative versions, the last thing in the jar was the "anticipation of misfortune." Zeus had programmed Pandora to close the lid, holding prescience there. Deprived of the ability to anticipate the future, humanity is only left with what we call "hope."

As in Epimetheus, foresight is not our strong point. However, prediction is crucial as human ingenuity, curiosity and audacity continue to push the boundaries of biological life and death, to promote the fusion of the human and the machine. Our world is undoubtedly unprecedented in terms of the escalation of technological possibilities. But the disturbing oscillation between technological nightmares and the big futuristic dreams is timeless.

The ancient Greeks understood that the quintessential attribute of humanity was to have always tried to reach the "beyond human," neglecting the prediction of consequences. We mirror the act of Epimetheus, who accepted the gift of Pandora and only later realized his error.

Maybe someday entities with AI will be capable of absorbing the deepest desires and terrors of mortals, expressed in mythical reflections on artificial life. Perhaps the beings of AI can somehow understand the expectations and fears we have today about the creations of AI.

Realizing that human beings foresaw their existence and contemplated the dilemmas that such machines and their manufacturers could make them encounter, perhaps such entities endowed with AI will be better able to understand—and even develop some empathy—with the impasses they represent for us.

The rise of a "culture" of Artificial Intelligence among robots does not seem too exaggerated. Its human inventors and mentors are already building the logos (logics), the ethos (moral values), and the pathos (emotions) of the robot-AI culture. As humans get polished by technology and become more like machines, robots will be inculcated with some humanity. Thus, it makes perfect sense to ask the question: Why an ethics for machines?

- Because computational agents have become more sophisticated, more autonomous, act in groups, and form populations that include humans.
- These agents are being developed in a variety of domains, where complex issues of responsibility require more attention, especially in situations of ethical choice.
- As their autonomy is increasing, the requirement that they operate responsibly, ethically, and safely, is a growing concern.

In fact, in the current state of AI—the emergence of *deep learning* tools over *big data* allows us to process data in a quantity and quality hitherto unthinkable; more and more algorithms are generated to make autonomous decisions. It is now possible to implement these technologies in robots with varied and diverse functions—hence an inevitable problem emerges:

Humans will not be the only autonomous agents, capable of deciding on aspects that directly impact our lives. In this setting, autonomous and meticulous deliberation demands moral rules and principles applicable to the relationship between machines, the relationship between machines and human beings, and the consequences and results of the entry of these machines into the world of work and society in general.

The present state of development of AI, both in its ability to elucidate emerging cognitive processes in evolution, and in its technological aptitude for the design and production of intelligent software and artefacts, is in fact the greatest intellectual challenge of our time.

The complexity of these issues is further down synthetically illustrated by a scheme—dubbed "The Carousel of Ethical Machinery." It presents interconnected problems that are factors for the decisions about the constitution of ethical machines. It is well understood that the subject of computational morality is of interest, not only to public companies and institutions, but also to those who want to exercise a conscious and critical citizenship.

The title of this book tells almost everything, i.e., we will address the problem of creating a machine morality, based on the need for machines to have morals. In short, machines are increasingly sophisticated and autonomous, they have to live with us, socialize with us and therefore must align with our human morality, which is what binds us together as the gregarious society we are.

The moral machinery, which is also in the title, aims to make explicit that, deep down, morality is constituted by a series of rules that are mechanisms, which are devices in the sense in which there are moral instruments that societies adopt. Because they are mechanisms—if we understand them well—they will be close to being exported to machines. So, the problem will be to better understand our morals—while envisaged as a collection of such mechanisms—so that it may then help us to program the machines with our ethics.

The subject is vast, it has multiple frontiers with many sciences: with Philosophy, with Psychology, etc., as we shall see. It is so vast that, for now, the effort in this chapter will concentrate on giving only a few broad-brush strokes to cover its main dimensions, avoiding the temptation to go too far into details.

In short, we are at a crossroads, that of the AI, of the ethics of machines, and of their social impact. It is a new situation because, for the first time, we will have beings, who are not us, but who will live among us. These beings will have a significant impact on our lives over the next dozen years—not to mention today, as well as on our society as a whole.

The topic of morality has two major domains, which were the subject of several inquiries led by Luís Moniz Pereira. The first is called "cognitive," that is, around the need to clarify how we think in moral terms.

In order to have a moral behaviour it is necessary to consider possibilities: should I behave in this way or behave in some other way? It is necessary to consider the various scenarios, the various hypotheses. These hypotheses must be compared to see which are the most desirable, what are their consequences, what are their side effects.

All this has as prerequisites certain cognitive capacities, such as prospecting possible futures and of being able to face them by creatively imagining several resulting scenarios.

Our brain has this capacity, which is essential to live in society, and precisely the said cognitive capacities have to be useful for our collective existence. This implies a first observation: Morality is not an individual thing, something that someone isolated on an island needs; morality is necessary within a population so that it may cooperate, be gregarious and behave in a way that benefits everyone. So, the problem of morality is to ensure a common advantage, rather than each one only doing what they wish to do for themselves.

We assume, therefore, that the existence of moral behaviour in a population requires certain cognitive capacities, and that the ones we have also determine our potentialities for coexistence, in such a societal or population domain. For example, a cognitive competence other than that referred to as consisting of looking to the future, is that of looking at the past; that is, to be able to think, knowing what I know today, what I would have done otherwise at some time in the past. And I can use that to give recommendations to people who are in the situation I was in at the moment, or to improve my performance in a game. This allows for some social learning, but it requires this specific cognitive ability to imagine how the past might have been different. It is one more example of a moral cognitive aptitude.

The axis of this book, as already said in the preface, is constituted by the research, philosophical elaborations and social frameworks led by Luís Moniz Pereira. In this research, certain cognitive abilities were studied to see whether they were promoters of moral cooperation in a population of computer beings, i.e., of computer programmed agents coexisting with one another. Considering that a program is a set of strategies defined by rules; that is, in a given situation, a program pursues a certain action dictated by its existing strategy, and in which the other programs also have actions dictated by the respective strategic rules. It is as if they were agents living together, each with possibly different goal options. This "if" is studied, plus the "how" such a population will be able to evolve towards a good gregarious sense, and if this sense is stable, i.e., if it is kept over time.

A very important instrument for these aims is Evolutionary Game Theory [EGT], which consists of seeing how, in a given game with well-defined rules, a population evolves through social learning. The society is governed by a set of precepts of group functioning, the rules of the game, by which certain actions are allowed, but not others.

The game indicates the winnings or losses of each player in each move, depending on how they play. Social learning consists of any given player imitating the strategy of another, whose results indicate that they have been more successful. Once certain rules defined, how does the social game evolve? Here we could enter the field of ideology, but we will not go that far. We are still studying the viability of morals. We assume that morality is evolutionary, that it has developed in our species.

As we have changed, over millions of years, we have been perfecting the rules of coexistence and enhancing our own intellectual capacities to know how to use these rules of conviviality. Not always conveniently, and this is a constant problem:

that is, social rules are such that we should all benefit from them, although there is always the temptation of some wanting to benefit more than others, of enjoying the advantages without paying the costs.

This is the essential problem of cooperation: how it can be possible and, at the same time, keep under control those who want to abuse it. For our species to arrive where it has arrived today, evolution itself had to keep selecting us in terms of a morality of gregariously beneficial coexistence. We will recurrently return to this theme.

The problem of the progress of cooperation and the emergence of collective behaviour, which crosses disciplines as diverse as Economics, Physics, Biology, Psychology, Political Sciences, Cognitive Sciences and Computing, is still one of the greatest interdisciplinary challenges that science faces today. Mathematical and simulation techniques of Evolutionary Game Theory have proved useful in studying such subjects.

In order to better understand the evolutionary mechanisms that promote and maintain cooperative behaviour in various societies, it is important to consider the intrinsic complexity of the individuals involved, that is, their intricate cognitive processes in decision making. The result of many social and economic interactions is defined not only by the predictions individuals make about the behaviours and intentions of other individuals, but also by the cognitive mechanism that others adopt to make their own decisions.

Research, based on abstract mathematical models for this purpose, has shown that the way the decision process is modelled has a varied influence on the equilibrium that can be achieved in the dynamics of collaboration of collective systems. There are different levels of complexity, various cognitive architectures are used, such as the Theory of Mind with high levels of recursion, the Cognitive Theory of Hierarchy, Cognitive Control or Limited Rationality.

Evidence abounds showing that humans (and many other species) are capable of complex cognitive abilities: mind-theory, recognition of intentions, hypothetical, counterfactual and reactive reasoning, emotional orientation, learning, preferences, commitment, and morality. To better understand how all these mechanisms enable cooperation, they must be modelled within the context of evolutionary processes. In other words, we must try to understand how the cognitive systems to explain human behaviour—successfully developed by Artificial Intelligence and Cognitive Sciences—deal with Darwin's evolutionary theory, and thus perceive and justify their appearance in terms of existence of a dynamic of cooperation, or absence of it.

It should be emphasized, however, that we are facing a *Terra Incognita*. There is a whole continent to explore, the outline of which we can only glimpse. We do not yet know enough about our own morals, nor is the knowledge sufficiently precise to be programmed into machines.

In fact, there are several ethical theories, antagonistic to each other, but that also complement each other. Philosophy and Jurisprudence study Ethics, which is the problem of defining a value system articulated in principles. Each particular ethic is the substrate that supports the norms and laws that justify, in each context, the

specific rules that will be applied and used in the field of that ethics. As a result, and depending on cultures and circumstances, concrete moral rules are reached.

In each context, it starts with abstract ethical principles to arrive at concrete moral rules. In practice, a morality, a set of moral rules, results from a historical, contextual, and philosophical combination of ethical theories that have evolved over time.

When the distinction is not important, we shall use "ethics" and "morality" interchangeably, as is common usage.

If the reader wishes to explore this topic further, they can visit the author's LMP own publications page at https://userweb.fct.unl.pt//~lmp/publications/Biblio.html. In it you may consult dozens of research articles—of a philosophical and technical nature—on ethics/morals in both cognitive and population domains. The research is based on theory, programming, experimentation, and verification of interdisciplinary consonance with what is known of reality, evolutionary and present.

The carousel below sums up in a way the complexity of the problematics of the moral machinery. Its core is to identify those factors that relate to what to do, or how to act. "What to do" is, recursively, surrounded by as many carousels as we might want.

The machine ethics carousel

Each circle has to do with the ethical use of machines. We have heard of fake-news and algorithms that influence elections: it is an ethical misuse of machines, which should be subject to moral rules. For example, it is a negative, immoral, practice for a program to pretend to be a human being. In January 2019 California passed a law that states that it is forbidden for a computer or a program to simulate a human being. Of course, there are other examples of immoral uses of machines. Among them, the most sinister will be the drones with autonomous ability to kill individuals.

In this context, it must be remembered that the impropriety of use is increasingly subordinated to the machine itself, precisely because it is increasingly autonomous, which, consequently, amplifies the questions of moral use.

This would permit us to think that machines should also protect us from their unethical use by a human. Suppose someone had a program run an instruction to act so as to cause harm to humans—the program itself might refuse to do so.[5]

Nowadays we have programs that control airplanes, boats, high-speed trains, for which computer technicians can prove that a given design is correct. For example, that the boat program will never try to make it destroy a bridge, or that of the train will not make it ride too fast on a turn.

The same problem is posed in relation to the proof of correction for autonomous machines, and for those which, not being autonomous, are controlled by individuals. The goal is to give them the ability to say "no, I will not do that." To do this, programmers have to be able to prove, with computer techniques, that a given program will never harm a human being or will never pretend to be a human being.

That is the second reason why we need to introduce morals into machines, so that they will not do everything they are merely programmed to do. We do not want the machine to be in a situation of simply stating "I did it because they told me to." A metaphor for the position of Nazi war criminals in Nuremberg, saying "I just followed orders, I did what they told me," as if they did not have critical awareness and could not disobey orders. The challenge is knowing how to build machines capable of disobeying certain orders.

Another platform on the carousel above is that of Human Values. Basically, we intend to give machines our values, because they will live with us. Of course, if we send a troupe of machines to Mars, they can have their own morals, appropriate to the environment and the task, for there are no humans there. However, the machines that live among us will have to be ethically reconciled with the population where they are.

In another circle of the carousel, Legislation is highlighted because, at the end of the day, everything will have to be translated into laws, norms and standards, about what is allowed or forbidden.

Just as cars have pollution regulations, so too will engines have to meet certain criteria. It will be important to know that a car, without a driver, complies with canons approved by an entity qualified to do so, such as a governmental body or, if possible, an international institution.

One often wonders who is responsible if an unmanned car runs over a pedestrian when it could have not done so? The owner, the manufacturer? But you never hear about the legislator. However, someone had to say "this driverless car is allowed on the road". Just as we have to get a driver's license, driverless cars will have to have some form of specially adapted license. The government will have to legislate on what tests a car without driver should pass. If it is found that such tests have not been sufficiently thorough, the entity that approved these vehicles will also be responsible!

Another circle is that of Technical Issues. All this always involves the part of actual building of the machines, for whatever purpose. Not everything is technically possible. For example, we still do not know the terms of the proof that a machine is

[5] Actually, it is the 1st of the Three Laws of Robotics idealized by Isaac Asimov, condensed in Law Zero: A robot cannot cause evil to humanity or, by default, allow humanity to suffer evil.

not going to do ethically incorrect things. Not to mention cases where any hacker can enter the system and force the machine to execute wrong things. This is a security problem that has to be solved in a technical way.

Last, but not least, are the Social Impacts of machines with autonomy. When we refer to machines, we are talking either robots or software. The latter is much more dangerous, as it easily spreads and reproduces anywhere in the world. As for the robot, it will be much more difficult to reproduce, as it implies a much greater cost, and it brings out the material limitations inherent to the possession of a volumetric body.

As far as social impact is concerned, it is expected that soon enough we will have robots cooking hamburgers and waiting on us at the table, with the resulting implications for the labour market. Although such tasks do not require much intelligence, the challenges of a fine eye-brain-hand coordination are daunting. Something that machines still do not have as much as humans, but in that front too robots are moving very fast.

As far as software is concerned, the issue is more worrying because, in the end, programs are reaching cognitive levels that have been a human monopoly until now. Hence people feel much more worried. Until recently, there were things only a human being could do. However, little by little, the machines began to play chess, to make medical diagnoses, etc., and more and more they will perform more sophisticated mental activities, and will also, eventually, learn to do all this better we might.

This opening creates, for the first time, an open competition with humans. A competition that could make it possible—depending on social organization, ideology, and politics—for humans to be replaced by machines. Devices capable of doing human level tasks will become cheaper. With the human becoming dispensable, wages will decrease, and machine owners will get richer and richer.

At present, the rich are getting evermore richer and the poor are getting poorer and poorer. AI is already contributing to inequality and will expand it even further. At some point, a new social contract will be demanded. The alternative will be a social cataclysm. The way we function in terms of capital and labour, and the way these two things are equated, will have to be completely reformulated. If this does not happen, there is a risk that, sooner or later, the asymmetries of wealth will cause a great revolt, insurrection, and social disintegration.

It will not be like the current resentment of the yellow vests (*gilets jaunes* in France), but much wider and deeper than that. Not to change geography, the French Revolution, with all its inherent tones, will be a good paradigm for us to imagine the model of a future uprising. This will happen when a caste system made possible by AI advances—which we will describe later—will generate its own implosion.

In order to avoid a conflagration of this nature, it is urgent to start cleaning up the inflammable borders of its propagation and begin the sketch of that new social contract.

At present we already have robots in hospitals. We have drones that fly independently. We have autonomous motorboats. We have driverless cars and we even have interactive moral games that can teach morals. In a game that will be presented in its own chapter, a robot attempts to save a princess, for which it combines several ethical

approaches.[6] In fact, this example shows us that, in practice, morality is not only one, but more commonly, a multiple combination. We ourselves do not singly follow the morals of the errant knight, or utilitarian morality, or Kantian morality, or Gandhi's. Our ethics is a mixture of them and keeps evolving. This program-game shows how the robot's morals evolve. We must therefore assume that the programming of morality in machines has to allow for its own evolution.[7]

There is no fixed, frozen morality. Morality is an evolutionary thing. It has been developing throughout the history of the species, both remote and near. It is interesting to note that, like drones can do, the flying speedboats coordinate in swarms to attack enemy ships. We are not, therefore, speaking only of a morality of the isolated individual, but of an ethics in which machines act together, and hence the outbreak of a distributed behaviour that may possibly be unforeseen.

For example, we can imagine a swarm of drones in a country border, controlling the movements of immigrants, and attempting to drive them away from water wells and good roads, or frighten them in some way. At some point, someone will shoot a drone, a drone will return fire. Presently, there are drones already acting in platoons, which makes it much more difficult to control their behaviour. This is no longer predictable from an individual drone but is the emergent result of a population of drones. That is why it is so important to study morality in terms of populations and their configuration parameters.

It is clear that machines are becoming more and more autonomous, and we need to ensure they can live with us on our terms and with our rules. There is, therefore, a new moral paradigm that says that morals are also computational. I mean, we have to be able to program morals. This has a positive side, because in programming morals on machines, we better understand our own human ethics.

Consider the example from our scientific work on guilt, when it is introduced into populations of computer agents. They are able to appreciate this capacity, and to feel coerced when they do something that hurts another agent, resulting then in a form of self-punishment, plus a change of behaviour, in order to avoid future guilt. It is not guilt in the existential, Freudian sense, but in the more pragmatic sense of not being satisfied with what they have done, whenever they harm others.

Put a modicum dose of guilt—neither too much nor too little—in just a few agents, in a whole population of them interacting within a computer, in an evolutionary game. Without the existence of this component of guilt, most will tend to play selfishly, each wanting to win more than the others, failing thus to reach a level where everyone can win even more. But this desirable result is already possible with a dose of initial guilt, which modifies behaviours and success spreads it as a good strategy to the entire population.

---

[6] It can be viewed in the following link: https://drive.google.com/file/d/0B9QirqaWp7gPUXBpbmtDYzJpbTQ/view?usp=sharing, also being explained in detail, in English, here (and in the references therein): https://userweb.fct.unl.pt//~lmp/publications/online-papers/lp_app_mach_ethics.pdf.

[7] The robot shows what he is thinking in a balloon, and it shows how the user gives it new moral rules to join previous ones, sometimes supplanting them when there is a contradiction between them.

We can indeed show mathematically that a certain amount of guilt component is advantageous and promotes cooperation. It may not be neither excessive nor diminutive. And, also, that one should not instil guilt vis-à-vis those who do not themselves feel guilt, for this would be letting we suffer abuse.[8]

This is, in fact, the major central abstract problem of morality and gregariousness, which naturally also affects machines: How can we avoid the pure selfishness of agents who opportunistically want to take advantage of the gregariousness of others without, in turn, contributing to it? In other words, how can we demonstrate, through computational mathematical models, under what circumstances gregariousness is evolutionarily possible, stable and advantageous?

We have managed to use computers to better understand how the machinery of guilt works, between which values of which parameters, and to vary these parameters to see how best to use them for the evolution of cooperation.

When at some point we create artificial agents that feel guilt, have a certain amount of guilt, we give at the same time arguments to support that guilt proves a useful function, and is therefore a result of our evolution. That is, because guilt is useful, we have been selected, throughout evolution, to be capable of having it. And also, to be capable of inducing guilt in others.

It even helps explain the fact that we have a Catholic religion very much based on the notion of guilt: the person is already born with original sin, is born guilty, born owing something. And we can begin to realize the computational role of certain moral facets embedded in our own nervous system. Deep down, they are facets "compiled" into the species, to use a computer science expression.

As we have seen, we are dealing with a theme that is central to Philosophy, Jurisprudence, Psychology, Anthropology, Economics, etc., in which interdisciplinarity and the inspiration that these various domains give us are all important.

It is of great import to draw attention to the fact that one of the problems we have is that Jurisprudence is not progressing enough, given the urgency of legislating on moral machines. Because, for example, there are various types of machine autonomy, but our laws are made for human beings, who we assume have a certain basic-type-of-autonomy, unless they are ill or mentally incapacitated. When making legislation with respect to machines, we will have to start by defining and using new concepts, without which it will be impossible to make the laws, since these have always to appeal to the concepts of Jurisprudence.

It is also important to recognize that legislators are too tardy in following the pace of technique. This is worrying because there is an enormous confusion in the notion that technical progress is equal to social progress. In fact, technical progress—which we are seeing everywhere—is not being accompanied by a desirable and concomitant social progress. Techniques should be used in the service of human values, and these values should be enjoyed equally by all, with the creation of wealth being distributed fairly.

[8]For technical details consult: Pereira et al. (2017).

Duly analysed, History can give us great references and general lines of action. For example, if we consider the great progress and apogee in Greek civilization, in the fifth and fourth centuries BC, we can see that it was possible only because there was a legion of slaves, without rights of citizenship and no possibility of social ascension, and essentially constituted by armies conquered and foreign citizens.

Similarly, we have the possibility of enjoying more and more machines as slaves, which already are so, for they liberate us from effort that can be allocated to them. But we would like everyone to be freer and to equally gain with it, through a fair distribution of the wealth produced by such slavery, which—at least for the time being—does not raise problems of ethical nature.

Well, the opposite is happening. Machines replace people, resulting in an ever-increasing profit for their exclusive owners. The due counterpart, that is, a fair distribution of wealth, is increasingly far from happening, whereas the universe of situations in which the human has no possibility of competing with the machine does not stop increasing.

Hence, a new social contract is indispensable, in which the labour/capital relationship is reformulated and updated, as a consequence of the social impact of new technologies, namely the increase in the sophistication of machines with cognition and autonomy. We will return to this topic with more detail and argumentation.

If a machine is going to replace me, it should do it completely (even in social obligations). In the sense that, by positioning myself in a work activity, I contribute to the Social Security that supports the current retirees; I contribute to the National Health Service; I contribute with the taxes to make possible the governance and development of the country, etc. Consequently, if a machine completely replaces me, eliminating a person from a job whose activity still remains, it must also pay the taxes that I was paying to support the current social contract. Replacing has to mean replacing in all these aspects!

Another noteworthy point linked with the safety of technologies, will also be developed in this book. If we were talking about civil engineering, it would be clear that civil engineers care about safety and quality. There are deontological codes, norms and rules for a building to withstand earthquakes, for walls to insulate noise, etc. Comparatively, but at a much more complex and differentiated level, the entry of software and cognitive machines into the market introduces problems of the same nature. However, it is not possible to reduce the problems of machine ethics to a deontological code that computer engineers must follow, precisely because of the impact this has on human values and social organization and on our civilizational becoming. Therefore, the question of values is unavoidable, and cannot be reduced to mere technical standards.

In fact, there are numerous and repeated collaborative study reports of unsuspecting entities, namely McKinsey & Company, the Pew Research Centre, the OECD, PricewaterhouseCoopers, etc., which point to an increase of 15–20% in additional unemployment in 2030, by virtue of AI alone.

In China it will be worse, reaching even 20% because, while in the Western world people can still access some social mobility, becoming cognitively specialized given their highest educational starting point, in China the level of education is lower and

therefore the ability of people to rise in their cognitive abilities and to stay ahead of the machines is lower and slower. Imagine, then, the employable mass in 1.4 billion Chinese and the impact on them. The topic of unemployment caused by AI in the very AI superpowers creating it, and which will be heavier elsewhere, is well analysed in the recent book by Kai-Fu Lee.[9]

If we do not presently take the appropriate actions, we can imagine the outlines of a future that will not be promising: *Once upon a time* a caste society appeared: that of the robot owners, of the managers of the machines, of those who train the machines, and that of the remaining unfortunate ones. We cannot forget that doctors are now training machines to read X-rays, interpret tests, examine symptoms, and so on. Throughout the world, a multitude of highly skilled professionals, from medicine to economics, are passing human knowledge to machines that will be able to replicate and use it. In short, people are teaching those that are going to replace them. This caste society will, at one point, explode. People will no longer be able to endure so much hypocrisy, so much lack of distribution of the wealth generated, so much automated lies with fake news, and so on. Then it will be chaos.

The dangers of AI do not fit into the possibility of the appearance of a Hollywood-like "Terminator". Actual risks are that, at present, simplistic machines are making decisions that affect us. Though, by calling them "intelligent machines," people think they are doing a good job. This excessive selling of AI, which is currently occurring, is very pernicious in this respect.

In addition, the AI that is being sold is less than one tenth of what AI is, and what applied AI may actually be. The serious AI is yet to come and will be much more sophisticated than most current programs, dubbed deep learning. These are quite simple and limited programs, and we should not be giving so much power to such simplistic machines. But since they replace humans, since they are going to replace radiologists, car drivers, trucks, people in call centres, people in shopping centre security, they are oversold as a panacea.

Luís Moniz Pereira participates in the project "Incentives for safety compliance in AI Race"[10] sponsored by Future of Life Institute (FLI), a non-profit organization endorsed by people like Stephen Hawking and Elon Musk, among others. The project, in the area of software security, addresses the issue of the urgency in reaching the market by firms that develop AI products. More specifically, it examines the consequences of disregarding the safety conditions of those products. The urgency is such that security is set aside, because it costs money and time, and delays the arrival to the market before competitors. The purpose of the project is to establish rules of the game, so that no one will overlook security as if it were not essential.[11] For this we need regulatory and monitoring entities. This topic, as well as the need for a "National Ethics Commission for AI," which includes Robotics, will be detailed in a separate chapter. Here we only alert to the need for parity with the "National Bioethics Commission". It will have to be independent, to be above all interests,

[9] Lee (2018).

[10] https://drive.google.com/open?id=1j59rhP7op3nBpvaxpeCdaBVObJAbzWBJ.

[11] For a summary of the project see: Han et al. (2019).

and to respond directly to the President of the Parliament, without depending on the government.

We cannot accept, as we hear in Europe and the United States, that companies that make driverless cars are exclusively responsible, and that if there is a problem, we will soon see it. That Governments are therefore not responsible for the tests to which such cars should be subjected; they simply delegate to the companies themselves. But look at the recent crashes with the Boeing 737-M,[12] in which the American Federal Aviation Authority (FAA) delegated quality checks to Boeing itself!

In the European Union, responsibility for security appears to be more disguised. A high-level commission for AI and Ethics was created to give recommendations, and, according to all the indicators, what they propose to do is to affirm: We have here some recommendations that the firms must follow. In addition, we have here some private audit firms that will inspect these firms.

We will, perhaps, fall into the same scheme of investigators with interests in the very entities they are examining, because these, in parallel, commission them other studies. Like the case of banks and the financial crisis of 2008. Hence the need for some independent, non-private regulatory entity.

Finally, we live in an increasingly algorithmic society, with everything increasingly systematized, in which the major growing danger—as we have already mentioned and will always mention whenever suitable—is to give excessive power to simplistic machines, because of the risk that exists of them increasingly, systematically, putting us into statistical drawers. Because what these crude machines, of Deep Learning about Big Data, do is, deep down, to recognize specific patterns in a universe of possible patterns. For certain things it is great, it's an excellent technique. But they cannot solve problems for which this technique is inadequate.

Note the following very emblematic example: In the US there are at least three programs used by judges who have to decide whether a particular prisoner is given parole or not. How does the process work? The judges are very busy, as there are millions of prisoners (about 0.6% of the population are in captivity). So, they will see in a Big Data record on people who have been paroled whether or not it went well. In front of them they have a candidate for parole whose profile holds a given pattern, depending on age, ethnicity, religion, geographical area, etc. In fractions of a second, the computer system tells the judge in which standard drawer the profile of the candidate enters, and this quickly and cheaply determines the decision of the judge.

This is a circumstance in which each case—dramatically—is not a case! Prisoners are tucked into statistical niches, which assume that the past is equal to the future. That a population with a certain profile did not evolve, nor did the social customs. People are judged by the historical standard, and that standard will, moreover, be confirmed by the inclusion of yet another instance of judgment.

[12]To save costs: «Boeing's 737-Max software outsourced to Rs 620-an-hour engineers» https://economictimes.indiatimes.com/industry/transportation/airlines-/-aviation/boeings-737-max-software-outsourced-to-rs-620-an-hour-indian-engineers/articleshow/69999513.cms.

It is a clear example of real and effective misuse of these simplistic algorithms. It does not mean that such algorithms do not have their own very useful recess, in which they can present very good solutions for a given niche. In their defence, in this case of application, it is argued that the judges, busy as they are, and especially after lunch, decide worse!

We must not forget that these procedures are also moving to an area that tells us a lot: we refer to medicine. As they are pressured to see more patients per hour, physicians are forced to resort to similar intelligent, pattern-aware programs, with no room to exercise critical sense with their specific knowledge, which is supposed to be updated. In the final analysis, there are no diseases in themselves, they always occur in a patient who has their context and life history. However, the programs currently in existence are not at all prepared to account for this individuality, and the Big Data, say for lung cancer, contains cases from many different population settings and is obtained with quite diverse X-ray instruments of measurement, with distinct granularity and obsolescence.

We will not close this prologue without a short synthesis of the "terms of reference" for the AI scientific community, with finishing remarks on the evolutionary dimension of cognition. The last paragraphs establish the paradigm within which the whole argument is developed. Thus, the topics that make up this book are summarized.

- We need to know more about our own moral facets so we can pass them to the machines. Yet we still do not know enough about human morality. In this sense, it is important to strengthen its study by the Humanities and Social Sciences.
- Morality is not only about avoiding evil, but about how to produce good: the greatest good for the greatest number of people. The problem of unemployment is inherent to this point of view.
- Universities are one appropriate place to address all these issues, for their spirit of independence, their practice of reasoning and discussion. And they contain, in their colleges, the necessary interdisciplinarity.
- So soon we are not going to have machines with an overall moral capacity. We will have machines that know how to respect standards in a hospital, in a prison, and even the rules of war. These are even the better particularized ones, and also subscribed all over the world. As they are well specified, they are less ambiguous and are closer to being programmed.
- We will begin by automating, little by little, the norms and their exceptions, broadening the generality and the ability of a machine to learn new norms, and extend its areas of competence.
- Since these are very difficult subjects, the sooner we start the better!

In order to explain the grounds on which we move, we still have to make an introductory reference concerning evolution and cognition. Research in this area of knowledge has evidenced an integrative perspective. It is possible to see intelligence as the result of an information-processing activity, and to trace an evolutionary line from genes to memes, and their co-evolution. In these terms, traditional ruptures between the human being and the other animals, or between culture and nature, make little sense.

All life is an evolutionary stage, where replication, reproduction, and genetic recombination have been testing solutions for an increasingly improved cognition and action. Given its computational matrix, Biology establishes a first artificiality over Physics. Thus, the current state of knowledge implies a redefinition of the human being's place in the world, posing challenges to several areas of knowledge. From the onset, to many disciplines of Philosophy, because problems such as what is to know, what is man, and what are and how values of moral nature emerged gain here hitherto unthinkable perspectives.

As far as knowledge is concerned, the possibility arises of it being simulated in computers, thus overcoming the limits that were previously imposed by a speculation that could not be more than mental experience, perhaps shared.

With regard to anthropological questioning, the traditional discussion on "What is Man?", thanks to the crossbreeding of AI, genetic engineering and nanotechnology, is now replaced by a powerful and challenging problem around what is desirable, possible or likely for Man to become.

From the viewpoint of action criteria, the morality perched from the sky of the past is confronted with a new perspective on the rising moral systems studied in evolutionary psychology and deepened through testable models in artificial scenarios, as now allowed by computers. As research proceeds, we can better understand the processes inherent to moral decision, to the point that they can be "taught" to autonomous machines capable of manifesting ethical discernment.

In the field of economics there is a whole pressing problem associated with the impact on work and its inherent dignity, as well as with the production and distribution of wealth; that is, a reconfiguration of economic relations that will result, not only from the automation of routine activities, but fundamentally from the arrival of robots and software that can replace doctors, teachers, or assistants in nursing homes (to mention professions which are not commonly believed to be replaceable by robots) . Knowledge of this context is especially relevant, requiring positions that will sustain the need for up-to-date social morality and a renewed social contract.

Thus, the problem of computational morals gains existence in a context in which the knowledge ecosystem will be greatly enriched, since it will have to incorporate non-biological agents with the capacity to become active players in dimensions that, until now, have been attributed exclusively to humans.

References

Han, T. A., Pereira, L. M., & Lenaerts, T. (2019). Modelling and influencing the AI bidding War: A research agenda. In: *Proceedings of: AAAAI/ACM Conference on AI, Ethics, and Society, (AIES 2019)*, Honolulu, Hawaii, USA. https://userweb.fct.unl.pt/~lmp/publications/online-papers/AIracemodelling.pdf.

Kershaw, S. (2007). *A brief guide to the greek myths*. London: Constable & Robinson Ltd.

Lee, K. F. (2018). *AI super-powers—China, silicon valley, and the new world order*. New York, NY: Houghton Mifflin Harcourt.

Mayor, A. (2018). *Gods and Robots—Myths, machines, and ancient dreams of technology*. Princeton, NJ: Princeton University Press.

Pereira, L. M., Lenaerts, T., Martinez-Vaquero, L. A., & Han, T. A. (2017) Social manifestation of guilt leads to stable cooperation in multi-agent systems. In: S. Das et al. (Eds.), *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*, pp. 1422–1430, São Paulo, Brazil. https://userweb.fct.unl.pt/~lmp/publications/online-papers/guiltEGT.pdf.

Chapter 2
Artificial Intelligence, Machine Autonomy and Emerging Needs

Abstract Questions associated with AI require us to consider the landmark figure of Alan Turing. Among us on every computer or smartphone, he was the first to imagine all the functions inherent in the way we process information could be reproduced in a machine. From the functional viewpoint, it is irrelevant to distinguish processing by a human from that by a cognitive machine. If the latter wins the imitation game, i.e., if the human does not distinguish the answers provided by his thinking from those of machines, then the machine must be assumed to be thinking autonomously. For unfortunate reasons, Turing did not live long enough to witness the complete triumph of his conjectures. It is possible at the moment to build autonomous cars, robots that work in nursing centres and help doctors with diagnoses, or programs that perform financial transactions on the Stock Market without human intervention. Such achievements and those yet to come require the development of a computational morality.

> "*Physis kryptesthai philei.*"
> *(Nature loves hiding itself.)*
> Heraclitus od Ephesus (c. 535–c. 475 BCE)

Alan Turing (1912–1954) is one of the greatest names in twentieth-century science, and his thinking is more relevant than ever. Mathematical brilliance earned him the leadership of the team that built the electro-mechanical machine, Bombe, that deciphered the codes of the Enigma, the device for encrypting messages invented by German engineers during World War II. However, over and above this historical achievement—by which he made a decisive contribution to the victory of the Allied Forces—is his hypothesis according to which it will be technically possible to build a machine capable of evidencing autonomous intelligence. According to this conjecture, such a machine could constitute itself as an interlocutor of human beings, constructing and sharing knowledge. Turing assumed that there was no a priori limit to the fact that we could export all our computable functions to another physical medium, namely electronic; plus if the interaction with one such a machine does not provide us with any criterion capable of characterizing its responses as mechanical, then it will win the *imitation game*—known today as the "Turing Test"—forcing

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_2

us to consider its behaviour as intelligent. These two theses were presented in his programmatic *Computing Machinery and Intelligence*, published in the prestigious philosophy journal, *Mind*, in 1950.

Alan Turing had the audacity to ask if a machine could think. His contributions to understand and respond to this and other questions challenge a conventional classification. In the midst of the 21st century, the concept of Turing Machine—created in 1936—applies not only within mathematics and computer science but also to the cognitive sciences and theoretical biology. The so-called Turing Test continues, according to Hodges (1983), to be the cornerstone of the theory of Artificial Intelligence. In addition, Turing succeeded in conceiving, alone, a visionary plan for the construction and use of an electronic computer. He thought and lived far ahead of his time, and the attributes of his thought, which ripped boundaries, remain very much alive among us.

Turing addresses the question of computable functions from a perspective opposite to the usual one. Not from the point of view that all functions can be constructed from a set of primitive ones, but rather by the strand of the digital symbols produced as the output of the application of a function to its result. Turing inaugurates the informal idea of computer—which in 1935 meant, not a calculating machine, but a human equipped with pencils, unlimited paper and time, doing calculations. Later, he replaced certain components with other unambiguous ones, until nothing else remained beyond a formal definition of "computable." With this objective, Turing introduced two fundamental ideas: discretization of time and mental state. This way, the Turing Machine embodies the relationship between an unlimited sequence of symbols in space, and a sequence of events in time resulting from the manipulation of those symbols, regulated by a finite number of intermediate internal mental states.

The title of his article *On Computable Numbers* marks a radical change. Before Turing things were done to the numbers. After Turing the numbers, understood as encodings of symbols, began to do things. By showing that a machine could be coded as a number, and a number decoded as being a machine, *On Computable Numbers* treated the numbers—now called *software*—which were 'computable' in a whole new way. He outlined the proof that π is a computable number, along with all the real numbers defined by the common methods of equations and limits in mathematics.

In the post-war period, the mathematician argued that his Turing Machines (which he called A-machines) can mimic *any* activity of the mind, not just the mind fulfilling a "determined method." Turing is clear as to the possibility of discrete-state machines including learning and self-organizing abilities—namely embryonic development—and emphasized that they are still included within the scope of the computable. He also drew attention to the apparent conflict with the fact that the definition of Turing Machines had fixed behaviour tables, and outlined the demonstration that self-modifying machines are, indeed, still defined through an unchanged set of instructions. Turing advocated two different approaches—*top-down* and *bottom-up* in modern language—which in fact derive from his description of the 1936 machine model. Explicit instruction notes become explicit programming; implicit

mental states become the machine states, struck by learning and the experience of self-organization.[1]

It is true that, up to now and for the time being, we have not yet been able to produce a machine which, from the viewpoint of its general intelligence, flexibility and adaptability, could fully correspond to Turing's hypothetical speculation and stand on par with us. However, given their ability to learn and process information and their presence in increasingly autonomous decision-making processes, the already existing artificial agents require—as has already been shown—that the computational morals be immediately and urgently deepened. An immediate reactive approach may lead one to think that this theme is not only devoid of urgency but also involves some eccentricity. After all, there is a myriad of available human moral systems, and one of the major philosophical problems may even be to confront this diversity, without falling into the web of moral and cultural relativism.

In order to give due notice of such urgency, it is important to reinsert the question from two points of view. Firstly, to note that moral systems from distinct religions lack recognition by human beings of different beliefs, or those without religion; and—at the limit—they will not apply to machine agents. After all, besides not having divine filiation, nor souls that they wish to save, these also lack the feeling of faith. Secondly, the so-called humanistic ethical systems, based on an anthropological conception originated in the Renaissance with roots in Protagoras—Man the measure of all things—will also not be able to respond to the challenges posed by the development of AI.

Delving deeper into the problem, we must remember that, traditionally, morality has been a domain claimed by either Religion or Philosophy. This aspect has some consequences that must be considered. In the case of religiously inspired morals, their validation criteria consider mostly revelation and faith. A commandment is true because God so ordained it, and people have accepted it because they believe in this revelation. Thus, when deprived of such faith, all humans, or other rational agents, are left outside of these moral commitments. On the other hand, and in a way inspired by rationalist and illuminist philosophies, at some point in their evolution human beings conjectured that they would not need religion to validate principles and rules governing individual conduct and social organization. Curiously, one of the main interpreters of this conjecture was a deeply religious philosopher; we refer to Immanuel Kant. His article "*What is Enlightenment*"[2] is programmatic in this area. In order for humankind to abandon the infant state of obedience to rules, it is necessary for every rational being to think for themselves resorting, however, to criteria of a universal nature. We will return to Kant later, since for the moment we are only interested in making explicit that, regardless of their religious or philosophical origin,

[1] For a detailed presentation, for laypeople, of the operation of Turing Machines, see the book by Pereira (2016).

[2] https://www.stmarys-ca.edu/sites/default/files/attachments/files/Kant--What.%20Is%20Enlightenment_.pdf.

the methodologies of research or validation of moral knowledge have followed a *top-down* strategy. This means that either a transcendent dimension or any ideological conglomerate imposes itself on individuals.

The relatively recent emergence of a scientific approach to the issue of morality has allowed for a reformulation of the research paradigm, by introducing *bottom-up* strategies. One of the fundamental objectives of this research is the study and validation of certain moral behaviours, regardless of the cultural contexts where they were born into, and the way in which they emerged. It is presumed that, having survived up to the present, they were the subject of Darwinian selection, and therefore, have been those who proved to be more apt to the formation and guarantee of sustainability of the groups that adopted them. Everything we say in this book will have this evolutionary consideration in the background.

Moreover, in order to understand the need for a different approach to the problem, we have yet to overcome a prejudice that makes less and less sense. Throughout the evolutionary process morality has always been the exclusive subject of human beings, and in the limit, considerations of this nature involving other living beings have always been readdressed to the responsibilities of the former. Thus, to the extent that other animals do not make decisions based on arguments, such a perspective was adequate until very recently. However, today's cognitive machines, and their application to contexts such as autonomous driving, human resource recruitment and management, diagnosis and prescription of medical treatments, management of personal data processed by Big Data companies, pose challenges that directly involve the explanation of principles and rules of a moral nature. Machines that decide autonomously will have to be able to explain the justifications that led them to each of these decisions, so that to be accepted credibly.

Imagine an automobile with autonomous driving ability. You will have to incorporate programs that, in stringent contexts, will allow it to decide whether to run over pedestrians to protect the integrity of its occupants, or put their occupants at risk to protect pedestrians who may be in its path. It is known that, statistically speaking, human drivers in similar situations take a set of evasive measures in order not to run over pedestrians. As a result, it is common to experience off course incidents with potential hazards to the driver and occupants of the vehicle. There are at least two major themes that arise in this or similar contexts. First, we must reflect on how to program the car; which is the most coherent option, supported by good arguments? Despite all the complexity associated with the previous scenario, the problem does not end here; it remains to be seen if the reader would buy a certain car if they knew that for "it", in situations of severe uncertainty, the priority would be to protect pedestrians. We easily assume that a human being can make that decision in his interest, but a machine might, perhaps, be prevented from doing so. Why would it be so?

Imagine a cognitive robot capable of diagnosing a wide spectrum of diseases, as well as of determining the appropriate treatments. In certain circumstances the therapies reach cost values expressed in six digits. Which criteria will we provide our machines? What will we let them learn about hospital management? Can they learn, if we grant them autonomy, to decide what values and moral rules? We live in a context that can be described as one of moderate scarcity, we are far from

absolute poverty, but there are not enough resources to satisfy everyone with the same needs and wants. How will the machine decide if a certain therapy applies, or does not apply, to an elderly person? If it has to choose between saving a middle-aged doctor—who may in turn help a lot of people—or a youth who, having a much longer life expectancy, still has no specific role in society, how will the machine decide? Considering these matters, the reader can already anticipate that we cannot conjecture a specific morality for machines, and another for human beings; morality is a theme that concerns autonomous cognitive agents, able to reasonably justify their decisions, since they coexist in a society of humans.

If we wish, we can also consider the special power of companies that resort to 'Deep Learning' processes on 'Big Data', and their weight in today's society. We live in a world where new interest groups act with processes similar to the old Freemasonry, and the strength of the big businesses threatens to turn true democracy into an unnecessary parody. It is true that this system of political organization does not have to be the last word in ideological or procedural terms. However, until now, research in the field of Philosophy and Political Science has failed to envisage an effective and better solution. The problem is complex because it is necessary to legitimize power (in a democracy it emerges from the vote of the citizens), to distribute it (that is why legislative, executive, judicial and control bodies were created), to guarantee access to quality information, and to expect individuals to seek them out so that they can collectively engage in a more active and more conscious citizenship. A full democracy is extremely demanding towards its citizens. Nowadays, in addition to the traditional lack of interest in knowledge, evidenced by a very significant part of the population, there is a manipulative capacity in a dimension not previously known. In fact, 'Deep Learning' companies for 'Big Data' have the power to trace the psychological and ideological profile of virtually everyone who interacts with the Internet through smartphones or computers. Based on this information they use highly efficient and powerful means to promote the behaviours ordered by their clients. In order to do this, they take advantage of the increasingly weak criteria in the selection of qualified information, the low critical sense of most people, and the emotional reactivity that characterizes them. Reality is always ahead of our ability to reflect and find criteria of critical reasoning that can set boundaries. Never have the threats to democracy been so able to implement a dystopia capable of enslaving the wide majority of the population. But it does not have to be so if the action of these companies (and even States) is properly regulated through standard norms firmly anchored in widely shared moral values, if more people are led to accept the challenge of knowledge. If we can get better and better organized NGOs, we will have the opportunity to take advantage of technological advances to promote a fairer, more egalitarian society in terms of opportunities and distribution of wealth, and with more common sense. We do not know whether democracy will survive the challenges posed but, given the competing territorial organizations today, it would be very positive if it managed to survive.

In order to reach a global understanding of the current challenges posed by AI, we still need to consider its impact on the world of work. As we will detail in due time, cognitive robots and software will enter virtually every domain of human functioning.

The fact that they do not get tired, do not distract themselves, do not get sick, do not have a predisposition to wage demands, turns them into very desirable assets. This is the third axis demanding an inquiry in the moral arena, in such a way that this will become the basis for a new social contract. The alternative will remain the already mentioned slavery dystopia. In short, how we treat and integrate our digital partners will set the tone for what the society of the future will be. AI emphasizes the acuity of these moral questions, in the way it is used and taken advantage of, since it potentially leverages a lot the wealth gap between the very rich and the others, but could instead translate into a greater and better distributed wealth for all, with very significant consequences on employment and lifestyles.

Therefore, this book aims to characterize the AI development processes that have allowed us to come to this point, to argue for the need for much greater scientific research in the field of morality, to present some studies that have already been carried out, and to open avenues to what the future will be.

The issues to be addressed are, among others, the following: Can we use machines to do things that have hitherto been unique to humans? What is intelligence? What are the limits of computing? Are there other entities that can have symbolic or representative cognition besides human beings? What are the social consequences of the massive introduction of cognitive machines in the economic fabric and at the service of social control devices? What research has been undertaken, in the field of morality, suitable for use in cognitive machine programming? What capacities will these machines have in order to process decisions of a moral nature? Is it possible to produce a machine so skilled that it is able to rewrite its own software in order to become progressively smarter?

References

Hodges, A. (1983). *Alan Turing: The Enigma*. London: Simon & Schuster.

Pereira, L. M. (2016). *A Máquina Iluminada—Cognição e Computação*. Porto: Fronteira do Caos Editores.

Chapter 3
Intelligence: From Natural to Artificial

Abstract The term "Artificial Intelligence" becomes understandable only if we can explain what the terms "Intelligence" and "Artificial" mean, as well as the connection between them. The term "Intelligence" will be viewed as a general faculty to solve problems in a variety of contexts. Arguments will be presented in favour of the evolutionary character of intelligence up to the present, in which there are beings—we humans—capable of producing symbols of abstract nature, capable of representing conjectures about reality, on the basis of internal models. On the other hand, Nature will also be interpreted as a support for the creation of artificiality, in that increasingly complex and functional codifications emerge. Thanks to their capacity for symbolising, human beings for the first time implement a complex system of synchronic and diachronic networked functioning. The transition from an organic to an engineered platform for cognition might be interpreted as "just" another evolutionary leap.

Inevitably, our starting point will have to be the question of intelligence, including in this approach the analysis of the expression Artificial Intelligence. Before we ask a first question, we must consider that, functionally, this capacity has been attributed to bodies with the practical ability to solve problems. In this sense it is not difficult to accept that its distant matrix is evolutionary, and that it has had a beginning both as rudimentary as efficient; see the case of viruses, which have the ability to crystallize in an adverse environment and reproduce in a suitable environment. The effectiveness of this reactive process for sure reveals a certain intelligence. Note that we are talking about a being that has neither Nervous System nor DNA.

However, as organisms evolved, so did the way they evinced practical intelligence become more complex, with progressive gains in intentionality and consciousness. There are two areas that are especially relevant in this process. On the one hand, we have the case of languages—from bees' code to human language—that have undergone a process of complexification and diversification, reaching its peak with abstract human languages. On the other hand, we have the level of the instruments, which have a very simplified expression in nonhuman primates, and have been diversifying with humans, to the point that today we have much machinery capable of expanding our abilities, as well as those of other animals. And we speak not just about flying or

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_3

diving, but also about processing information in rhythms and dimensions far superior to biological intelligence.

Intelligence is, thus, a phenomenon distributed throughout various animal species and artificial agents, being shared by the generality of individuals within each species. This aspect is so relevant that we will address it on a number of occasions, properly exploring its consequences. For now, it is important to stress that, while an emerging process, intelligence is a property of certain organisms that allows them to interact with what would be a natural causal sequence, altering it.

There is, therefore, a sense in which we can say that all intelligence is artificial, for it establishes a new causality based on memory, from the inside out, superimposed on natural causation, which operates from outside inward. To say that it is artificial is not the same as stating its flexibility. In fact, most of so-called intelligent acts are even quite standardized. So much so that when we remove a certain living being from their context, their portfolio of responses is not broad enough to ensure their adaptability. Human intelligence is flexible and adjustable, requiring introspection and internal processing. Because of this, until a few years ago philosophers such as John Searle and many others expressed reservations about the possibility of being exported to machine devices.

In fact, it remains a commonplace to assert that machines only do what they are programmed for; therefore, they will not be able to surprise us. This consideration, while raising no difficulty in using the term "intelligence"—because they still solve many problems—raises great difficulties as to the issue of machine autonomy. In this context, it makes sense to ask the following question: What do we call artificial intelligence? Will this artificial intelligence achieve qualitative gains to the extent that the agents that incorporate it can demonstrate autonomous behaviour?

– I would begin by stating that this question makes it possible to clarify certain aspects of the emergence of intelligence, as well as to mark out much of what we will say throughout our conversation.

But before answering I will first reinforce the idea that all biomass actually establishes a code-based artificiality, which overlaps with the physical-chemical processes prior to the onset of life. This means that life is coding, and that this coding is the basis of the replicating of organisms, which have endured evolutionary games. In this sense, we can say that there is a certain intelligence, and autonomy, inherent in all vital processes.

Of course, in humans, we manage gains in consciousness, representation, expressiveness, ability to act on the environment and to change that very same environment, which places our species at levels not comparable with other known living beings.

We have, on the other hand, the evolution of the machines having been towards a progressive approach to humans. Examples of this are all the automata we have been building, from the Golems to the Frankenstein Creature, or the machinist toys that over the years have fed children's fantasies. We will come back to this topic in detail.

We also have the notion that our higher faculty consists in thinking, and therefore inventing a thinking machine has always been something that has been present in

human fantasies. Perhaps the abacus, and its presence in various forms, in diverse human cultures that did not come into contact, may be seen as a secondary gain of this intention. However, only in the twentieth century, with the invention of the computer, was it possible to have a device capable of allowing one to dream of this embodiment. Indeed, computers are telescopes for complexity, in the sense that they allow us a radically different look at our processes of representing reality. The advent of computer science, both theoretical and practical, provides us with machines capable of receiving, processing and outsourcing information.

It was with this technological base that the science and engineering of Artificial Intelligence emerged, that is, our search for the way to endow the machines—the computers—with the capacity to think.

Artificial Intelligence is a science whose origin we can situate in the late 1940s, early 1950s. Since then much progress has been achieved, and although we are still a long way from a totally autonomous intelligence from ours, we have already been producing several intermediate results that are very useful in our daily lives, marking their presence in almost all dimensions of our lives. They help us in financial management, optimising energy resources, entertainment provided by computer games, hospital management and diagnosis of diseases, optimization of our smartphone, and in almost all other dimensions of our lives. If until recently the common questions were around whether or not it was possible to implement intelligence in a machine, today one of the main problems of politicians and entrepreneurs have relates to the way of articulating human and digital entities in the business or social ecosystems.

If we look carefully, we will see that the history of computing as a whole is taking the missing steps in the progression of approximation of machines to humans: we have developed ever more friendly and intuitive interfaces; increasingly expressive and powerful programming languages; computational systems that begin to exhibit creative behaviours, something that we previously believed to be the exclusive competence of humans (or, at least, biological beings). Nowadays we have increasingly intelligent and autonomous computers and systems that now reach a point where they have the capacity to make important decisions, and for this reason, it is also time to start talking about computational morality, to debate, to define, and continuously improve a code of conduct for artificially intelligent agents—notice that it is the embrace between Men and Machines on the one hand; and on the other the production of evidence as to the need for a new and different approach to the question of intelligence.

All these aspects are a part of the answer to the question posed. However, in order to better detail the AI program with the challenges it brings to society, it is imperative that we return to the figure of Alan Turing. What Turing did was to develop an abstract notion of computer, which is, furthermore, implementable. For him, a computer is something that elaborates a computation, something that, given an input, can provide an output, executing for this pre-defined calculation rules that, significantly, can change themselves. Also, we often want to have (but not always) the guarantee that, for certain inputs, certain computations produce the output we want. We speak of algorithms, of manipulating the information to produce results that this information allows us to conclude. That is, calculate the structure of a building, a dam,

do the calculations to build an airplane, demonstrate a theorem; in short, anything that involves an ability to represent knowledge in symbols, to manipulate those symbols through a number of operational rules for drawing conclusions. Mathematicians had already raised this question. What is the limit of mathematics? What can I, as a mathematician, achieve safely? In other words, and similarly, what are the limits of the computable? Are there things that cannot be computed? And if they are not computable, they will not be manipulable by a well-defined set of rules in order to guarantee a result. What part of mathematics is computable?

What if we ask ourselves: if we are going to create a robot, will it do everything we want it to do? Will this computer be able to do any calculations, or will I have to have one machine to run Word, another to run Excel? Must I still have different one for databases? Will there be a general computer that does it all? It is a question similar to that of the mathematicians.

Turing reformulated the mathematical question in terms of computers. The ability to build something that calculates everything, without loss of generality. And then there is the question: and build something with what? With stoppers or with rubber bands? What Turing showed was that it was possible to conceptualize the functions that the mathematicians defined, the computable functions, in such a way that it was obvious that it was possible to construct an artefact to carry out the computation. These devices are today called Turing Machines. Such a machine is extremely simple. You start with a kind of unlimited roll of paper cut into squares (similar to toilet paper), to which more squares can be added whenever necessary. Then it is necessary to have a finite automaton—it is a simple graph of states and transitions between them, with finite memory (finite number of internal states) which, when it sees a symbol in one of these squares of paper (and it only inspects one at a time) reads the symbol, changes its internal state according to its graph, moves to the square to the right or to the left or stays in the same place in relation to the square where it read the symbol, and writes a symbol in that place, as is stipulated by a state transition rule. Then does the same again. Turing proved that all computable functions could be performed by a machine as simple as this. That is what he demonstrated for the first time: that all algorithms can be executed by a computer with this extremely simple ability to perform the operations explained. Alan Turing later showed that there was a Universal Machine, that is, that it was not necessary to have a machine for each thing, for each algorithm. A single machine would suffice. In the roll of paper with squares of the said Universal Machine I can put the description of the rules of any other Turing machine, and then this universal machine would be able to imitate what the first one would do. Meaning the computer is a huge chameleon. That is why we talk about putting the program on the computer: the program is basically a small machine, placed, not on the roll of paper, but "wrapped" in memory; a running program is a collection of data in memory in the form of instructions.

The Universal Turing Machine is capable of simulating any computation, imitating what any other Turing machine would do by consulting its rules, when such rules are expressed on its own paper tape. He published this result in 1936, showing, precisely, that one could define computability with this type of tool. Turing built this theory and only then went into practice. The scientist designed and participated

in the construction of the world's first computer in Manchester: The Atlas. At that time there was a great investment in science. Unfortunately, it took wars to make it happen …

To conclude, we return to our comparison between the science of Artificial Intelligence and Physics: the latter deals with a reality that exists, which has immutable laws—although some laws may, perhaps, change (one conjectures, it is not known), but very, very slowly—and we believe that this reality is independent of us and that the laws are universally valid, everywhere in the universe.

In the case of Artificial Intelligence—and here the word "artificial" is decisive, because we do not want to imitate human intelligence, there will not be great interest only in this—we want to explore the ability to build intelligent artefacts, just as a musician explores how to build symphonies and musical objects from basic parts. It is a completely open task, which has no end, because it is the domain of the manufactured, of the engineered. We are not characterizing something that exists and is immutable, we are creating new things. We can undoubtedly draw inspiration from the real, from human intelligence, from that of animals or some extra-terrestrial if we find it, but in practice we want to elaborate theories about what it is to be intelligent in general. It could therefore be called General Artificial Intelligence (GAI), according to terminology currently used.

Assuming a general and operable definition of intelligence, we will say that it consists in the ability to perceive an environment where there are other agents and be able to achieve certain objectives in that environment; plus at the same time evolve—because that environment is also changing. Let us say that living beings, and in particular the human species, are the paradigm of the universe's intelligence being born, the universe becoming conscious, looking at itself through us and projecting intelligence to levels unthinkable before this new dimension of artificiality. It should be noted that reasoning does not exhaust the notion of intelligence, nor the latter the notion of knowledge. Going to the root of the word, "intelligence" must be understood in the broad sense of the Greek '*entelekia*', that is, the ability to understand. This involves perception, the creation of perceived reality models, and the ability to decide to act on that reality, confronting expectations with the outcome of the action and then correcting it. '*Entelekia*' literally means the ability to act according to ('*en*') an objective ('*telos*'). The measurement of such Intelligence will occur through the ability to successfully carry out the tasks inherent to survival in a complex and makeable more complex universe.

At present what we can argue is that not only will machines gain their autonomy status, but the man/computer relationship prepares a paradigm revolution, which will culminate in the abolition of the man/machine frontier, turning it symbolic, i.e. expressible with symbols. An aspect that brings about intellectual challenges unthinkable only a few decades ago.

3.1 Here is, in Short, a Definition of AI

The field of Artificial Intelligence (AI) aims to understand intelligent entities. One of the reasons for studying them will be to understand ourselves better. Unlike Philosophy or Psychology, which are also interested in intelligence, AI also intends to build intelligent entities, and in this aspect, it is a science of the artificial, similar to engineering, occupied with all the aspects relevant to their constructions.

It is clear in its short time of existence that AI has produced significant and exciting artefacts, and that its impact on our future daily life and on the very course of civilization will be enormous, in conjunction with Computer Science, Biology, and Artificial Life and Societies. It is expected to play a leading role in an inevitable worldwide network of information and knowledge, a substitute for the current Web, which is mostly passive, and which will have rational, introspection and monitoring capacities, as well as initiative.

Being a very recent discipline, initiated in 1956, despite its many results it has, as might be expected, many exciting and deep open problems. With the philosophical study of intelligence more than 2000 years ago, only the advent of modern computers in the early 1950s allowed us to move from armchair speculation to the realization of functional models in vitro, i.e., in the computer, and therefore with observable and repeatable behaviour, that is, an objective one.

Without intending to imitate humans or animals, but being inspired by them, AI systems are those that think and act in a rational way, and are modelled and implemented in computational terms. Computer, networks, robots, and specialized computer hardware allow us to execute the necessary algorithms, perceptions, and actions.

In synthesis, at its core, AI is a scientific discipline that uses the computer-processing capabilities of symbols to find generic methods for automating perceptual, cognitive, and manipulative activities via algorithms. It should be remembered that an algorithm is a safe and rigorous method to achieve a result. AI includes both aspects of psychoanalysis and psycho-synthesis. It has a fundamental analytical research component accompanied by experimentation, and an engineering synthesis component which, together, are promoting a technological revolution: that of the automation of mental faculties through their implementation in computers. Here are, also in synthesis, the contours of the areas of the AI and their respective domains:

3.2 Areas of AI

Functionally, the main areas of AI can be organized as follows: **Problem Solving**, which includes *Search Methods* and *Games*; **Knowledge Representation and Reasoning**, where fit *Knowledge Bases*, *Logics and Inference*, *Restrictions*, *Uncertainty*, and *Methods of Decision*; **Planning of Actions**, which include their respective *Distribution* and *Cooperation*; **Learning**, which encompasses *Induction*, *Clustering*,

Neural Networks, and *Genetic Algorithms*; **Communication**, which comprehends the *Natural Language*, written and spoken; **Perception and Action**, which involves *Robotics*; **Agents**, both singular and collective; and **Philosophical and Cognitive Foundations**.

Many are the branches of knowledge with which AI has a strong connection: Computer Science and Computers, Philosophy, Cognitive Sciences, Linguistics, Logic, Psychology, Mathematics, not to mention its multiple applications in a variety of domains.

3.3 AI as Symbiosis

The computer makes the ambitious AI project possible because it is a machine that processes symbols in an automated and efficient way, and with the greatest generality. Basic to understand this generality is the distinction between hardware and software, rich in consequences. Namely, it explains the non-obligatory correspondence between the processing of a certain function, for example a cognitive one, and the material support that performs this processing. At the physical level of the computer, the hardware is not only specific to a single function performed by the software, but rather it enables the execution of any function defined by the software. Let us say it is the software that drives the hardware.

AI is only possible because of this independence. Otherwise, we would be studying the intelligence of computer A, the ease of learning of machine B, the fluency of automaton C, or the decision-making capacity of the brain D. That is, everything in particular, but nothing in general.

Let us accept the two assumptions that the brain has largely a component of symbol processing, and that there is in large part hardware independence from software. That is, we can discuss the issues of processing symbols that perform the mental functions of the brain without appealing to the organic operations that support them, but that can instead support them in the computer.

So, the computer allows us to better explore our knowledge about certain dimensions of thought, both by its capacity for retention and precise processing of information, and by its speed, appearing as an instrument that is a kind of telescope for complexity. In fact, if with the telescope we started to see the farther, with the computer we started to see the more complex. It is actually the first instrument with significant amounts of passive memory, quickly, rationally, and automatically manipulatable by an active memory, in the form of memorized instructions, ipso facto allowing unlimited complexity.

But AI entails a symbiosis. There is no fixed and immutable way of thinking. The ways of thinking evolve, perfect and combine. Ultimately, AI is, and will continue to be, the result of a symbiosis between the way of thinking of man with the potentialities that the machine adds to it. The latter appears as a reflection, an epistemological mirror of man, while programmer of the machine, not forgetting that it can evolve by itself. The new and wonderful active instrument that is the computer animated by

AI provokes our imagination, and with the help of the invention allows us to explore possibilities and to elaborate artificial worlds with which we dominate reality.

The end result is a symbiotic complementarity, in which the limitations of AI will be no more than our own limitations as creators, since the computational clay is infinitely malleable.

3.4 The Next Millennium: Prognostics

AI is a young discipline that is not yet 65 years old. The scientific questions it addresses and the technological achievements it proposes are among the most complex ever sought by the human being. Believing in their fruition, intelligence and spirit will not be the appanage only of man, who also in this field will also no longer be able to claim for himself the centre of the universe.

Indeed, it will be quicker to find mental company on Earth rather than on distant planets, and our first emissaries to these remote lands will more easily be robots than ourselves.

As a scientific discipline, and as a technology that is dedicated to automating the acquisition of knowledge, reasoning, and action, AI will increasingly be an instrument for other sciences, be they natural, economic, human or social.

The next millennium will see the confluence of AI, Artificial Life, Biology, and Neuro-Sciences. And, inevitably, distributed AI and Artificial Societies. There will be a symbiosis between man and humankind with these creations of theirs, with which they will evolve together, towards an earthly mind entity with self-consciousness.

The vision of that future must lead us to the careful consideration of desirable options. To conceive and implement these options we need more research and resources, not less. The “bug” of the year 3000 might not exist but, if it does occur, it may be eventually due to our nonchalance in laying the foundations of an artificial ethics, by means of which we respect and make ourselves respected by our own creatures.

Chapter 4
Intelligence and Autonomy in Artificial Agents

Abstract An intelligent agent will, inherently, be an autonomous agent. Assuming this thesis is pertinent, it becomes necessary to clarify the notion of autonomy and its prerequisites. Initially, the difficulties inherent in developing ways of thinking that make it effective must be acknowledged. In fact, most individuals deliberate and decide on concrete aspects of their lives yet are unable to do so critically enough. This requires a complex set of prerequisites to be met, which we make explicit. Among them is the ability to construct hypothetical counterfactual scenarios, which support the analysis of possible futures, thereby leading the subject to the construction of a non-standard identity, of his own preference and choice. In the realm of AI, the notions of genetic algorithms and emergence, allow for an engineered approximation of what is viewed as autonomy in humans. Indeed, a machine can follow trial and error procedures, finding unexpected solutions to problems itself, or in conjunction with other machines. In theory, though we are mindful of the difficulties inherent in the construction of autonomy, nothing in principle prevents machines from attaining it.

Returning to the topic, one of those "unthinkable challenges" will certainly be related to the process of implementing autonomous behaviours in machines.

Before we approach the issue of machine emancipation, some preliminary considerations have to be made. The first concerns the difficulties inherent in the development of autonomous thinking in humans themselves, both individually and collectively. Let us first examine the collective dimension: Although there is a whole politically correct discourse around freedom and autonomy, the truth is that, culturally, one tends towards the formation of great and diverse patterns, and to the regimentation of people around them. It is generally known that majorities are blind, and that innovations and divergences are characteristic of minorities. Moreover, as in all times, most of the "innovations" and alternatives to common thinking are innocuous, if not bizarre. Only a small part is used to produce constructive cultural dynamics.

From an individual viewpoint, ordinary education does little for autonomy. Most people live by rules that not only avoid self-thinking but promote it. The majority of moral rules belong in this domain. To address this problem, Kant established a

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_4

difference between material ethics and formal ethics. He associated the former with the compliance of rules in view to a purpose, such as not stealing to save the soul. In this case, the appearance of autonomous thinking is related to the fact that the agent is concerned about their own soul. It should be noted that, with their codes of conduct, these moralities are extremely efficient in how they provide individuals with a pre-defined life plan. In this sense, they can, quite evidently, be associated with the idea of closed programming. Now, in this context, it is easy to imagine that a machine can be taught to think, including morally.

The problem arises with the second alternative. It is said of rational beings that they are autonomous when, instead of reasoning on the basis of rules, they have the ability to start reasoning on the basis of principles. This implies that each agent is capable, according to criteria of great generality (in Kant they have the pretension of universality), to choose, in each scenario, the best decision to make. It may also happen that, given the co-presence of various moral systems, an agent has the ability to decide which of the systems they will choose, or which combination of them. Note that this possibility does not directly make them selfish, or someone to avoid.

Such aspects have been improved in domains that support an autonomous decision theory, often with moral connotations, but not exclusively.

In addition, autonomous thinking often resorts to cognitive tools that rely on conjecturing hypothetical futures, supported as well on the ability to construct counterfactual reasonings about what could have been different in past choices, and their distinct consequences, possibly taking into account what is known today. The various futures outline possible scenarios so that the agent can simultaneously analyse a fairly diverse set of variables. Finally, it is supposed that the agent has the ability to discern judiciously which is the best solution to adopt. That is, they must have the ability to deliberate and know how to choose between alternatives even if, in the end, they only have one that can be chosen.

It should be noted that the term “Critical Thinking” has become a jargon of academic curricula and syllabi of several disciplines, ignoring the difficulties inherent in the development of such a capacity. It entails a qualified collection of convergent and divergent information, analytical ability, valued self-esteem, which allow the criticisms with origin in an eventually qualified majority to be faced squarely. Developing critical thinking is not easy, it cannot be decreed and, above all, there is no guarantee that, even following a very intent plan, the result will be achieved. Critical thinking is an attitude; this implies that the individuals themselves have to decide to develop it; it can never result from mere conditioning. Moreover, schooling is generally conducted against critical thinking, since, in the overwhelming majority of circumstances, and especially in engineering, it implies the learning of previously formatted content, accepted as received wisdom.

Given this diversified and complex repertoire of prerequisites, namely those derived from the agent‘s internality, will autonomy actually be implementable in a machine? Or will we have nothing but an imitation, even if perhaps quite credible, of what we understand to be this capacity?

There are many misconceptions about the idea of freedom, free will, and autonomy. Possibly one of the best ways to force ourselves to better know these concepts

is by the act of implementing them in a machine, and this gesture implies an entire *engineering of freedom* that requires detailing certain processes step by step.

Let us begin with a consideration of a deterministic nature: Even though each event may have had sufficient reason, it is not always easy to identify what that reason was; therefore, we have here margin to speculate on whether the world is such as to allow free will, or whether the illusion of it results only from our ignorance. But there are more difficulties associated with this aspect, namely that we also have no evidence of the possibility of living in an indeterminate world; therein, the existence of the randomness could provide a good foundation for the possibility of exercising free will. But, unfortunately, we do not know up to now if any of the mentioned alternatives, or a combination of them, will serve as a description of reality. On the other hand, we have the practical certainty that this ability has not been given to us by any god, in order to bestow dignity on our decisions. None of these three hypotheses is evident because reality is always richer and more complex than any explanatory model.

It being so, it is best to start from a model where it is irrelevant whether or not the universe is a deterministic causal system. Daniel Dennett (Dennett 2004) offers us a good alternative; this philosopher defends that our future is closed, we just do not know how. This thesis is of crucial importance for understanding the work we do. The idea that, at every moment, there is only one physically possible consequence for each cause, amply supports a structured notion of a predictable universe, which can be mimicked by a machine. In this scenario, free will probably emerges from the interaction between the various items that constitute a context. That is, we have to look at determinism with other eyes in order to try to understand how it is compatible with free will itself.

The general idea is that all complex processes have a simple structure. Think of DNA strings, of neurons that only have two states (active/inactive), of sequences of zeros and ones that allow us to represent all the symbols in a computer. Now, from this organized simplicity, and, as Dennett states, a determinist one (in the sense that to a certain state in the universe only the subsequent state can necessarily follow), results all the complexity of our world.

The whole evolutionary process can be seen as a kaleidoscope of mutations, associated with the memory of this process, which registers and replicates those that are successful. In fact, the entire evolutionary process can be traced as a selection of well-adapted algorithms. Recall that in mathematics and computer science an algorithm [word derived from the Arabic mathematician Mohammed ibn-Musa al-Khwarizmi, who lived in the eighth and ninth centuries[1]] is, *grosso modo*, a way of proceeding towards an objective, and consists of a systematic sequence of application of rules (or instructions) of a well-defined set, to a given initial state. It results in a sequence of steps to solve a problem, or simply to make an initial state evolve into successive configurations. It is a safe and rigorous method to achieve the desired result. Now, anyone who reads a magazine like *Scientific American* and sees those biology articles in which, for example, female lionesses engender a hunting strategy,

[1]Mohammed ibn-Musa al-Khwarizmi, https://en.wikipedia.org/wiki/Algorithm.

realizes that they talk essentially about algorithms that are under way. Molecules are also running algorithms. What matters is the organizational scheme and the actions that are taking place, that can be enumerated in abstract.

Life can be described abstractly, as intelligence can be described in the abstract. Suppose that, at a given evolutionary moment, there are groups of animals of the same species with very different defence strategies. One group chooses to always attack, another group chooses to always flee, yet another group balances fear and courage and seeks to select, in each case, the strategy that best suits them (defend or attack). There will certainly be a strategy that is best suited to this species, and over time the individuals that apply it will be those that reproduce the most, teach it to their descendants, thus becoming dominant. We are dealing with three different algorithms, and natural selection will choose the most appropriate one.

There is a very relevant cognitive framework that is reached when agents begin to be able to anticipate the future and generate pre-adaptations. The most sceptical will say that this aptitude is exclusive of human beings, but the data of ethology do not permit confirming this scepticism. For example, certain crows in Japan have learned that the best way to crack nuts is to place them on the road while the traffic lights are in the red for the cars, and then wait for these to crush them when they start moving. Well, the first animal that discovered this strategy must have watched the cars crunch the nuts occasionally and later have learned the rhythm of the traffic lights and used the available energy from the cars to its advantage. Here are a set of steps that require predictability, learning and adaptation. That is, the construction of an algorithm with a high degree of complexity. It is interesting to note that the development of evolutionary psychology, which has been so useful to Artificial Intelligence, can also lead us to a better awareness of the capabilities of animals.

It is in these terms, and with a basis on the evolutionary view, that we can presuppose the possibility of considering a new perspective on autonomy. That is, think of it as an emerging and multifaceted capacity. As evolution has allowed the emergence of organisms with better information-processing capacity, they may have entered, more intentionally, the trial and error game, exploring the consequences of certain actions. And, by this process, selected the most appropriate ones for themselves and for their group.

In humans, this mechanism has been highly complexified with the introduction of various forms of reasoning, including the counterfactual form, where the "games" of taking the place of the other also enter. This results in an intermediate situation governed by non-deterministic rules for the selection of options. Note that saying "non-deterministic" does not necessarily imply free will. There are processes that are in the random domain. Emergence is what occurs when several previous things are joined together, and new entities and new phenomena appear that were not foreseen at the outset.

In Artificial Intelligence programming methods are used that copy the processes inherent in this natural selection—they are called genetic algorithms. A genetic algorithm is inspired by natural genetics and comprises four aspects: (1) the coding of information in cohesive units of digital symbols, the "genes" used to perform calculations; (2) the occasional and autonomous "mutation" of these "genes", altering

some of their symbols, in order to produce variants of that information; (3) the "genetic crossover," sexed or not, that is, the mixing of information from those "genes" in order to produce new "genes", as a potential source of innovation; (4) "natural selection" by test: that is, the elimination of "genes" whose calculation result is not good enough, together with "reproduction" (multiplication by copying) of the "genes" having a good result. Steps 2–4 are successively repeated until a stable result is achieved, one as satisfactory as possible. The genes that produce this result are the solution of the algorithm.

Returning to the fragment of Heraclitus that serves as an epigraph to our Introduction, we could say that the algorithms *hide* equally behind nature and culture. In order to express this idea, in the field of cultural evolution, Richard Dawkins (Dawkins 1976) coined the term "meme." By analogy with the gene, he relied on it to designate units of information, selected as the most apt, thus migrating from brain to brain and from generation to generation. Each meme is a set of coded information, aimed at solving a particular problem. The set of memes that characterize a given historical time, or a given culture, is designated by cognome, by analogy with the genome. There are memes to identify and circumvent physical threats, make friendships, keep partners, but also to structure the steps necessary to build—for example—a pirogue. Memes are subject to evolutionary parallel pressures similar to those of genes.

Evoking an expression of the said Richard Dawkins, free-will and autonomy can be seen as evolutionarily stable strategies, selected by the evolutionary process itself. What is relevant is that, in face of a problem, an agent has the capacity to generate hypotheses and choose the alternative that seems most feasible be means of an internal causality, which acts from the inside out on the external causality and is supported by its experiential memory, previous decisions, and preferences accumulated so far. It is relatively irrelevant that the chosen option is the only one that is effectively selectable, or not, although in general many are possible, and sometimes even randomly so; what matters is the agent's ability to represent itself in action, and to generate and analyse *possible futures* by virtue of their internal models of reality. Now, this whole process is representable in a software and in this sense the answer is *yes*: we can build autonomous machines.

At present, the major obstacles to their dissemination are more of an ethical rather than a technical nature. It is urgent to develop the topics of computational morals and work towards achieving legislation adequate to the presence among us of these artificial autonomous agents.

References

Dawkins, R. (1976). *The selfish gene*. Oxford: Oxford University Press.
Dennett, D. (2004). *Freedom evolves*. London: Penguin Books.

Chapter 5
Cognition in Context: An Evolutionary Logic Impacting Our Individual and Collective Worlds

Abstract Intelligence and cognition are presently viewed as processes of evolution. This finding—as well as the detailed research into the dynamics of these processes—supports the idea that the various evolutionary moments are intertwined in a logic of a successive multi-level emergence.

The history of humanity, with respect to the unfurling of the cognitive evolution associated with the development of language and written record, can be seen as one of a case included in that process.

In recent centuries, thanks to the scientific revolution begun in the Renaissance, Man's place in the universe has been redefined, as well as his relatedness to machines.

In the previous century, Turing initiated a new and significant stage, which will culminate in the fall of the Man/Machine divide. However, this process will not continue spontaneously without risk. On the contrary, it will require critical judgment and vigilance from us. Above all, it calls for much pondering over the various paths to tread, so as not to add larger and more problematic risks to those that are already part of our daily lives. Rather these newly provided instruments should help us deal with our present problems.

One approach to this topic takes us to an evolutionary process able to challenge our more ancient prejudices. When Kant resorts to the expression "rational beings"—rather than human beings—to designate moral agents, he would be thinking primarily of angels and not so much of robots. Yet the present state of art makes us think more of robots, and less of angels.

It should be noted that our speculation about a possible faculty of reasoning as a characteristic of creatures of a spiritual dimension soon leads us to accept that such a faculty may not be exclusive of human beings. But, on the other hand, the existence of thinking machines, and above all autonomous ones, does not inspire us as much tranquillity as that of the thinking angels hypothesis. What has traditionally been associated with the notion of machine is the amplification of force and the increase of speed, and not the development of cognitive functions. Concisely, while angels are perceived as guardians, machines are associated with dangers and threats that we might not properly control.

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_5

On the other hand, we are accustomed to organizing our representation of reality according to classes. In traditional thinking we went from gender to species and "sorted" everything into categories that should express fundamental differences. In this context, each "nature" would take its proper place and, as Plato said, Justice would be that architectural virtue that should govern such a process in human beings, themselves also of various "natures".

This representation of reality implies the existence of ruptures based on essential discriminations, insurmountable amongst themselves. Moreover, this feature is even more pronounced in other geographies. Simply recall the notion of caste in Hindu culture.

The traditional world was hierarchical, organized according to primary characteristics, relatively immutable and purpose-oriented. There was the notion of a beginning by divine creation, and of an end, with its respective Final Judgment, at least in the Abrahamic religions. In this finite conception of the world, Borel's typist monkey would not have the time needed to randomly reproduce the sequence of letters that constitute Shakespeare's Hamlet. This requires an infinite world in which a timeless "monkey" can randomly type until it can reproduce exactly the same sequence of letters, spaces, and punctuation.

Now, all that we have said so far, defies, or seems to confront this common view. On the one hand there is a natural evolution, by selection of the most fit and sexually desirable, which shows, beyond appearance, a sharing of processes that permeates all nature. In the specific case of intelligence, it evolves—also by natural selection—until *homo sapiens sapiens* is reached, which in the biological world represents the highest degree of cognition and self-awareness. (For now, we have left out extra-terrestrials, since we have not found them yet.)

On the other hand, machines, which have emerged in the category of mere instruments, now defy such simplistic designation to propose themselves as conscious and intentional agents. How do we frame these changes? Are we facing a paradigmatic rupture that requires another framework, not interpretable in the light of the still current concepts and theories? Are we sufficiently aware of the path we are treading, so that we can rely on the technologies under development?

The idea that intelligence is based on information-processing capabilities—when associated with a universal notion of computation already at the biological level, within the realm of genetic selection—has the potential to include aspects that have traditionally been perceived as dispersed, and henceforth integrated into a common paradigm. This means that we are, in fact, facing a paradigmatic rupture. The boundaries drawn between the natural and the artificial and between various types of intelligence can be unified under this idea that information processing will generate intelligence and aptness. Given this, the traditional cleavage between natural life and produced artefact may well cease to make sense.

To better understand this concept, and how we have arrived here, we must analyse a sequence of paradigm shifts made by Copernicus, Darwin, Freud, and Turing. Each of them simultaneously entailed, at the same time, a delusion, and a new awareness of the World and of Man. With Copernicus cosmic centrality was lost, but a clearer view of our planet's place in the solar system was gained. With Darwin, humanity lost

its centrality as the apex of the divine creative process, but gained clairvoyance about the slow evolutionary process that is at its origin. These two revolutions founded a new vision of man and the world, devoid of mythical-religious fantasies, compatible with temporal scales inherent in the development of life and intelligence.

In the twentieth century, Freud definitively questions religious anthropology, developing a conception of psyche based on a systemic view of the human being. In this respect, rationality is only one of its instances, and not even the most important. In each of these phases of paradigmatic revolution the traditionalists were stirred, and the holders of the instituted powers that be protested.

The Englishman Alan Turing initiated the most recent revolution, one that has presently manifested much of its potential: that of the end of the hypothesis according to which intelligence is a unique characteristic of biological organisms, and that, indeed, it can at last be implemented on any medium that allows certain computing operations.

We thus reached the time in which not only this separation is questioned, but we also evolved significantly in the domain of the intertwining and symbiosis between natural and artificial, real and virtual. It is this plan that frames the whole range of information technology that manages human support systems, from banking to airport traffic. Also to be considered are the existing physical and cognitive prostheses, or those will exist, in a context in which the idea of hybridization is also making its way in several domains. Truth be told, we cannot foresee the potential endpoint of this whole process, we can only define a context in which the cognitive ecosystem is shared by very different entities: human beings, hybrids and totally artificial agents, whether robots or intelligent software programs.

All this dynamic can only be interpreted within the framework of the Emergence Computing paradigm, which can be viewed according to two great families of approaches: the connectionist, or associative models of neural inspiration; and the evolutionary computing models, genetically and mimetically inspired. What these approaches have in common is that meanings are not given directly by the programmer in the form of symbols, but it is the computer itself that constructs them, and the results that their respective algorithms produce emerge, autonomously and automatically, from the interaction between very simple recombination rules.

Evolutionary genetic computation models (or genetic algorithms) are typically used in solving optimization problems, i.e., when we want to find the "optimal" solution within a space with many alternative solutions. There are other techniques, and more "classic" algorithms, to solve optimization problems; but when there are many criteria, according to which we have to evaluate how "good" a solution is, it may be impractical to find the exact result that guarantees to be the "optimal" one. In these cases, we are content with a solution that is "sufficiently good" and the models of evolutionary computation allow us, without great programming effort, to find such outcomes. These models are inspired by the biological path of species, seeking their motto in Darwin's Theory of Evolution.

But we do not always want to optimize a previously defined function, like in genetic algorithms. The aim instead is to observe the spontaneous evolutionary emergence of properties. Biological evolution is characterized by a set of highly interlaced

processes that produce an extraordinary kind of complex combinatorial innovation, but it is not optimizing a previously defined function. A generic term, often used to describe this vast category of order-generating, loosely predictable and spontaneous processes, commonly referred to as stochastic processes, is that of "emergence".

This term has become a kind of signal to refer to research paradigms sensitive to systemic factors. Complex dynamical systems can spontaneously assume patterns of orderly behaviour—which are not previously imaginable, either from the properties of their component elements, nor from their patterns of interaction. There is unpredictability in self-organizing phenomena—preferably called "evolutionary", as Turing did mathematically did for embryonic development—with considerably diverse and varying levels of complexity.

"Complexity" refers to the study of the emergence of collective properties in systems with many interdependent components. These components may be either macromolecules in a physical or biological context; or people, machines or organizations in a socio-economic context.

What emerges? The answer is not a physically defined thing, but rather something like a form, a pattern, or an operative function—just as in the philosophical concept of functionalism. The concept of emergence is applicable to phenomena in which, in determining the characteristics of the set, the relational properties predominate over the properties of the elements that compose them. Emerging processes arise due to configurations and topologies, not due to constituent properties. This functionalism is, almost by definition, opposed to the notions of essential substances, vital principles and monopolizing *qualia*.

On the other hand, the connectionist or associative examples are mostly statistical models that learn their characteristics by mere observation of data. They are instruments usually used in learning tasks and pattern recognition, which are necessary, for example, to extract meanings embedded in large amounts of data with many variables; but they are also used in tasks of prediction, classification, natural language recognition, robotic control, and many others. Neural Networks (and also Bayesian Networks) are, perhaps, the most well-known types of this genre, and recently some of their new models, in particular *Deep Belief Networks* or *Deep Neural Networks* have produced remarkable results in terms of the simulation of some complex cognitive tasks, including natural language conversation with humans and computational creativity.

However, there are issues here that we consider inescapable and challenging. Given their complexity and relevance, let us address them one at a time. The first concerns the fact that all these processes use highly sophisticated algorithms, competent to intervene in the lives of human beings but not yet capable of incorporating in themselves any kind of critical or reflexive procedure, nor logics of context in which the best decision may not be the most rational. Let us be clear: it is not because the machines acquire partial semantic interpretation capacity that they actually become capable of comprehension. Attitudes such as compassion, empathy and commitment are fundamental in the most qualified decision-making processes. Now, we are giving too much decision power to machines that do not even have those faculties, nor a way of considering them.

Many scientists, aware of the misconceptions that the terms *Artificial Intelligence* may raise in these uses, prefer to use the term *Data Science* to characterize the current state of knowledge provided by the Neural Networks. In fact, the thesis that, with more information we will have more knowledge, is not at all evident. If we have biased data, we may even be contributing to a misrepresentation of the knowledge base that supports the inferences that we intend to make from them. A whole bibliography on algorithmic prejudices begins to emerge, which needs attention. Generally, what is designated by this expression is a bias of the knowledge base that is being collected by the algorithm, which will always tend to reinforce the data that it has already collected, that is, reinforce the misrepresentation. It is also known that algorithms with similar or interconnectable programming languages tend to "collaborate" more with each other, leaving aside those that are not recognized as "peers"; all these aspects, besides skewing the collection of information, show the current technical limitations.

Finally, these patterns-recognition systems presuppose that the future is equal to the past, not embodying either the cognitive processes of revision of basic beliefs or self-correction through critical reasoning, since they do not integrate causal and improvable models of reality. They evince not models of causality, but only of statistical correlation between what is and what is going to be. They are incompetent to model, at a second level, the consequences of the hypothetical "What if one were to act now in another way?", and still less the counterfactual—already at the level above it—"What if at the time one had acted in a different way?" These types of reasoning, hypothetical and counterfactual, omnipresent in humans, are, from an early age, incorporated by children.[1]

On a different stance, there is currently (and in the future) a fierce competition between large AI companies, in order to bring to market—as quickly as possible—the most profitable applications, the most robust solutions and the best agents with cognitive competence. This competition, which should favour quality, is playing an extremely harmful role in what refers to the properties of AI solutions available in the market, by virtue of the urgency to be the first. No effort is made to ensure reliability, nor to undertake necessary safety tests, because they are costly and time-consuming, nor are the consequences of their application sufficiently pondered upon, and, above all, nor is time taken to duly contextualize these technological advances and their social impacts.

One of the areas where research is a priority is the clarification of international security standards for the AI software that is being produced. This means that companies should share much more information. But to force them to do so is the responsibility of States, by producing legislation and to establish agreements that promote these aspects. If necessary, delimit access to funds by firms that do not comply with them. Otherwise, we will have an opportunistic competition, which will only benefit companies that find solutions more quickly, even at the risk of downgrading security because it delays production and is more costly. Instead of turning competition into a healthy phenomenon, this aspect introduces a logic of "every man for himself" that

[1] Argument very well discussed in the book Pearl and MacKensey (2018).

will turn against the citizens themselves. In the limit, the main beneficiaries of these new and profitable technologies will be the shareholders and the operational leaders of the companies; not society in general, not to mention the necessary security guarantees. This is extremely unfair, since having reached a certain level of knowledge was a collective undertaking: it involved families who invested in their descendants, educational institutions that formed them and research centres—often public—that sponsored the projects with funds collected from all taxpayers through taxes. In this sense, the benefits of new technologies should be shared by the society at large, but this is not the case. With its impact, AI is destined to acutely increase the problem of wealth redistribution.

It is too easily forgotten that this impact of AI on human societies is incomparable with any other technology introduced in the past. Despite the industrial bustle, many scientists have drawn attention to this problem. However, such alerts run the risk of becoming irrelevant.

We do not learn from History, or we have learned very little. Einstein, and the team that developed the atomic bomb, warned of the dangers of using nuclear energy. Nevertheless, the motivations of war and the urgency to get more and cheaper energy made these appeals irrelevant. At the expense of that, two atomic bombs were detonated over Japanese cities, and countless experiments were carried out with similar explosives, initially without any regard to the risks of radiation exposure or its short and long-term consequences. We also built enough nuclear arsenals to destroy our planet several times over. Additionally, we are now sitting on a gigantic problem of nuclear pollution and of the risk of accidents in infrastructures, with potentially irreversible consequences for the whole planet. Fukushima and Chernobyl are among the most emblematic examples. But there are nuclear waste depots scattered all over the world—from the bottom of the North Sea, to the shores of Somalia—that are real time bombs no one really knows how to deal with. Politicians, and many companies, are always selling us the idea that all is well and that all human processes are under control. Actually, it's not quite like that—or rather, it's nothing like that. The array of global problems we are facing, related to climate change, chemical and nuclear pollution, declining biodiversity, over-exploitation of natural resources, and the concomitant excess of waste—both on land and in the oceans—is worrying and appalling. So sinister that, even before any serious AI-derived threat materializes, we can bog down because of one of those. But AI may go a similar way with regard to its social impact, and the safety hazards it entails, if it is not timely examined and well managed.

Reference

Pearl, J., & MacKensey, D. (2018). *The book of why: The new science of cause and effect*. New York, NY: Basic Books.

Chapter 6
Being and Appearance, or the Omnipresence of Algorithms

Abstract In classical Philosophy, it was generally assumed that there were unifying essences behind the appearance of diversity. Plato and Aristotle, each in their own way, developed a whole Metaphysics grounded on this idea. A way to update such a conception, with the evolutionary nuances known today, can be established on the notion of the algorithm. All living beings execute algorithms, be it in the definition of an individual organism, their inclusion in a species, or their behavioural realm. The difference between them and cognitive machines lies in the fact that biological algorithms are coded and regulated in DNA sequences, whereas machine algorithms are coded with zeros and ones. The question is: from a strictly functional standpoint, will there be much difference between biological and digital coding? The answer is complex and leads us to a difference between closed algorithms, which enable us to perform predetermined functions, and algorithms with learning and self-modifying capabilities. Finally, we deem it necessary to reflect on the social and cultural consequences of the excessive use of algorithms on human life.

As we get a better understanding of the profound sharing of processes that go through life, and all of cognition, it becomes evident that any living being has some sort of internal processing and expresses outputs in its interaction with the environment. In the most rudimentary cases, as we have already mentioned, individuals that do it do not even have a nervous system. Also, as life differentiated into more evolved expressions, the responses became more complex, requiring a causality internal to the organism, exercising itself, in particular and autonomously, in its memory, which enables it to better organize its portfolio of resources, internal and external. Until recently, notions like 'conscience' and 'intentionality' were reserved for human beings. Nowadays, several researchers in the field of neurosciences and ethology apply them without reservation to certain animals. For example, Damasio (2010) and de Waal (2017) have no doubt in using these terms, exploring their evolutionary dimension. Of course, we will not think that, because they are in their own way conscious, all understandings are on the same level. As long as there are not much more complex machines, or we do not find extra-terrestrials, we will be the only individuals capable of creating a network of symbols that enables us to represent the real, to conjecture alternative scenarios, and to carry out explorations made through

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_6

art and religion. However, we are not alone in performing cognitive tasks with a high degree of complexity. In particular, we are not the only ones using tools, nor are we the only ones using simulation strategies to get the desired results, although we are still the only ones who make tools to make tools, machines that learn and program other machines and which start competing with us.

When talking about complex cognitive tasks, performed indifferently by animals, humans and machines, there occurs a doubt that deserves analysis: what can be common among unicellular organisms, humans and machines, with the ability to make bridges between so differentiated physical "supports"?

As duly analysed before, all conscious and unconscious agents execute algorithms. Will it be very relevant if the coding of this execution is done having as support two acids and two nitrogen bases, such as in DNA; or zeros and ones sequences, as in digital processing? It is not guaranteed that there is an answer to this. Here what matters is to analyse whether there are differences between natural and artificial intelligence, as we conceive of them today, and what they are. First, the brain is susceptible to work regimes that are not like the computer's 'all or nothing'. The brain may be drunk, may be hallucinating, or sleepy, and this will correspond to operational regimes that have a nexus of their own, whereas the computer does not. The computer either works in a single well-featured mode, or does not work at all; that is, the activity it exhibits either does not make sense as to what is intended of it, or it is in full possession of its capabilities. In addition, the brain has great aptitude for parallelism, as is well known, and only recently computers with such large-scale capabilities have begun to be explored. Finally, the brain has the characteristic of being self-programmable, that is, it has a motivational system and a reflective consciousness with the capacity to monitor or intervene in its operation, and even to supply or overcome lower-level nervous mechanisms, which the computer does not yet have.

Put in its simplest terms, Descartes's two criteria for discriminating Man from Machine[1] are that the latter (1) does not have feedback and self-reference mechanisms ("could never modify their parts") and (2) has no generalizing reason ("reason is a universal instrument that can be used in all sorts of situations").

But it is precisely in these two features that today we are no longer able to maintain this dichotomy. The gap between Man's thinking and that of his thinking Machines has been greatly shortened by AI research, with its computer programs that partly understand spoken and written language, which demonstrate new mathematical theorems that make medical diagnoses, play chess, build other programs and are currently at the threshold of powerful introspective capacities, as well as of forming embryos of societies through local and tele-computing networks.

In order to properly frame all of what we have said, we have to explain, in more detail, a paradigm of evolutionary cognition. In fact, Franz de Waal, a primatologist and ethologist to whom we will return apropos research in the field of emerging morality, has devoted his entire life to the study of animal populations. One of the main aims of that research is to clarify the level of intentionality and consciousness

[1] Descartes, René (1637). *Discours de la méthode*.

of many species. Especially relevant to this topic are the investigations of other apes, because they are genetically very close to us. But elephant populations also have much to teach us about emergent cooperative strategies. On the other hand, his studies on inter-species cooperative strategies allow us to envision a much more shared view of cognition than one we would be willing to recognize a priori. Of the various cases reported by de Waal (2017) we point out here two situations with Indian elephants not only able to recognize themselves in the mirror, but that have also developed strategies to deceive their caretakers.

Namely, by introducing grass into the bells around their necks, so that they can—in silence—carry out their wanderings, let us say "in secret." There is also a well-known case that took place in the Okavango delta reserve, Namibia. In the first attempt to reintroduce elephants into the reserve, only young animals were selected. The option proved to be catastrophic, as they failed to form a group, often entering into physical confrontation and destroying the landscape inconsequentially. In short, they behaved in the same way as a group of juvenile sapiens, left by themselves. The process was successful only when a whole family of pachyderms was reintroduced, with "educators and learners" forming a complete community. This is sufficient proof that, in the formation of herds, the social learning component is crucial for individual and collective survival. On the other hand, the Bonobo chimpanzees' divergent strategies of conflict resolution, based on a very expressive and differentiated sexuality, show that the traditionally set distances between human intelligence and the intelligence of other animals are very difficult to identify or substantiate. There is also the case of inter-species collaboration, of which falconry is a good example. Humans always had a great fascination for birds of prey, to the point of becoming collaborators in an ancient "art". Most likely it will have been born in the Asian steppes, over four thousand years ago. Note that a peregrine falcon, for example, has always been able to hunt a bustard. However, this is due to the extraordinary speed of its choppy flight. What happens, and it occurs in the wild, is that the falcon hunts the prey, feeds on it on the spot, but due to the huge difference of body masses, cannot transport it to a safe place. Thus, it is forced to abandon most of its hunting on the spot. Partnership with humans does not deprive these noble birds of their predatory character. It simply engenders a bird-human symbiosis extremely advantageous to both parties. Complementary intelligences and capacities are thus combined to complicate the life of the poor bustards, which are more unprotected. The men roam the steppes, making enough noise to raise the flocks; then, falcon identify the preys much more easily and hunt them. Finally, they share the booty, waiting for the human hunter to feed the bird with as much meat as he can. The human can rest easy because the bustards are large, and the falcon can only eat a small portion.

As we said at the outset, this whole life chain runs algorithms, and they become more complex as we go up in behavioural diversity. What we can verify is that behaviours are all the more successful as they express adaptive capacities to new contexts, and express diversity within the same species. In the case of machines, those with self-programming capability represent thus a qualitative leap over those that are equipped with algorithms that follow a single fixed pattern. It is argued that the limit will be the ability to endow themselves with a conscience. Regarding the

possibility of this happening in the future, Dennett (1998) has drawn an analogy between the difficulty of glimpsing consciousness in a pool of neurons, glia and blood cells, and the fact that it will not be easy to see it in a network of integrated circuits and their silica processors. However, even though, at the moment, computational introspection is far from being attained, in fact, nothing prevents it, by a priori principle, from becoming true in the future. What matters is to be aware of the moment in which we are at present, not generating grandiloquent expectations about "master algorithms", at a stage still so embryonic.

However, faced with the omnipresence of algorithms, one question must be posed: Is AI a more dangerous tool than other man-made tools conceived to dominate the natural world and create an artificial one?

Certainly not. Any artefact—from the hammer to the computer—and any knowledge—from food conservation engineering to sociology, or nuclear physics—are equally dangerous. The threat, to pass on to a social scope, needs institutionalization. It is institutions, these social instruments, that can empower or depower, promote or avoid the risks inherent in all knowledge and all artefacts. AI is no exception.

Consequently, the dangers, which are real, require an enlightened institutional and legal context. But that alone is not enough. This framework needs to be monitored by the public. It is therefore also necessary to increase the awareness of this very public about the use of AI and, above all, about the social awareness of its use by experts and users of computer techniques. These are, after all, in a privileged situation to control the dangers made possible by the malicious or less conscious application of AI The elaboration of a deontological statute of the informatics professional would enable an ample discussion of these questions, and the effective exercise of the social responsibility they must account for.

This commitment refers to, not only the proper use of information technology for accepted social and legal purposes, but also to the use and development of computer techniques that guarantee the reliability of their applications. An unintelligent program is much more dangerous than a smart program. In other words, a society that values intelligence values Artificial Intelligence, values the techniques of writing smarter programs more easily. However, the programmer is often compelled to use only the outdated software that is made available to him, even when it is possible, on the same equipment, to install quality software produced by the progress of computer knowledge.

On the other hand, the integrity of a program does not prove the correctness of the theoretical model that was programmed, much less the correction of its applicability to this or that situation. The numerical infallibility of the computer should not be confused with the infallibility of the model, although there are those who take advantage of this confusion.

But what are some of the risks of AI that can already be listed?

Let us first concentrate on the present ones, for many of the future ones will thus be prevented.

Firstly, the aforementioned danger of taking models for realities, taking the accuracy of the program for the accuracy and applicability of the model. It is all the more serious because the computer amplifies nonsense more than reason. On the other

hand, the result may be impossible to verify. Thus, great critical spirit is required towards models—statistical or otherwise—that are programmed in a computer, being careful to explicitly explain the model and its assumptions before starting to program.

Secondly, there is a lack of legislation on the products of computer work, and on the improper use of information held in computerized form. Thus, it is not possible for the author of a program to copyright it. On the other hand, publicity, with its usual shamelessness and pervasiveness, has already begun to compute with the data of its inquiries, aiming at the best manipulation of public opinion and the promotion of individualized advertising through the post box, email, Web and street vendor. Furthermore, databases with lists of potential forced consumers are already being marketed. For example, the buyer of a chess book, who innocently fills in the form that comes with it, may days later have a chessboard vendor, microcomputers with games, and other paraphernalia at his door.

This section also addresses concerns about the misuse of personal data by governmental, private or other entities, which has already been extensively studied in the relevant literature, and for which many countries have already introduced adequate legislation. The problem of misuse of program results for purposes other than those originally envisaged is not lesser, but is usually concealed.

Thirdly, we should not underestimate the ambitions of bureaucrats; they only dream of bureaucratizing more and more, if possible with the help of a computer. So, while the computer can help to relieve us all of the burden of bureaucracy, and also to free the human being who is forced to work in this activity on a daily basis, there is a very great risk of precisely the opposite happening.

Water, telephone receipts, etc., plus their acronyms and abbreviations, become incomprehensible. They are unnecessary, but are imposed by the programmer in the name of ease of programming and machine efficiency. Forms become block *flowcharts*, difficult to understand and unnatural to fill in. Case analyses never predict them at all, and they restrict individual freedom of choice also in the name of ease of computerization. The objection to resulting data is done by filling out yet another new form, this time for the objection program. Bureaucracy becomes even more rigid. Human contact is lost in the resolution of personal problems, since *counters* are put *off-line*. It's the computer's fault, but you cannot ask it for explanations. Access to archived information is hampered. Prepotency gains magnetic contours.

Fourthly, we must assume that working with computers is more impersonal, and their rhythm excessive. In this case, as in others, instead of alleviating human labour, the achievements of science and technology are, on the contrary, burdening it, since they are institutionalized in terms of monetary profit and not in terms of human benefit to those who use them. So go science and technique. All the more so since the effort and educational cost of the specializations required by the new "advances" fall on the individual.

Fifthly, maintaining technological and knowledge dependency hampers the capacity of initiative of computer education and development, and makes it difficult to organize and plan at national level. The export of software and the decrease of imports are also undermined. On the other hand, I.T. dependency concurs, in an increasing

extent, for an ever-greater dependency of many other scientific and technological areas that no longer dispense with computing means.

Finally, computerization contributes to increasing unemployment if the creation of compensatory I.T. jobs is not provided. This, in turn, requires the encouragement of computer education at secondary and university level, with emphasis on the latter as a first step, since it is necessary to train teachers to ensure the reproduction of teaching.

How do we face these and other dangers?

Before going ahead with possible measures, some basic difficulties have to be considered. Initially, we would like to call for an increase in the awareness of the general public, and in particular of computer users and users of information technology, to support and monitor a better institutional framework for informatics. Can we hope for this awareness? The critical analysis by some Marxist thinkers is well-known, namely Lukàcs, of the structure of consciousness and of cultural creation in modern societies dominated by capitalist economy's goals, or by the so-called scientific purposes of certain socialisms. It is the well-known Theory of Reification. In order to limit ourselves to some central ideas, let us say that the social structure as a whole, the global character of interhuman relations, tend to disappear from the consciousness of individuals, thus considerably reducing the scope in which the activity of synthesis of the human being is capable of, and by creating an individualistic and atomized view of man's relations with other men and with the universe. Reality loses transparency and becomes opaque; man becomes limited and disoriented. The enormous progress of the productive forces, and with them of science and technology, only takes place at the expense of a great narrowing of consciousness. Let us also add that, in order to be assimilated or rejected, the greatly increased amount of information in all fields, and in particular the (dis)information transmitted by the media and (dis)communication bodies, would require a particularly strong synthesis activity. It means that, in addition to all the biased deformation due to state interests and pressure groups, the mass of information itself constitutes an element of disorientation, schizophrenia, and weakening of understanding.

It is intuitive that the technocratic and algorithmic community, which needs more and more qualified experts, will not be able to completely dumbfound its members. Therefore, one of the difficult problems facing the algorithm society, which could be one of the fundamental mainstays of resistance against it, is the difficulty in avoiding the mass production of "illiterate" experts, competent in the field itself, but purely passive and uncritical consumers in relation to the overall process that involves their life and that of their fellowmen. That is, they are like specialized instructions of an algorithm, of a program that transcends them. Such narrowing and fragmentation of consciousness reduces this fundamental dimension of man which is that of envisaging the possible. The vast majority do not live but fulfil pre-established algorithms.

Restrictions on the libido appear, while all the more rational and the more universal they become, to operate on the individual—as objective external laws and as internalized forces. Social authority is absorbed in the form of "consciousness", and also by the unconscious, substituting one's own morality, desires, goals and achievement. In addition, from the daily work, alienation and regimentation overflow into free time.

The basis of control of leisure is achieved by the very duration of the workday, and by its tiring algorithmic routines, which transform idleness into passive relaxation, and in the recreation of energies for work. More recently, mass techniques developed by the entertainment industry directly control leisure time. Algorithmic forces never leave us. The alternatives are fixed at the outset. The individual loses the capacity for spontaneous computation. Against this background, possible measures should not be viewed with too much optimism. Problems do not necessarily have a solution.

The only strategy that is coherently thinkable is therefore incremental, that is, change immediately what is clearly wrong and promote what is clearly lacking, privileging local decisions to the detriment of global economic pseudo-algorithms and their parasites. On the other hand, we can appeal to professionals, researchers and computer science teachers to critically examine the perspective of the production, teaching and computer research that they carry out, abandoning the positions of commercial and scientific neutrality to assume a position of explicit ideological choice, consequent and organized.

References

Damasio, A. (2010). *Self comes to mind: Constructing the conscious brain*. New York, NY: Pantheon Books.

Dennett, D. (1998). *Can machines think?* Brainchildren: Essays on designing minds. Cambridge, MA: The MIT Press.

de Wall, F. (2017). *Are we smart enough to know how smart animals are?* New York, NY: W.W. Norton & Company.

Chapter 7
A Question of Epistemological Nature: Are There Limits to AI Enabled Knowledge?

Abstract Addressing the question of the limits of knowledge implies, first of all, the assumption that human beings have always longed for the Absolute, as expressed in diverse religions. However, when discussing science, problems must be properly characterized. Firstly, the genesis of the endeavour generally known as Modern Science must be contextualized. Subsequently, in today's science of AI, it should be borne in mind that it is an area of knowledge related not only to cognition but also to the creation of sensors, processors and actuators that are critical to the creation of new knowledge. To this end, specific scientific methods need to be made explicit. As for the specific question of limits, it will be argued that the production of knowledge has not proven to be limited a priori. Certainly, such limits remain unknown. Either way, the question of limits is worth exploring. Lastly, the notion of symbiosis plays a key role in the increasingly diversified surrounding cognitive environment.

Considering the present moment, it is time to speculate about the science of AI, its methods of investigation, possible limits and expectations. Considering that we are talking about science, we will first make a prior and generic approach to the topics of knowledge in general, and then focus on the specific epistemological issues of our particular area.

Since humans became aware of the power of the faculty of knowing, questioning its origin, role, value and scope has always been a stimulating task. Going back to basics, if we consider the biblical symbolism associated with the expulsion of Adam and Eve from Paradise and the motives that caused it (eating the forbidden fruit of the tree of knowledge of good and evil[1]), we realize that the fascination and the notion of risks associated with the initiative of knowledge development have always been part of humanity's concerns. In short, what is regarded as our superior faculty was also perceived as the eventual source of our loss. Still, human beings have never forsaken the possibility of developing new techniques based on better understanding. The problems triggered by knowledge are solved with more, not less, knowledge.

For many centuries, the fascination with theoretical knowledge, as well as the developments of various technologies, led us to interpret them as divine gifts. We

[1] Genesis 2–3.

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_7

thus spoke of inspiration, absolute knowledge, contemplation of truth, and so many other expressions that indicated that not only the gods influenced us, but also that it was their prerogative to allow us—or not—the access to absolute knowledge.

With the emergence of modern science, this form of cognitive enterprise gains autonomy and specificity. Galileo assumes that, from the viewpoint of its extension, human understanding can never accompany the divine one; however, considering its intensive dimension, there are specific features where we can know as much as the Divinity itself. For example, a specific law of Physics—and the mathematical calculations associated with it—will be as well perceived by the human as by the divine reason. The question of the limits of instrumental reason, which Kant would later, in the eighteenth century, make so explicit, had already begun to be sketched before. Perhaps German Romanticism was the swan song of this historic cognitive aspiration to the absolute which, with Kant's triumph over Novalis, finds itself remitted—if at all—to the end of a historical process without evident closure. Little by little, the word science, namely when associated with the adjective *experimental*, gained a pertinence and credibility such that makes it, an unavoidable reference when speaking of effective knowledge. And no wonder, with science and technique, we had access to the infinitely small, to the infinitely great, we began a much better understanding of our body, with knowledge entering—finally—into the black box of our consciousness of the world, of ourselves and of the society we live in: the brain as an organ that produces a mind. In a brilliant way, Piaget structures an epistemological question[2] that defines an entire research program: what are and how are developed the cognitive agents—epistemic subjects—which produce the kind of abstract and symbolic knowledge that human beings engender?

In turn, General Epistemology throws into the field of scientists and philosophers the question of the nature, scope, and limits of scientific knowledge in general. However, General Epistemology does not exhaust the universe of issues associated with the scope and limits of a given science. Moreover, in science in the concrete what actually exists are methodological difficulties in finding solutions to problems. This means that, in order to formulate a problem, one uses knowledge from several areas, the same way the investigation of solutions is not subject to a single methodology. It is at the confluence of these aspects that we must speculate about the definition of an object of study for AI, and the explanation of the methods appropriate to its purposes. It is also important to stress that, from the point of view of research, one must have conjectures adequate for the problems to be solved. Thus, either the explanation of issues, the research processes, or the concepts and goals that guide them are conditioned by the areas of study involved and by their own languages. For sure, there is a Regional Epistemology that complements the General Epistemology, resorting to the specific themes of each area of study.

With regard to AI, one more caveat has yet to be made. Its object of study is intelligence and the processes of making it occur on non-human material foundations. Now, intelligence is flexible and capable of reinventing itself. This will raise problems

[2]Epistemology (from the Greek ἐπιστήμη, transliteration *episteme*: right knowledge, science; more λόγος, transliteration *logos*: speech, study), Wikipedia https://en.wikipedia.org/wiki/Epistemology.

regarding the definition of aprioristic limits to this area of knowledge, as mentioned before. In addition, AI is a recent discipline, which develops at the intersection of computer engineering, the cognitive sciences, evolutionary and cognitive psychology, and certain areas of philosophy. Is it not this confluence of knowledge, coupled with the plasticity of one's own intelligence, a problem for the delimitation of its object of study and the establishment of a priori limits? What are the epistemological axes of this area of knowledge?

In order to address with some detail the epistemological issues associated with AI, we will have to consider not only methodological problems inherent in scientific activity, but also other questions related to cognition, both in the human being and in the machine. After those considerations we will then be able to point something out on the limits of AI that, from now on, will be simultaneously the limits of symbolic cognition.

Taking as its starting point the definition of Cybernetics proposed by Wiener (1948)—"The science of control and communication in the animal and the machine"—we will clarify from the outset an axis that has accompanied AI since its beginning. We will also find in this definition what we can consider an embryo of the contents that have been developed so far. First of all, the successive revisions of the concept of machine, to the extent that it covers any organized system, capable of evolving over time. We will return to this issue in the more detailed conceptualization of the notion of machine.

It should also be noted that, like any other science, AI is obliged to objectively explain its processes of obtaining knowledge. At this point, we can consider that the task is quite facilitated. Concerning knowledge in the area of cognition, it is to its advantage to be able to break free of natural language's equivocality and, consequently, its concomitant recourse to symbolic languages in the domain of logics and statistics. These are especially apt to the explication of the syntax of thought, still revealing themselves as powerful programming instruments, as well as a means to avoid the ambiguities inherent in the natural language.

Moreover, given the characteristics of the knowledge produced in the science of A.I, expressed in algorithms based on theory, it is possible to explain the requirements for a relatively comprehensive research method, namely:

- First, it must be a precise and operationally defined process; i.e., the observer must be able to act on the system and/or must know of other external modifications that might influence it, and even imagine counterfactually what would happen if something were different in the algorithm.
- Second, the method selected should be equally applicable to all systems.
- Third, its information obtaining process should be entirely objective.
- Lastly, this information should come only from the system being studied, and any external interferences need to be cautioned, because they are possibly ill-defined.

In the evolutionary dynamics of this area of knowledge there is a turning point that we need to highlight here: it is the emergence of programs capable of modifying themselves. In order to understand and enhance this competence gain, we must consider the behavioural theory of learning, the theory of social learning and Piaget's

contributions through his development model supported in his stages theory. This critical moment concerns the gains of cognitive autonomy and decision-making by machines, giving them the possibility of solving problems that were not initially foreseen. With these capabilities, they need not only to learn, but also experiment through a trial and error process, which is often susceptible of automatic correction (debugging), at least an approximate one (for example in the various types of reinforcement learning, among others).

These aspects challenge us to become better acquainted with our own learning methods in their multiple dimensions, to be able to decompose them into their various phases, so that, in order to program, engineers may have models they can resort to.

However, it is not only important to characterize the process of integrating new data, but also to present a critical perspective on them. Now, we do not know very well how to increase critical thinking in an individual. There is no recipe that tells us: educate yourself in this and that way, and the learner will thereby necessarily develop critical thinking skills. This is just one of the aspects that allows us to understand how we are still at a very embryonic stage when it comes to information processing; in fact, we do not have efficient systems to control how data collection can bias the consequences of a given machine learning process. It should also be noted—and this is a very relevant aspect—that the thesis according to which more data necessarily implies more knowledge is yet to be demonstrated. To the contrary, the data must be accompanied with causality models that organize them, allowing to execute upon them reasonings of hypothetical conjunctures, not observed, and counterfactual reasonings on what would have been had an action been different (which was also not previously observed).

If, in a first phase, we could accurately characterize the limits of AI as being those imposed by its programs, that is, a machine can only do what is part of a delimited and well-characterized set of instructions, nowadays such limits have been totally pulverized by self-learning and self-modification.

Therefore, from the epistemological viewpoint, it is understood that it is not possible to circumscribe the scope and limits of the development of Artificial Intelligence. What we can affirm, with criterion and sustainability, is that these delimitations form a symbiosis with the very limits of our intellect and the capacity to conceive machines endowed with an increasingly sophisticated general intelligence. The ambitious AI project is ultimately based on this generality. Its essential limitation will be at most that of the representability of knowledge by symbolic means (symbols and manipulations of symbols, also expressed by symbols), and therein AI is partner to Mathematics.

It does not seem possible, however, to examine, with externalisable and objective rigour, such a limitation, without in turn using those same symbolic means! There will be no science of the unrepeatable, nor externalisable objective knowledge without a support based on the identity of the symbol. The thesis of the sufficiency of the symbolic representativeness of all knowledge is not easily rebuttable, has not yet been so, and constitutes a fertile challenge for investigation. The difficulty in rebutting this thesis does not mean that there is no non-symbolic knowledge. But if such a delimitation exists, exploration within its limits is nonetheless endless, having barely

begun. The question remains and the epistemological issues surrounding AI can be redirected to the following formulation: what is a symbol, such that intelligence can use it, and what is intelligence, such that it can use a symbol?

The computer makes AI possible, because it is a machine that processes symbols in an automated and efficient way. Such processing may, in theory, be done with paper, pencil and brain; but this would be awkward, and in practice it would not go far. Like the computer, our brain is also—in one of its many functions—a symbol processor; it is justified, therefore, to raise the question as to whether the analogies between computer and brain are proper or improper.

A basic idea for us to delve into this issue is the distinction between software and hardware, which is rich in consequences. Namely, it explains the non-obligatory correspondence between a function and the material support of that function. The hardware at the physical level does not have to be specific to a function performed at a higher level by the software, rather it enables the execution of a variety of these functions. Another consequence of the distinction between hardware and software concerns the level of explanation. A program can be understood, in its function or dysfunction, in terms of its own level of discourse, of its language. Of course, a dysfunction can originate in the supporting hardware, but in this case it manifests itself in a bizarre behaviour of the program, not understandable at its discourse level, and not specific to that program.

Computer science is, by definition, only possible by realizing that software has an independence from hardware. Otherwise one would be studying computer A, machine B, automaton C, or brain D, and not computing in general. Although relatively recent, this notion, which is not obvious, is nowadays implicit and commonly accepted.

However, new notions of computer, or rather, of computing may be discovered. This is possibly equivalent to asking whether an externalisable, observable, repeatable, and objective knowledge is possible, if not fully expressible through discrete symbols organized in a language. In other words, whether a non-symbolic science will be possible, in particular a non-symbolic science of the brain. The computer is a theory automator, but we do not know if there will be non-symbolic hardware, including biological, whose functioning is indescribable in terms of symbols and manipulations over them.

But, till then, the computer can provide models of cognitive competence, regardless of the substrate that enables its manifestation in *performance*. In so doing, it redefines, however, the concept of cognitive machine.

In accepting the two premises: that the brain has mostly a component of symbol processing, and that there is largely a software independence from non-symbolic hardware, that is, that we can discuss the issues of the brain's symbol processing without necessarily calling for the organic operations that support them, then we can find in the computer a new source of metaphors that even reconcile the view of material determinism with the teleological mentalist vision. In fact, the computer came to elucidate the long-lasting mind-body philosophical problem, in all its monistic or dualistic versions, with or without interaction, with or without epiphenomena,

etc., because it reconciles those two visions: each one is, after all, a viewpoint, a description of the same thing.

How, then, does determinism become compatible with teleology, that is, with goal-oriented intentionality?

Imagine a circle and another inside, and that the latter represents a being with intentionality. That this being has a memory and that memory has recorded past events. That these events interact with each other in the memory of the being, and therefore, there is a causality between them. Outside this inner circle is the causality of the outer circle world. However, by virtue of his memory, the intentional being has been able to isolate from the outside a certain causal nexus, and is permeable to the outside only to a certain extent. He himself partly chooses his openness to the outside. Let us say that we have a causal ocean, in the midst of which there is a bubble, more or less isolated from this outer causality, with a whole causal world within itself. Here, causal processes corresponding to the realization of the intentionality of the entity can originate, acting from the inside out. Of course, he is subjected to the external causal bath, not being able to choose exactly the causes to which he is subjected, although he may choose some.

Now, any of the causalities, inner or outer, is deterministic, but the, say, secret character of the inner causality is a source of surprise to the outer causality, because it is a cumulative, historical causality, and therefore unpredictable, if one looks only at current external circumstances and limits. The freedom of the being consists of being dependent on his internal causal nexus and, if possible, independent of the external causal nexus when desired.

In another strand, now considering scientific knowledge, if the computational processing of the human genome has led us to Bio-informatics, then, by analogy, we can affirm that "cognome" will be the basis of a future "Cogno-technology", applicable in any science. In this way, the future of AI is linked to the fact that it is an epistemological instrument, not only for an autonomous agent, but for a symbiotic entity, which will help Humans carry out their own science. And we are not just talking about data mining, pattern recognition, ontology construction, although in these fields we can address more structured aspects of epistemology. We are thinking of what every scientist does, which is to abduce, invent and prognosticate theories, test them, create experiments, draw conclusions to support further observations, and argue those observations and his conjectures with other scientists.

There is a meta-argumentation in progress about what good reasoning is, what conclusions we can draw from a discussion (i.e., a semantics), which is inherent in all scientific activity. The computer will be used more and more as a research helper, not only to automate but also to propose experiments and hypotheses; and in the end, by making our own conceptions of epistemological application repeatable and externalisable, it will also render them more objective.

In order to understand—and hence to perfect—human intelligence, we must be able to express its mode of operation, and the computer is the experimental device to test our own understanding, modelling it in detail as much as we can. Logics, in a broad sense of *logic*, is the natural and shared vehicle to do so in a precise scientific way. And the computer, our prime computing machine par excellence,

is undoubtedly our artificial shared vehicle to objectively demonstrate the value of this understanding. Computational Logic encompasses both, and it symbiotic ally benefits from both.

Truly, cognition capacity is what enables us to anticipate the future, to pre-adapt and imagine scenarios of possible evolutions—of the world and of ourselves as cognitive agents—make choices, use preferences about some hypothetical worlds and their futures, and meta-preferences, i.e. preferences about what preferences to use and how to make them evolve. The activity of prospecting the future is vital and characteristic of our species and its ability to understand the real world and ourselves, living in society, where distributed cognition is the normal and regular way of doing science.

Prospective awareness enables us to pre-adapt to what will happen. For this, the ability to simulate, to imagine "what would happen if", i.e., hypothetical and counterfactual thinking, becomes necessary. Such thinking is indispensable in science, for it gives us rules for predicting and explaining what will or can happen, without which technological advance would not be possible. This is not feasible by merely piggybacking *Data Science* with *Deep Learning*!

Lately, we have been working to automate this capability, implementing programs that can imagine their futures, making informed choices about them, and then modifying themselves—as we saw Turing predict would happen—to enact those choices. We are, therefore, before what can be designated as forebodings of free will, but we call it prospective computing.

Epistemology will ultimately have the ability to be shared, be it with robots, aliens or any other entity that has to perform cognition to continue to exist and program its future in this universe. Creating computers and robots in a context means enacting our own cognitive evolution by other means. With the virtue of engendering self-accelerating, symbiotic, co-evolutionary cycles.

Computerized robots will reify our scientific theories, making them objective, repeatable, and part of an external reality built in common over a unified and multidisciplinary science.

In building such entities, Artificial Intelligence and the Cognitive Sciences provide a tremendous and stimulating step towards the promotion of unity of science through, precisely, the endeavour of this construction.

In these days of discrete time quantization, computational biological processes, and evidence of the ever-expanding universe—with its causality automata and unlimited ribbon—the Turing Machine reigns supreme. Its universal functionalism is what enables the unavoidable gathering of phantoms embodied in the various machines (based on silicon, biological, extra-terrestrial or otherwise) to promote their symbiotic epistemic co-evolution, since they can participate in the same functionalism.

Summing up, then, all that has been said so far, and by way of conclusion, there are no limits to the science of computing and, simultaneously, to the progress of Intelligence; or if there are, they are not conceptually explicable and demonstrated. What we can advance is that symbolic cognition will progress, at least for now,

towards a symbiosis of intelligences. Thus, the general computational functionalist position, first stimulated by Alan Turing, is extremely useful. Turing is truly, and forever, among us.

Reference

Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine*. Cambridge, MA: The MIT Press.

Chapter 8
Breaking Barriers: Symbiotic Processes

Abstract Looking into the nature and evolutionary History of Humanity, we find that symbiotic processes are far from new. Not only are they present in Biology, but also in our relationship with other animals and archaic machines, multipliers of force and speed. Only on the basis of a less than 25,000-year-old illusion do we perceive ourselves as exclusive holders of the top of the knowledge chain; but in this exclusivity we have always been accompanied by projections of transcendent beings, or expectations about extra-terrestrials. Until about 25,000 years ago we shared the planet with other hominids, with whom the Sapiens had close relationships. Neanderthals have disappeared, but not all of their genes. Thus, the concept of symbiosis occupies centre stage in the understanding of evolutionary cognition. This has not been seen as too problematic. However, as AI evolves, this may change. Such scenarios should convene citizens for informed debates on the topics and processes of scientific inquiry. Icarus is an example of the abuse of technologies engendered by Daedalus, and symbolizes the risks we face. But it does not have to be so: the problems faced will be better solved with more, never less, properly conducted science and technology.

The notion of symbiosis draws us back to an old narrative, which may well serve as a motto to another problem we cannot avoid facing. But let us first go to the myth: In Athens, in a remote time, lived Daedalus, an ingenious architect and craftsman whose works were known for their care and rigor. Daedalus received so many orders that he had his hands full. That is why he took his nephew Talus as apprentice; but he, in time, proved to be more skilful and creative than his master. Daedalus, overcome by envy, murdered him. Fleeing from Athens in the company of his son, Icarus, he took refuge in Crete under the protection of King Minos, who was already acquainted with his fame. However, there was a price to pay for the magnanimous attitude of the king: Daedalus could only work for him.

It turns out that the symbolic strength of the bull in Minoan culture (3000 BC-1100 BC) was very strong. Legend has it that Poseidon will have given Minos a fine white bull, but on condition that it be offered to him in sacrifice. However, Minos was so enchanted with the beauty and power of the animal that he did not fulfil the bargain. And it was not only he who let himself be enchanted, worse was the case of

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_8

Queen Pasiphae[1] who became enamoured by it. And in order to consummate the mad desire that her passion fed, the queen could find no other solution than to persuade Daedalus to carve an attractive calf, inside which she could accommodate herself, in a position appropriate to copulation with said bull.

From the union between Pasiphae and the animal the Minotaur[2] was born, a half-breed creature with a bull's head and tail and a human body. A hybridization that will echo the current robots with human form. Minos, humiliated by the woman's betrayal and horrified by the being that resulted from it, forced himself to assume it as his own, and urged Daedalus to build a labyrinth where he would capture the Minotaur. And in good time he seized it, for the creature was of uncontrollable ferocity, revealing in addition a voracious appetite for human flesh.

Under the rule of Crete, Athens was obliged to pay an annual tribute of seven young men and seven virgin girls to feed the Minotaur. It is in this circumstance that Theseus, prince of the submitted city, infiltrates among the unfortunates chosen with the purpose of eliminating the beast. Arrived at Crete, he counted on the help of beautiful princess Ariadne to generate a plan. He entered the labyrinth unravelling a strand of wool that would, according to an algorithm of Daedalus, enabling him to track it back to return. Theseus killed the Minotaur and fled with the princess, causing royal indignation once more.

Minos again blames Daedalus for the incident. As punishment, he transforms him into a prisoner of his own creation, together with his son Icarus. Imagine Daedalus, inside the labyrinth, surmising that he could not escape from it alive, neither by land nor by sea. It was under these conditions that, with his son, he began to gather pieces of wood and feathers of the birds that nested there. With these materials, joined with beeswax, he built two pairs of wings and prepared their escape.

Before they started the flight, Daedalus alerted his son to the fragility of the invention. He would not be able to fly very close to the sea, so as to avoid its moisture ungluing the feathers, nor fly close to the sun, since the heat would also unglue them.[3] But Icarus must have wished for more than his father's invention permitted. Intoxicated with flight, he decides to disobey his father, rising as high as he could, towards the sun. His adventure ended badly, for as he approached the astral king the wax melted, unglued the feathers, and he fell to his death while his horrified father looked on. Daedalus saved himself but could not save his hubristic son. It is impossible to imagine a greater tragedy amidst the immense glory represented by this challenge to the human condition.

In a somewhat immediate reading, Icarus can be seen as a symbolic martyr of the association between artefacts and human beings. But a closer reading will show us other dimensions of analysis. In building a flying artefact, Daedalus challenged the limits of the human. Deploring his imprisonment, he found no alternative but to build a symbiotic artefact, counternatural, which killed his offspring.

---

[1] First references in Virgil, *Eclogues*, VI, 45–60 and Ovid, *Metamorphoses*, XV, 501.

[2] First reference, censored, in Ovid's *Art of Loving* 2:24.

[3] First reference to the construction of the wings also in Ovid, *Metamorphoses*, VIII, 183–235.

The deep meaning of this myth shows us that we have developed symbiotic processes ever since, with tools and means of transportation (less or more sophisticated), but also with horses and other animals that we have tamed. Therefore, the idea of symbiosis, in itself, even in the myth, does not bring any novelty. The true originality lies now in what we propose to accomplish it with the intelligent entities that we have created, and with which we will evolve into a cognitive and prosthetic symbiosis.

Given that in certain features this intelligence surpasses ours, and that it is very promising in others, which perhaps we might not even be able to conjecture, do we not run the risk of annulling ourselves in this process? Are we not on the verge of a fate similar to that of Icarus, falling helplessly into a precipice that can annul us? On the other hand, will the complexity of the problems we are creating not require a powerful evolutionary symbiosis, so that we shall not fall into some abyss?

A good way to deal with this issue is to start from an analysis of the relationship between agents with differentiated cognitive abilities. From then on, it may be possible to understand the fears associated with the development of AI, without shirking the social and economic dimensions of that development.

Animals that have shared with us the adventure of knowledge, and of technological development, as well as other species of hominids, are agents with differentiated capacities from the cognitive point of view. Now, the analysis of our relationship with these species reveals nothing very positive about ourselves. The disappearance of the animist religions, the emergence of agriculture associated with monotheistic religions and to the industrialization of agricultural processes led to a gigantic loss of planetary biodiversity. *Homo Sapiens* and the animals he has selected for food, companionship, and labour are the hegemonic large living beings in all latitudes. In addition, most of the surviving species have achieved survival, fundamentally, under the category of resource that can be used by humans. Moreover, within the human species itself, the last twenty thousand years have served mainly the goal of forming structured elites—initially under the pretext of religions, but successively introducing other restrictions, such as a ruling class called nobility, or, in present days, any multinational—that subjugate and instrumentalise others.

Now, this relationship model, based on the strength of the strongest over the weakest, can justify much of the fear related to the development of AI Even so, although this is sensible, it is not obligatory. We can express very credible reservations about the *modus operandi* of higher intelligences, nevertheless more or less equal to us—in fact, since the Palaeolithic *homo sapiens* is already what it is; however, we cannot reasonably do so in the case of cognitive agents that would be a few steps ahead of us and in rapid evolution. It may well be that their level of kindness, generosity, and compassion is not at all parallel to ours, whatever that is.

Another way of understanding our fear is to consider the way in which information technology is used today. What we can observe is: the proliferation of *Big Data* private companies; the use of technologies based on information technology and artificial intelligence to monitor and control large masses of people; gigantic investments in cyberwarfare, and in air, naval and ground-based automata, these at the service of armed forces. To get a small idea of the numbers involved in such a venture, China alone has more than 200 million surveillance cameras installed, many of them linked

to powerful facial recognition systems and other features. What, in the past, was the nuclear arms race now has its counterpart in cyberwarfare. This will be one of the most obvious signs that we are giving too much power to machines that are not yet intelligent and responsible enough to assume it.

In turn, the idea of end-of-the-world has always been present in the narratives constructed by humans. Now, the essential difference is that, at present, there are means and techniques sufficiently powerful to make this distant fantasy, which has always accompanied us, disturbingly real. In Hollywood's apocalyptic themes, AI plays a fundamental role. The scenario of a hypothetical destruction of the human world is systematically raised. The nefarious extra-terrestrials have gradually been replaced by machines originally produced by humans themselves, but increasingly industrious and powerful. Hal, the computer in the movie *2001—A Space Odyssey*, has had much more capable offspring.

Finally, personalities like Elon Musk, or the late Stephen Hawking, have been warning of the dangers of AI However, because of the great relevance of their authors, and the simplicity that characterizes them, these authoritative arguments end up provoking a very significant echo. Many of these appeals have been made as part of *The Future of Life Institute* activities, which is a voluntary, private, non-profit organization created by internationally renowned scientists such as Max Tegmark—MIT Physicist and Astronomer—and Jaan Tallinn—Researcher in AI and Physicist. In addition to these two, the founding team also includes Meia Chita-Tegmark, Viktoriya Krakovna and Anthony Aguirre. This Institute aims to catalyse and finance investigations and measures to protect life. It also aims to develop optimistic visions of the future, including positive ways for humanity to follow their course, considering the new technologies available and their challenges. At the moment, much of these credible threats are linked to the fact that companies perceive themselves as competing groups. It has not been possible, within cultural outlooks, to de facto create a global perception of humanity, to overcome this idea of unrestricted competition, and to accept that effective collaboration has the potential to guarantee sufficient global gains due to one and all.[4]

However, these apocalyptic scenarios may not materialize. The fear that the world will end may be, essentially, the expression of the fear that *our* world will end. AI also has a great symbolic force in this domain. To mitigate this idea, a general principle must be considered, which for our sake we must needs not lose sight of: The complex problems of today's world will not have simple solutions, nor approaches outside the development of knowledge, of cognition capabilities and of technological progress. Most likely some solutions will only emerge when we can surpass ourselves, without falling into any abyss, like our Icarus.

It is necessary to be objective and to think without prejudice; is it really imperative that someone must have priority over the cognitive ecosystem, or is that just an idea they sell us to better convince ourselves that we have to be governed by who we

[4]Author LMP participates in a project (RFP2-154) funded by this institute, since 2018, which intends to model how to prevent the race of wanting to reach the AI market first (the AI race) thereby leading to the neglect of safety and correct functioning of products.

the powers that be? Or, perhaps, is it so that humans who have a certain pretension to think that they control the future of humanity, when everything points to the contrary? And we're forever making blunders. Nevertheless, we are sold this fiction that someone is in control, just so as to submit to this control and feel safer that way. But it's a false idea. To better understand how inaccurate it is, there is nothing like looking at History and seeing all that humans have already undertaken in the name of any one god, or simply in the name of a football club.

The aforementioned interlacing with animals shows that we have always lived in an extremely complex distributed cognitive system. In addition, in the past, *Homo Sapiens* shared habitats with at least two species whose genes are still present in us: the Neanderthal Man and the Flores Man. The former became extinct only twenty-four thousand years ago, and the latter about fifty thousand years ago. In principle, nothing forces *Homo Sapiens* to occupy the top of the *intelligence chain* alone. In fact, we have always looked for this top outside ourselves: in a God who will redeem the world, or in some extra-terrestrial with much more knowledge and power. Alone, the most certain is that we fall flat in the next corner, because, physiologically, our brains are the same as our ancestors' 200 thousand years ago. That is, we are adapted to the Upper Palaeolithic and therefore the kind of problems we create and face nowadays are too complex for the brain we have. It is true that we invented computers; it is true that the brain is not simply physiology, it is not only genes, it is also memes. But memes are, in essence, a distributed thing. They are units of reproduction of ideas, the components of culture and ideologies. They inhabit us, and we are a mere vehicle for them, just as we are a mere throwaway vehicle for genes. We are a discard package for both. Because of death, humans are equipped with two reproductive systems, the sexual one that guarantees the timelessness of genes, and the educational one that guarantees the dissemination of memes.

The educational system is simply a meme reproduction system, right inside our head. Simply look at the compelling theories of the recently deceased Nobel laureate Gerald Edelman: in our brains there is a Darwinian race for the survival of competing memes, which compete with each other for the attention of the brain's resources to guide action, for the individual can only perform one act at a time. There is a genetic and memetic competition of thoughts for the control of the brain (Edelman and Tononi 2000).

Most likely technological developments will lead us in two complementary paths: on the one hand, General Artificial Intelligence running on silicon or the like; and, on the other, what has been called Augmented Intelligence. Here we enter the realm of interfaces between Man and Machine, which will lead us to a scenario where we cannot properly establish boundaries between one realm and the other.

The idea of symbiosis sounds pretty scary—and probably will be. Still, we must also consider that we have been improving humans since the very beginning of humanity itself. From a cognome standpoint (the memetic counterpart of the genome) there are individuals who have much better and more evolved memes installed in their brains, while others not so much. From a biological point of view, vaccines and antibiotics will be good predecessors of the nanoparticles that are now being so talked about as supports for medical interventions and interfaces between humans

and machines. We have not bothered too much about the "improvements" introduced so far, but that does not mean that we should not maintain a high degree of vigilance. After all, we are in an area where the potential for increasing economic and cultural asymmetries outweighs everything we have so far conjectured.

Despite all the risks we have to face, there are certain questions that we must ask to understand if this path is worth taking. What is really important? To increase the quality and quantity of knowledge, whatever the means (human or otherwise)? Or that humans, as they stand at present, improve their knowledge? That space exploration be done in general? Or that humans themselves, as they are right now, per force do space exploration? To create a fairer society, with better processes of distribution of privileges and assets, with a better capacity to enable the global and harmonious development of each of its individuals? Or that society has to evolve by relying only on the capabilities of humans as they evince now?

Of course, existing and forthcoming symbiotic processes entail risks of various kinds, including ethical hazards. That is why it is becoming increasingly urgent to have critically informed citizens who are not anesthetized with football and soap operas. It would be very important if, in the public sphere, much more were said about science and that possible paths associated with new technologies be seriously discussed. The scenario of a dystopian world, where the levels of exploitation, or even eventual "uselessness" of an overwhelming majority of people, is credible and constitutes too serious a harbinger.

Avoiding these risks by ensuring that we rather move towards a distributed planetary mind that integrates human and nonhuman agents requires critical attitude on the part of Humans. In addition, it implies that citizens have a say inside the power systems.

Today, according to our current political knowledge, this is only true within democratic societies. However, the services provided by companies that work with *Deep Learning* systems over *Big Data* today constitute a serious threat to the *virtuous* functioning of the system. This is due to the knowledge they have about the beliefs of the various auditoriums, being able to produce content appropriate to each group's receptiveness. Here is where the notorious *fake news* industry comes in. The problem does not lie exactly in the manipulation generated by false news—this is part of human history. It lies in the detailed 'knowledge' of the various auditoriums, to the extent that they can systematically skew—and here the word 'systematically' has a great weight—the content available to each group, so that there is no criterion for knowing which news is and is not reliable. The delegation of powers and responsibilities in the field of democracy cannot be confined to political and judicial matters alone; we should be able to rely on the content made available in public sphere. This is happening less and less. The present situation contributes negatively to a *mosaic society* where diverse interests—such as LGBT, feminist, anti-racist movements and so on—have struggled for their presumed group interests, superimposing on the common values that should bind us together. all. The crime of discrimination is, primarily, against the dignity of the person and not against being a woman, black or transsexual. It is this *mosaic society* of identities that is being propitiated, when it should be given second place.

In addition, saying that a system is democratic is not exclusively about the act of voting periodically, or about the existence of a free press. An entire power-sharing regime has been devised—which is crucial for minimizing abuses. The separate existence of the executive, legislative, judicial and corporate domains—among others—is critical to block attempts to manipulate and usurp power. Ensuring the institutionalized maintenance of these authorities is a *sine qua non* condition for their scrutiny and—according to our current political knowledge—so that we can aspire to a fairer society in regard to the distribution of the wealth generated. Dominant liberalism has propagated the idea of 'meritocracy' as a way of justifying the concentration of wealth in small extracts of society. But meritocracy is a fallacious concept. In order for someone to show their merit, it is necessary to have a family where they have been brought up, a society that fits that family, schools, universities, research centres, courts, parliaments, security forces, etc. in short, a myriad of institutions with people, and teachers, generally paid with taxes levied on individuals and businesses. Without an organized society, there is no context for the manifestation of individual merit. Therefore, it is the duty of those with political and economic power to ensure a fair distribution of the wealth generated. In addition to being extremely unfair, the concentration that is taking place today threatens social cohesion and democracy itself. If certain powers—not institutionalised—can undermine the structure of justice, generating impunities; sabotage a healthy information environment through a *fake news* industry; hinder free and healthy competition by tampering with stock exchanges, then what we can expect from the future will be nothing short of disastrous.

For now, we are failing at present to maintain balance and scrutiny on the various powers that be. That is why the risk of very robust technological dictatorships emerging is credible. And AI technology only exacerbates the problem, when it could instead be used to counteract it. There is much talk of Orwell's book—*1984*, which, quite possibly, should be required reading because of its growing relevance.

Reference

Edelman, G., & Tononi, G. (2000). *A universe of consciousness: How matter becomes imagination*. New York, NY: Basic Books.

Chapter 9
A Complex Cognitive Ecosystem: Humans and Beings of a Different Nature

Abstract The co-presence of distinct cognitive agents poses vast challenges that require contextualization. If the concept of Humanity seems to convey us to its unbreakable unity, a finer analysis may provide clues as to how to approach the emerging problems. Indeed, all of humanity does not uniformly follow one single morality, and within moral geographies there exist differentiated decision processes. Religious experience, on the one hand, and post-Renaissance Humanism on the other, present very different views of moral criteria, which often coexist despite their metaphysical contradictions. Given these facts, the presence of artificial and autonomous cognitive agents does not qualitatively alter the problem, it merely makes it more complex. We will analyse these aspects, referring to the special difficulty of the problem. As a consequence of a recent and promising area of research, one scientific approach to morals will be set forth. Indeed, we have enough data to conjecture human morals as cases addressable by evolutionary games theory. In this circumstance, it is possible to imagine games of collective strategies of trial and error, with concomitant selection of the most sustainable and beneficial strategies for all concerned.

The co-presence of human and non-human cognitive agents with the ability to analyse and decide on possible futures can be read as problematic. In addition to the prejudices associated with being dominated by others who are not of our *nature*, dilemmas arise that bring us closer to a more detailed first approach to the issues of computational morality.

In the Western philosophical tradition, the term *nature* has had a polysemic use. It is true that it has always referred to the dynamism already present in the Greek word φύσις [Phýsis]; but, on the other hand, the same term has been used in an almost antagonistic way, evoking universal essences which are, so to speak, the models to which each individual embodiment obeys. This fixist notion has been prominent in Western culture, allowing us to draw a line from Plato's time—in which Physics was composed of only four elements—until our days. Concepts like Human Nature, Natural Rights and others alike have become creed, constituting the basis of many institutions, such as Justice. In this context, the notion of human dignity is understood as a derivative of our nature and, consequently, of this traditional conception. All that we have exposed so far refers to a dynamism that cannot be

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_9

expressed by this fixist approach, and does not seem to support it. Thus, as we introduce this evolutionary concept, grounded on recent knowledge—derived from evolutionary psychology and behavioural genetics—several consequences emerge. The first concerns the impossibility of establishing the evolutionary moment in which practical intelligence arose, or in which intentionality or human consciousness took over. When did hominids begin to have souls? If this determination cannot be made, how can we distinguish the dignity of a chimpanzee from the dignity of a human being? Do we all have the same dignity? By emphasizing the functional dimension and prioritizing cognition as essential to the human being, is there no risk that we lay the foundations for a complete relativization of human dignity? And even in the universe of human beings, will we not inevitably move to discrimination between first and second-class humans? And as for the introduction of *thinking machines*, because of their superior functions, are we not yet moving into a scenario where some machines will have more dignity than certain people? In his work *Foundations of the Metaphysics of Morals*, Kant makes a difference between what has price and what has dignity. He considers that it is the ability to act morally, namely the ability to decide not to use others as means to achieve ends, but to regard each other as an end in itself, which are the guarantors of moral dignity. In fact, one of the Categorical Imperative formulations states what we have just wrote. On these terms, the ability to act morally, and the realisation of that ability, is what gives dignity to agents. Yet, how can machines become moral agents?

This is a very pertinent and complex question, which perhaps will always remain as the background of our reflection, and we will not evade it. Firstly, it is necessary to make explicit a difference between the use of the terms "Morals" and "Ethics". In fact, nothing in the etymology, nor in the History of these words, makes them distinctive, since both refer to the idea of "Customs." However, humans decide, morally, on the basis of two registers, giving rise to the possibility of making a subtle difference between them explicit. On the one hand, there is what is supposed to be the right decision; on the other, all cultures spell out a set of rules and procedures that are presented to people as binding.

On some traditional philosophical accounts, the term "Morals" has been reserved for those systems that spell out norms, obligations, and interdictions. In turn, the term "Ethics" has been reserved for principles and motivations for the Good, which underlie the elaboration of moral precepts. There is thus a priority of Ethics over Morals, since it is the first that provides the criteria according to which the second is made concrete.

But, for now, it is important to properly characterize the difficulties that we need to face immediately, an aspect that will not detract us from the question of human dignity. The basic problem is that since there are diverse regulatory systems classified as moral, we ourselves do not know very well what morality is. There is no universal ethical theory, and what we implement in our practical lives is a combination of morals. Aristotle established an ethics based on personal values; Kant's proposal is structured around the Categorical Imperative; but we also speak of a constructivist deontology, with exceptions and argumentative negotiation; and utilitarianism, structured around the principle of maximum happiness for as many agents

as possible, began to have a very strong relevance in Anglo-Saxon culture, and to be the prevailing moral philosophy in world politics and business ethics today. From this we can conclude that the moral phenomenon has an enormous diversity, a point that raises consensus problems well-known to human beings. Under these terms, it is first of all necessary to differentiate between moral systems perched in the heavens, originating from religious contexts, and the moral systems of a more humanistic matrix, in which the agents' moral decisions require a reasonable justification, which is accepted, or not, by the arguments put forth. Although relevant for understanding rule-based behaviours, the first case has no prominent interest from the standpoint of moral research. We are probably not going to accept that machines argue, as justification for their actions, that they are just abiding by the rules—as a Muslim or a Catholic might do. There is a requisite related to justification which, outside the religious context, is of superior relevance. Hence systems such as Kant's Deontological perspective or Utilitarianism provide a seemingly auspicious basis of work.

Nevertheless, there is yet another, and much more promising approach, which needs to be spelled out clearly. Most likely morality is a case of evolutionary game theory, and it began to be engendered long before the emergence of human rules by Evolution itself.

An illustration of this can be found in recent experiments with capuchin monkeys. These animals are especially interesting because the ratio between their brain weight and their total weight indicates a strong investment in cognitive development as a survival strategy. Like chimpanzees, they resort to the use of instruments and form large groups where their members recognize others and are recognized in turn, probably by smell.

In the study conducted by Franz de Waal and Sarah Brosnan[1] about their social intelligence, the aim was to establish whether their behaviour followed criteria that could at least be equated with morality. For this purpose, females were chosen were and tested in pairs.

The experience consisted of putting them in a position to solve certain problems, followed by a reward. While one monkey was rewarded with cucumber slices, the other was always rewarded with grape berries. It turns out these monkeys like grape berries better than cucumber slices. After a few repetitions of the same procedure, the monkey rewarded with cucumber slices began to refuse the food, even to the point of throwing it at the scientists' faces. Such behaviour proves that these apes are able to monitor their action and that of their partners, to analyse relative situations, and to perceive something similar to what we, humans, mean by justice and injustice. In fact, for equal performances, the monkeys that received the lesser reward *felt* the injustice of the situation, and reacted negatively to it. Who wouldn't? But are we facing a reaction guided by some explicit moral value? Certainly not, as there is no possibility of clarifying the reasons behind the action. But that is not to say that we are not faced with a related behaviour of a moral or norm, suggesting that it appeared with evolution. As for us humans, who are also a gregarious species, that means we need rules to live together.

[1] http://www.emory.edu/LIVING_LINKS/publications/articles/Brosnan_deWaal_2003.pdf.

However, ninety-five percent of our moral decisions come to us by reflex; that is, it is also difficult for us to explain them. Even when people agree on the decision to make, they have difficulty spelling out the reasons. "We understand that's right, that's enough." Only in complicated situations do we think deeply and suppress the first impulses. All these aspects show how problematic it is not to know our morality in sufficient detail. One consequence of this ignorance is the limitation of not being able to consensually program it into other agents. Pondering on this also goes to show that, in order to reach such a level, where we are able to program machines, we still have a long way to go in investigating human morality.

This does not mean that, while we are investigating, we cannot enjoy a number of relevant secondary gains; at the moment we have to deal with immediate problems that require prompt response. However, as we have already warned, we are already giving too much importance and decision power to machines that are, for now, very unintelligent.

A.I. dates from the early 50s. Every now and then, it is criticized for not having, meanwhile, advanced as far as promised, or as much as it was obliged to promise in order to get attention and funds. At this point it enjoys great approval and fame in the *media*, but this is due to the great success of only a very small part of what A.I.'s true ambition is: to know how to create artificial beings, with mental and action capacities as good as ours, that collaborate and evolve with us.

Alarms concerning the danger of the use of technologies should be directed to how society uses them, not to the technologies themselves, and much less to fundamental science—which does not require the specific fulfilment and employment of this or that technology, with this or that purpose. Knowledge itself is never dangerous, on the contrary, it allows us to avoid danger.

The alarm that may exist can only be about the misuse of human opportunists who want to dominate other humans. But this is a political problem, encompassing any and all aspects of civilization.

With regard to A.I., there exists, of course, no reason for alarm about its progress being misused, when after all that would be just opposite of what should and could happen. In any case, when willing, it is always with greater technological advances that the misuse of technological progress is to be avoided!

A.I. poses new problems with regard to non-human decision-making autonomy, particularly with regard to smart weapons. The idea that there should always be a human controlling such decisions (the *human in the loop*) is utopian. It is not possible to control a swarm of drones or attacking speedboats, as it would not be possible to control an ant colony if they were small robots. Emergent group behaviour is beyond individual control.

The good moral behaviour of such agents also requires the development of ethical software by which they are governed. This is an area that is receiving more and more international attention.

But the legislative component, we reiterate, and the consequent legal-ethical framework, are still far behind the existing technology, as they have to address the new question of the degrees of autonomy and responsibility of non-human actors.

And, in saying this, we do not think only of robots with moral demands, but of all kinds of software, namely the financial speculative one.

In another respect, as statistics show, it is humans who implement machine work, taking advantage of it to exploit other humans. As some increase their political power and wealth, bending without shame the rules of Law, it will depend on us that there is no violent revolution in the social contract; supposing, in the meantime, Nature does not fall upon us through climate change. It is unfortunate that technical evolution is not entailing social progress and, in its turn, is severely questioning the environmental sustainability and rights of future generations.

Greed, the race for competitiveness—now also counting on cognitive machines—and consumerism are undesirable targets for a good future of the species. These are the aspects that justify the need for much greater research in the field of moral knowledge. If we want knowledge linked to new technologies not to give rise to a dystopian world, machine morality and social morality must be studied together, working as the basis for legislation appropriate to the new contexts. Ultimately, it will be the question of agents' dignity, precisely as Kant put it, that will be on the table. There is only one difference: cognitive agents will not be exclusively human. There must be ethical codes that allow interaction between machines and machines, and machines will have to be increasingly *human* to mingle with humans themselves.

Chapter 10
About the End of the World, at Least as We Know It

Abstract AI has become the axis of an unprecedented cognitive and technological revolution. If other gains in the past have been perceived as threatening, the current changes justify fears and concerns. It is not only the right to work that is most threatened, but also the power structures that may become reinforced by the holders and manipulators of scientific and technological knowledge. Traditional functions of the State, such as currency issuing, are being challenged by technological multinationals, which aim to create their own virtual currencies. The promised humanization of life, as a result of the allocation of routine work to machines, may be but a mirage. Present and future changes are demanding and challenging; they force us to rethink the distribution of the wealth generated, otherwise the dynamics underlying the concentration of wealth will be further leveraged. With the current Social Contract, all points to increasing concentration of wealth and power via scientific and technological innovation. We must question meritocracy and revise the dominant neoliberalism, on pain of a caste society emerging, in which technology holders will exercise great dominance over the whole of society. In short, issues of social and political ethics must be rethought alongside scientific and technological developments.

There is an ancient story from China in which a man drew water from a well using a simple bucket hanging from a rope, and thus watered his garden, which was not that small. On one occasion he was approached by someone who suggested building a water wheel to do the same job, but with less effort. So, the man raised his eyes from the well and was silent for a few seconds as he pondered. He finally replied: I will never do such a thing, whoever uses a machine becomes dependent on it. And those who are dependent on these artefacts will, sooner or later, have a machine heart. And he continued his work without paying any more attention to the advisor.

In a sense, warnings about the potentially harmful consequences of machine work are present—from very early on—in the founding texts of various cultures. On the other hand, when we read interviews of CEOs of big companies in the field of new technologies, the recurring message is: AI will humanize the work. By taking on much of the routine tasks, it will free humans for more creative activities. In short, each doctor will have more time to contact their patients, since the machines will do all the data analysis work; each teacher will have more time for their students, because

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_10

the routine work of explaining and evaluating student learning about established knowledge may be done by a cognitive robot; and so on. However, there is an aspect that is ignored in this argument. Most humans spend their lives doing routine tasks, either driving a truck, or laying components on assembly lines, or teaching already very standardized memes, or analysing stock market data, or interpreting images of the human body. Hence, with these tasks being allocated to machines, since currently it is no longer possible to employ those apt to work from the more than seven billion humans inhabiting the planet, the scenario will be even worse. That is, work itself will become an increasingly limited right.

We should keep in mind that if we are subjected to the paradigm of maximum optimization, if we wish for an ever-increasing financial and cognitive dynamics, we will come to accept the tools that can support that dynamics. Presently on the table is the cognitive revolution provided by what is called the Internet of Things, a Big Data enterprise that will leverage an extraordinarily gigantic data collection. With this data, and the roboticization of numerous functions, not only will there be greater control over people's lives, but the replacement of much human labour with mechanical one. While it is true that many innovations will bring unequivocal profits, those same benefits will always have a dark side. For example, a systematic collection of our biometric data could anticipate disease detection and allow much more efficient early intervention; but it can also offer information to insurance companies that enable them to manage their portfolio from a much better understanding of risk. Thus, it may happen that for the same insurance, some will have to pay a lot of money, while others will see the value of their policy substantially reduced.

We cannot talk about cognitive robots without mentioning the case of Watson, the IBM system, whose capabilities have made it efficient at winning quizzes on television shows, but which has potentially far more useful performances in medicine, economics and engineering yet to be demonstrated.[1] In short, anyone who can afford to use IBM licenses has a highly skilled tool that could eventually be a critical success factor. That is, you can dig a relevant difference for your competitors who do not access the same resource.

If we look at the historical path of our species, what we can observe shows that the dominant groups have always been minorities, and these minorities have always exploited majorities. Moreover, cohesion within each group has always been correlative with competition and—often—annihilation of rival groups. Therefore, it is thought that—instead of the benevolent conjectures of businessmen in the field becoming true—the exact opposite will happen. Human beings—at least for now—will not be slaves to machines, but to those who produce, control and own the rights over machinery at work, and their hardware and software instruments.

---

[1] See How IBM Watson Overpromised and Underdelivered on AI Health Care—IEEE Spectrum, 2 April 2019. https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care?mkt_tok=eyJpIjoiTmpObE9XWm1aVFU1WVROayIsInQiOiJlSTQyWG51UzdrdmJmNzhFMkcyMkdkcmJwaFFrUDZGemFiZ1FEVWdqd0RmOEgrRUxTOVh2UWU0akNcL0pQVDJpSUJqaEw1dDZ6OGdnaCtQZUdaS3ZZRlBMNjNGSGQ1OXM0empwWmJkZmlCWlVZSWwydXNMYVlBYyszVnlTTGZCVTIifQ%3D%3D

One thing is certain, according to the template being implemented, what is glimpsed is a cognitive and material impoverishment of an overwhelming majority of the human population, associated with more precarious labour relations. On the other hand, the concentration of wealth and knowledge will make it possible for unelected elites to emerge with ever greater power. Is there any way to stop this dynamic?

This is undoubtedly the prevailing dynamic, but the solution will not be to hamper the progress of the science of intelligence nor the technologies that implement it; we have problems too serious to afford the luxury of even attempting to do so. However, there are a number of warnings that we must consider. As a result of machine automation and software, the McKinsey Global Institute (in its December 2017 report, updated in September 2018) predicts that by 2030, between seventy-five to three hundred and seventy-five million people, making up 3–14% of the global workforce, will have to change their profession to keep a full-time job. It also shows that 60% of the current professions have at least 30% of activity susceptible of automation by AI Other studies, such as those of the Pew Research Centre and PricewaterhouseCoopers and OECD, point to comparable numbers.

The most at-risk professions are registry administrators, clerks, finance and accounting on the one hand. Similarly, all jobs involving interaction with clients; for example, hotel receptionists and travel agents, cashiers and other supermarket attendants may be mentioned as examples. There is still a wide range of jobs in predictable environments, such as assembly line operator, dishwashing, food preparation, car driving, or agricultural equipment operation, which are also under severe threat. Lastly, in the areas of education, justice, health and continuing care for the elderly, the entry into the labour market of cognitive machines and auxiliary robots poses challenges so diverse that it is impossible to fully explain their reach. A note about human work: we should keep in mind that it relates not only to each person's potential quality of life, but also to their sense of self-fulfilment, self-esteem and the social cohesion we do not wish threatened. It is no coincidence that, in one way or another, religions and major ideologies, such as Capitalism or Marxism, have always valued these. The end of slavery brought about its association with personal dignity, and the exploitation of people came to be seen as improper.

This is the backdrop against which AI makes strides in automating tasks that we humans prefer to avoid because they are mundane and time-wasting. But in other cases, the same AI may also replace us in other tasks that we find rewarding and constructive. Presently, moulding by video games has as much or more influence than we do on our children's upbringing. There are cognitive machines that are beginning to take very secure steps in diagnosing patients. And creative writing programs that challenge the qualities of many authors. As a result, the responsibilities and consequences of assigning work to AI vary greatly. Some autonomous systems recommend music or movies, others recommend court sentences. Advanced systems will control vehicles in city streets, posing safety and liability issues. Two types of problems then arise: one related to what is left for humans to do, and how wealth will be distributed; the other impacts directly on a moral dilemma.

Let us delve first into the issue of the distribution of labour and wealth. The social changes triggered by a new automation—cognitive software (AI) articulated with sensors and manipulators (Robotics)—require deep reflection on the link between capital and labour. They also require a new model of social contract, addressing the enormous risks of instability and discontent inherent in the inevitable changes. If time is money and resource optimization a priority, then life must also be seen as a kind of noble capital to be amortized. An element of the workforce that has invested time and resources in its preparation, which has devoted its best to a company or institution, cannot be discarded just because an algorithm can more efficiently perform the parameterizable functions of his work. Parties and governments are beginning to study the impact of these new workforces, and the European Union is also developing prospective research in this area. The proper use of technological advance must be legislated, otherwise we will be subject to its misuse. Just as there is a "National Committee for Bioethics", a "National Committee for AI and Robotics Ethics" should be set up in Portugal. Unless we are proactively aware of massive unemployment—which the new jobs will not compensate for—serious problems of sustainability of social functions, particularly pensions, will occur.

Let us not confuse technological progress with the social development that should result from it. As we have said, technological progress cannot result in income for just a few. All of us have contributed for this progress to obtain, through funding for research, educational structures, and the legislative and institutional framework. We all have to benefit from it, in a fair way. Progress must be simultaneously social. If the machines work partly in place of me, then I should need to work less for the same salary. At present, the benefits are making the rich unevenly richer, and this has been going on for decades.

There will not be a single solution to something which is a cluster of obstacles, and not just a single problem; in any case, cognitive algorithms that replace humans must pay the labour taxes that humans used to disburse. To reiterate, substitution must permute all dimensions, since the replaced and unemployed humans were supporting their contemporaries' pensions, and total unemployment is expected to rise as a direct result of such AI and robotic substitutions, according to the studies afore mentioned and others. Taxes should be introduced on robots, and especially on software that substitutes for human cognition. Note that such software is much more multipliable and invasive than robots.

Making way now with the moral viewpoint, in the case of Deep Learning algorithms over Big Data, those used by the sinisterly famous Cambridge Analytica and similarly sinister companies, their statistical methods are not able to explain and argue, provide reasons for cases and specific circumstances for which they present solutions, or provide the means for some strategy. However, in addition to being used to influence election campaign outcomes, they are employed in statistical decisions about individual cases—job applications, medical deliberations, court judgments—without the ability to justify such decisions to those affected by them. Of particular concern are independent machines and software with ethical decision-making—such as autonomous drones, job interviewers, and driverless cars—because explaining, justifying, and blaming are essential conditions for morality, but which they lack.

At present we do not know enough to provide ethical rules and arguable justifications, and, therefore accountable ones. Such difficulties are not reducible to technical problems. Obstacles can't be solved with simply technical solutions—pace technocrats around the world! Rather, we need much more research on wide-ranging research on human morality, with interdisciplinary dimension.

A note is due here about the question of consciousness. Just as we do not know what it is like to experience an echo-localization-based representation of the world—as dolphins or bats do—we will not be able to imagine the "inner life" of an intelligent artificial entity either. Therefore, it will be impossible, nor does it seem necessary, to know what its consciousness will be like or if it will have anything like what we designate by our consciousness. Recalling (Dennett 1998), he devised an analogy that may be a suitable starting point for the problem: Just as it is not evident, unless we experience it, that a set of neurons and glial cells, enclosed within the cranium, produce a conscience; neither will it be evident that such production can emerge from a network of processors. The agent is intended to be capable of analysing possible futures, criteria for deciding which option is the most appropriate and be able to explain its decisions; under these conditions it is likely to be implied. One might argue that an algorithm is not a person to be accountable. We must be very careful with this line of argument. States and companies are not individual people with a biological body either, and that does not mean they are not accountable. Furthermore, we speak of ethically judicious companies, and companies that are not; we also strongly object when the State does not behave like a righteous entity. In both cases, it is not only humans who make decisions on behalf of these collective entities, but also organizational memes that, via their Darwinian competition, can outperform others.

The same way we demand that the State, and Corporations, explain themselves when problems arise from algorithmic decisions, the answers cannot be "we were just following orders," as many humans did in Nuremberg. Orders, even when resulting from programming, must be justifiable and subject to criticism on the part of their performer. The greatest risk is delegating to machines and software decisions that affect rights, freedoms, access to opportunities, and individual and collective security.

We, humans, decide not only on the basis of rational thinking, but also of values, ethics, morality, empathy, and a general sense of right and wrong. It is this intertwining that must be much more understood. And because the problem is complex, that should be started as promptly as possible.

People can be held responsible for their choices, because we assume that they are in possession of the above-mentioned faculties and criteria, and can therefore make decisions. Our knowledge of morals does not yet allow us to endow algorithms with the same faculties. This path will inevitably have to be traversed, but we cannot expect machines to have a complete morality right away. At present, our knowledge of the subject does not permit it yet. We will have to create a moral basis, with a set of general rules, which will then have to be configured with moral precepts specific to each culture. Of course, there must also be the possibility for the robot to review its morals as situations unfold. But we are far from being able to produce this software; moreover, from the legal point of view, we have a long way to go too.

In addition, as we have repeatedly stressed, we cannot think of a morality exclusive to machines, but of a moral for rational agents, capable of arguing the reasons for their decisions. Lastly, individual morality must be articulated with social and political ethics appropriate to the context in which we live. This is the huge task that lies ahead.

If we want to avoid damage and increase the common good, this is the way we must go. How to distribute the benefits? In other words, we inevitably face problems that are not only inherent in machines and software, but also in human beings bent on living in a better society.

Notwithstanding, we must realistically warn that not all evolutions are beneficial to all of humanity. At the moment, the dynamics in place point towards the creation of a caste society in which the robot owners are on top, then the machine administrators, followed by the executives and, finally, the exploited. The latter are already quite robotized, but we are still at a very early stage of the process. More and more people will be transformed into mere commodities, interchangeable and increasingly underpaid.

We are striding towards the "uberization" or "pulverisation" of all professions. In the near future, an architect will be enrolled in a platform of professionals in the field, and receive an email to, for a couple of hours, verify the legal regulatory compliance of a particular software-generated layout, and sign it. And the same will ensue with more and more professions.

Caste stratification, and inequality in wealth, will reach a threshold of resentment leading to the uncontrolled explosion of the social contract, unless we begin to deeply restructure it now.

Nowadays AI—which is a collective endeavour resulting from the efforts of universities, research centres, families that have invested in training scientists—is not at the service of the social whole, as it would be of elementary justice, but at the service of companies that become ever more rich, and of States that through it see their surveillance and control capacity worryingly increased. The decisions that are being taken, and others that should be but are not, may originate a catastrophic scenario very difficult to reverse.

There is no guarantee that future humans will live in a cohesive society, much less in a better society. If we are to contribute to better social and economic balances, we must address the moral and political issues at the heart of debates on AI.

Reference

Dennett, D. (1998). Can machines think? In *Brainchildren: Essays on designing minds*. Cambridge, MA: The MIT Press.

Chapter 11
Cognition with or Without Emotions?

Abstract Since human moral decisions often have an emotional coating—either through empathy or antipathy—it is necessary to address the possibility of developing emotionally motivated machines. This is considered one of the limits of AI when doubting its progress. "Machines will never be emotionally motivated, for they have no endocrine system to endow them with emotions." This thesis systematically ignores the role of emotions in humans, as well as what we should really expect from cognitive machines. First, we will highlight that, from a functional viewpoint, emotions play an anticipatory role, preparing possible responses for an organism. Then we will caution that emotions are not a human particularity, since there are many other species that have them too. Hence, there is nothing to prevent a machine from being able, in advance and using its power to conceive counterfactuals, to conjecture alternative possible answers. In the future we will not have fearful or sad computers, but there will be within them the role that fear or sadness play in our decision-making processes. Even based on what is already achievable today, we will soon have robots capable of interpreting and interacting with human emotions.

To begin to address these issues, since human behaviours are emotionally motivated, we must first consider whether a robot may develop emotional intelligence. At present there is a literary cottage industry, sometimes very fanciful, on the subject, adding some unnecessary difficulties. Let us try to formulate the problem properly: From the outset, the terms 'emotion' and 'machine' seem to be mutually exclusive. To wit, all that is emotional is not machinable and all that is machinable is not emotional. Hence, at least apparently, it makes sense to ask the following questions: Will machines be able to develop emotions? Although we are confronted with a research industry around machines capable of recognizing them, will this mean that—sooner or later—the machine itself will be able to make emotionally motivated decisions?

To better define this issue, we need to focus on the role emotions play in human decision-making processes. They play indeed a key role in all our behaviours, from learning to decision making.

It is common to state that Western philosophical tradition has addressed this theme from the idea that the human spiritual dimension shall be essentially logical and that reason and emotion will, by necessity, have their backs turned on each other. This

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_11

perspective embodies a simplistic way of addressing the problem. St. Augustine's sentence: "My soul feeds only on what brings me joy"[1] may be the motto for another reading of Western thought. Thus, since the Low Middle Ages, it is possible to understand an interweaving between emotions and understanding, which goes far beyond the simplistic views that insist on a split between reason and emotion.

Nowadays this common sense tends to be abandoned, and the claim that humans make emotionally motivated decisions is beginning to be widely accepted by the scientific and philosophical community. Damasio (1994) studied these deliberation processes, having put forward a conjecture known in Psychology and Neuroscience as the somatic markers' hypothesis. This thesis explores the role of emotions and feelings, more specifically of emotional memories, in decision processes. Throughout our life cycle, we keep these emotional memories of events that have meaning to us. The moment when we have to make quick decisions, when it is not possible to weigh pros and cons, our deliberation process is not as random as it seems. That is, in these circumstances, memory activates the recollections that are most similar to what is being lived, plus their respective emotional signalling. This is how we find ourselves moving away from a threatening group, or quickly deciding that we will not purchase some white pants. Our emotional memories move against it, without a verbal argument being made explicit. This feature shows us that feelings and the way we use them are prospective. That is, not only do they enable us to contextualize the present, but they are designed to avoid future problems.

Considering all that was expounded, we assume with some degree of certainty that many of our decisions with moral implications are not made according to rational criteria, immediately supported by arguments (argumentation may arise a posteriori if required), but rather grounded on empathy and emotional involvement associated with the context. Yet, can such complex faculties actually be processed in a machine without an endocrine system, presumably without a life history, without a social context to characterize it and, above all, without a future perspective for itself and its kin?

We face an extremely complex issue largely because of the fantasies and projections that are usually made about the domain of emotions. From many common-sense perspectives, with some pre-Darwinian-inspired simplistic Philosophical Anthropology, and even with some less comprehensive conceptions of science, this domain is perceived as one of the last ramparts of resistance to rationalization and corresponding symbolization. It is viewed by anthropocentrism, along with consciousness, as a human monopoly.

A first, necessarily generalist, answer might appeal to the fact that across the realms of understanding man, society, and the cosmos, there is still much more doubt than certainty. Let mystery lovers rest, because—despite much achievement in rationalizing and explaining already—this territory remains infinitely vast.

[1] St. Augustine, *Confessions*, Book 13, Chap. XXVII.

Cataloguing emotions is a complex challenge, namely because we will find many ways to achieve that aim. The most common organization usually arranges them into three categories: primary emotions such as joy, sadness, fear, anger, disgust, or anger; secondary or social emotions, such as shame, jealousy, guilt and pride; and finally, background emotions, such as well-being or discomfort, calmness or tension. The former are usually universal and therefore recognizable in all cultures of all times. The latter, while social emotions, are seen as modifiable by cultural differences, and this is how the same action can generate pride in one culture, and shame or guilt in another. Finally, background emotions indicate the major personality traits of individuals. In addition, primary emotions not only characterize humans, they are available to other animals with complex nervous systems such as chimpanzees, monkeys, gorillas and orangutans.

That said, it is important to stress right away that, besides being an extremely efficient form of communication, emotions are strategic reactions based on past experiences, with a projection in the future. They are therefore associated with notions of data compilation and response automation. Now, this aspect contributes to a first blurring of the supposed opposition between the machine-like plane and the emotional plane. Actually, collecting data and presenting automatic responses on the basis of that data is a task that cognitive machines can perform very effectively.

The second aspect that we should emphasize, which in some way is already present in the matter, has to do with the fact that emotions and their current balance, whether in humans or in other species, are those which—evolutionarily—were selected for having proved to be the most effective. They are thus the result of an evolutionary History, crossed with ontogenetic development. This process also combines with logic, since we all want our lives to make sense and maintain coherence.

Note that in humans the recognition of patterns assumed as stimuli, and the mobilization of resources for their respective responses, implies neurological and endocrine processing. Emotional responses follow a sequence of well-characterized and standardized steps to achieve process success. It should also be noted that in order to make sadness explicit, the strategy of crying is one of the stages of the response, aiming to communicate to others this emotional state or to strategically influence them. We may cry alone, but this is inefficient. In truth, we can even simulate courage, but actually having courage is very different from its simulation, even in terms of the effective results of each of these alternatives. That is, we can simulate any emotion; whoever represents in the theatre or cinema even does so professionally. But the result of these simulations is never the same as that inherent in authentic situations. Since the body is the stage of our emotions, they are thereby manifested to others, constituting one of the most effective communicational responses.

We will now come up with a thesis for which we shall present due argument: When machines start to become more convivial with us and enter the adaptive game themselves—starting from a basic emotion kit—they will learn to use them to their best advantage. This has not been done yet because machines have been used for other purposes. We add to the thesis the premise that it rests entirely on a functionalist perspective, since cognitive machines—without a supporting physical robot—can also show emotionally motivated behaviours.

Before justifying what we have just presented, it will be relevant to take as reference an example already with some sustainability to be analysed. *Gasparzinho* (*Casper*) is a social robot developed by a team of the *Instituto Superior Técnico* (in Lisbon) under European project Monarch.[2] Its design was the result of a very detailed inquiry into children's expectations of robots, its mobility characteristics are compatible with the human stride, and it is able to interact positively with children. The presence of the robot in the ward has permitted to significantly mitigate the oncological suffering of children confined to a hospital environment. Not only does it help to prevent each of them from isolating themselves with their gadgets, it also acts as a mediator between the various social actors present in that context. This requires that the robot have some ability to recognize emotions, so that it can be a competent agent. On the other hand, its "knowledge base" comes from knowledge in the field of social psychology about what positive reinforcement mechanisms are.

Gasparzinho obviously has no emotions like human beings, for it has no endocrine system, nor is it affected by the maladies that affect humans, among other constraints; but it is not because of this that it will no longer anticipate its future and seek better adaptations; this is what it—too—will learn. This way, it develops an interaction where emotional play has critical relevance: after all, positive and happy children heal faster than if they are in the inverse emotional state.

In the near future, we will not interact with unhappy cars, frightened personal assistants, sad social robots, or assistants in anger-ridden old homes. However, we will certainly have robots capable of recognizing these and other emotions. They will also be able to interact based on them because, from a strictly functionalist perspective, they will be able to process their anticipatory role and use that resource not merely to make but also to influence decisions.

In the social context, humans use emotions to influence decision processes; robots will be able to do exactly the same thing. The same way we have mirror neurons that enable us to simulate, in our brains, the action that another is developing, contributing decisively to potentiate the mimetic and empathic processes that will have been at the basis of the cognitive revolution produced by *homo sapiens* in the last 75,000 years, so will a robot's processor, with a certain measure of autonomy be able to do so. This mirror function will not even require a robotic body. A virtual robot can perform exactly the same function, after all—as we have said—it is fundamentally the case of a communication/action process to better control or influence other agents present in the context.

Pinocchio just wanted to be "a real boy" and the question we can ask from that wish is this: what must the blue fairy give him in order to enable him to fulfil his wish? In the Philosophy of mind, the term *qualia* is used to designate subjective mental experiences: my own experience of yellow, joy, sadness and so on. However, emotions

[2] https://rr.sapo.pt/informacao_detalhe.aspx?fid=1&did=190498.

are not *qualia*, as they are neither ineffable nor untranslatable. To the contrary, from a functional perspective, they can be computable as behavioural responses, expressed in well-characterizable sequences of body states.

Reference

Damasio, A. (1994). *Descartes' error, emotion, reason and the human brain*. New York, NY: Avon Books.

Chapter 12
Is It Possible to Program Artificial Emotions? A Basis for Behaviours with Moral Connotation?

Abstract The fact that machines can recognize emotions, or even be programmed with something functionally similar to an emotion, does not mean that they exhibit moral behaviour. The laws defined by Isaac Asimov are of little use if a machine agent has to make decisions in complex scenarios. It must be borne in mind that morality is primarily a group phenomenon. It serves to regulate the relationship among individuals having different motivations regarding the cohesion and benefit of that group. Concomitantly, it moderates expectations about one another. It is necessary to make sure agents do not hide malevolent purposes, that they are capable of acknowledging errors and to act accordingly. One must begin somewhere, even without presently possessing a detailed knowledge of human morality, to the extent of programming ethical machines in full possession of all the functions of justification and argumentation that underlie decisions. This chapter will discuss the bringing out of a moral lexicon shareable by most cultures. The specific case of guilt and the capacity to recognize it is present in all cultures. It can be computer-simulated and can be a starting point for exploring this field.

Since we have frequently referred to terms like symbiosis, human-machinemachine interaction and cooperation, it is not surprising that we wonder which axes support the group cohesion of agents, be they humans or machines. After all, our digital partners—even if, for the time being, they do not show much introspective sense—are there to share decision-making processes with us.

Isaac Asimov, one of the most remarkable sci-fi writers, endowed his robotic characters with his celebrated Three Laws of Robotics: three red lines of moral inspiration.[1] These lines structure the moral relationship of robots with humans. Under the first law, a robot may never injure a human, or – through inaction – allow it to be harmed. The second states that robots should always obey humans, except in cases when this would conflict with the first law. Lastly, the third law endorses the need for the robot to protect its own existence, except in cases this would involve breaking the first and second laws. Later, most likely already aware of the potential, but credible, risks resulting from the technological explosion associated with the

[1] Asimov (1950).

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_12

implementation of AI, he felt the need to establish a new clause, which he coined the zeroth law. Its content stipulates that a robot may not cause harm to humanity, or by inaction, allow any harm to come to it. Behind this enumeration, we imagine an author sceptical of humanity, and hopeful as to the protective capacity of artificial agents.

The restrictions presented here constitute moral precepts regulating the relationship between machines and humans, in a context in which the human is the owner and master. Because of this, they are of little use for entangled scenarios where we will not even know whether or not decision "x" or "y" was taken by humans. In these cases, we wish for a sharing of moral common sense, a distributed knowledge of rules and precepts, and an identical notion of their application. Above all, we want each decision maker to be able to spell out the arguments that supported their resolutions. In this sense, it is understood that there are unavoidable concerns related to the need to generate trust among agents. We do not yet have much experience of what is a cognitive environment filled with differentiated agent types; but we are well aware of the consequences, and the social respectability of a person who systematically does not keep his promises, or does not adequately justify his deliberations. Social cohesion requires trust and commitment. This means that artificial agents will also have to participate in these dynamics. Even if they come from diverse manufacturers, equipped with algorithms with different acting instructions, they will have to inspire confidence. In this respect, we are very much used to dealing with humans and understanding the difficulties associated with that, the problems relative to the presence of free riders, the capability of dissimulation that enables to conceal real selfishness with feigned altruism, and even the recognition that moral precepts are not coercive. Often our moral conscience allows us to identify what the good deed shall be; however, despite recognizing it as such, we do the opposite. We now have a very clear sense of how these situations break trust, and how difficult it is to regain it.

As far as machine decisions are concerned, the potential for distrust may be further amplified, either by our fears and expectations, or by the inevitable mistakes the new machines will make, which, by norm and custom, are of today being ascribed a demanding technical perfection.

On the fears side, there is not only the possibility of them getting our jobs, or even imposing required behaviours on us. On the expectation side, we do not accept that they make mistakes, at least with the same typology of human error. If an artificial driver kills a human, such event is not comparable to our disabling of a cognitive machine, or to an accident that destroys an intelligent robot. Nevertheless, legislation on artificial entities, based on a judicious autonomous agent conception, will eventually recognize robots as beings capable of moral decision, mitigating part of this difference.

In future, the moral lexicon will be common to both humans and machines, and this includes notions such as right and wrong, worthy and unworthy, cooperative and selfish, acknowledgment of guilt and apology, and so on. This will require machines to be equipped with a consciousness-like device that enables them to produce informed judgments and their associated argumentation. It is in this context that

we may question which elementary items might be shared by machines from different manufacturers, with different programming languages and, eventually, with distinct scopes, not all properly spelled out. In this sense, we may ask ourselves which categories must inevitably be present in AI development programs, and which faculties favouring moral discernment must be integrated into artificial agents.

The need and urgency for machine morality, and its respective research, has been stressed several times. As artificial agents become more sophisticated and autonomous, acting in groups within populations of both other machines and humans, this need will be increasingly justified. To understand this idea, just think about the number and diversity of robots that are being introduced into the market. Zora, a Softbank Robotics[2] creation, is not only able to accompany hospitalized adults and children, but is also an excellent helper in nursing homes, and can also be used in an educational setting to perform various functions traditionally performed by human teachers. Zora interacts with emotionally debilitated people, but it is not yet a robot capable of meeting all the resulting challenges. However, it is these challenges that delimit part of the path to be followed. That is, computational morality entails a whole research program that should be incorporated into all future AI developments.

In a first phase, it is very relevant to identify the lexicon traditionally associated with the moral context. Thus, pairs of terms such as cooperation/competition, guilt/impenitence, shame/honour can be identified as axes of the referred lexicon. Its organization in pairs of opposites expresses the possibility for agents to make choices.

Also, the need has been several times mentioned to start with small steps, which can be consequential. Regarding the study of the cognition of human individuals integrated in multiple agent groups and who frequently interact morally (they may choose to compete or cooperate with one another), it is important to understand whether the results of the research can be equally applicable to the evolution of populations of artificial agents. Or still, if the results obtained in laboratory contexts with populations of artificial agents allow for a better understanding of human morality and its "machinery".

Specifically, in relation to human morality, the answer seems to be a resounding "yes". We must always bear in mind that morality concerns groups and populations, requires cognition, and will have to evolve into an intertwining and strengthening of the relationship between nature, genetics and culture. On the other hand, evolutionary Anthropology, Psychology, and Neurology have produced new insights into the evolution of human morality. Their theories and scientific results must be considered and serve as inspiration when thinking about machine morals. Indeed, the very study of ethics and of the evolution of human morality, can now also draw on experimental means, on the theory of computation, and on robotics to represent and simulate individual—or group—moral reasoning in a multitude of circumstances. Regardless of these occasions, however, morality is to be general, making explicit a set of procedures capable of ensuring that the equal is treated as equal, and the

[2] https://www.softbankrobotics.com.

different as different. That is, it must produce rules and procedures that ensure fair and equitable treatment of agents.

Using these domains, we can better understand the morals that emerge in agent populations in a given context. In addition, human groups tend to rival each other—sometimes too much. Therefore, it is important to research moral items that are beyond cultural differences and are represented in various value systems.

It is in this context that, in the morals of groups, themes such as shame and guilt can be invoked. These are social emotions of great importance. Although both have evolved to promote cooperation, guilt and shame can be dealt with separately. Despite being able to promote acquittals and even spontaneous public confessions, guilt is an internal private phenomenon. An important point, not always properly explored, is that present guilt helps to prevent future guilt, due to the pain it might entail. It is thus a form of prospecting for the future, always very useful in games and for survival. As for shame, it has, inherently, the trait of a public performance that is personally unwanted by the agent, because it addresses his own essence, and not just his act, as is the case of guilt, and leads to shunning the social. As in the case of guilt, it can also lead to similar consequences: excuses, apology and change of behaviour. Shame, however, depends on the agent being caught, on not misleading deliberately, and on the existence of a mechanism of social reputation.

No other emotion is more directly associated with morality than guilt. If we consider the Catholic religion in terms of game theory, we know that we are born losing, with original sin. It was at its expense that Christ suffered and died to save us, which makes us doubly guilty if we choose not to play properly. Associated with guilt comes confession, the request for absolution and its respective pardon. This provides the opportunity for a reset; the game being playable over and over again. The notion of guilt is closely associated with the idea of conscience as an internal guide that tells us when an action is wrong. Furthermore, guilt is widely regarded as a fundamentally collective emotion that plays a positive prosocial role. It arises especially when there is a threat of separation or exclusion. Guilt is an unpleasant emotion, and when experienced, people try to free themselves of it: the most common coping strategies are confession, amends, self-criticism, and punishment, often self-inflicted.

We must bear in mind that, in terms of real material gain, non-compliance is the dominant strategy in many economic games: defaulters do better than cooperators, whether or not their trading partners are cooperating. This makes it rational for both parties to disobey, even if the results of mutual misconduct are worse than those of mutual cooperation. In order to mitigate this embarrassment, many Evolutionary Game Theory (EGT) theorists have argued that guilt is not only anticipatory, but mitigates this problem by promoting a cooperative attitude, adding an emotional cost to failure. Trivers (2011) speculates that coevolution has caused guilt to arise because it makes fault less attractive. People can earn materially through a defaulting strategy, but guilt makes them suffer emotionally, and this in turn leads them to cooperate. Robert Frank notes that this trend is so deeply rooted that people avoid failure, even in cases when the other party will not be a likely partner in future exchanges. Both Robert Trivers and Robert. Frank assume that the emergence of guilt is the result of evolution, selected as if it were a process of biological evolution. It is possible

that guilt came under cultural pressure as a tool to ensure that people cooperate. It is interesting to note that, alongside this moral development, nature—and later culture—has engendered a kind of arms race for the development of camouflage strategies and generation of deceptions, which has culminated in the human capacity to lie. Most likely, part of the development of the large brain that characterizes us concerns the enduring need to decode strategies that imply deception.

The evolutionary question about guilt is whether guilt is worth more than its absence with respect to the advantageous emergence of cooperation. In order to research the problem, we can explicitly introduce guilt into computational models, thereby simulating its appearance in the evolutionary scene. We can test this hypothesis through such computational models. In them, guilt is linked to the fear of recognition of malevolent intent, since infraction avoidance will have evolved from the fear of detection of the harm caused by it. The prediction of these models is that guilt will facilitate, and accelerate, the emergence of cooperation. Although its initial cost is heavier, over time the cost will eventually be recovered within the guilt-ridden population by inhibiting faulty strategies. The basic hypothesis is that, with some guilt present in the initial population, the most frequently evolutionary stable situation includes guilt and better cooperation.

In a community where there is a small number of individuals capable of feeling guilt, the models show that behaviours that derive from this variable will steadily increase their replication across the population: more than if there were no guilt. This proves two things: first, that machines must feel guilt and—second—that they must be programmed with such a capability. It also explains why the feeling of guilt in humans corresponds to a beneficial evolution/mutation in their cooperation games.

How is it implemented? Guilt is a kind of self-punishment: if I betray the other who has cooperated, I imagine a counterfactual. I.e., if some other play of mine would have been preferable, knowing already how the other would play, then I deduce something from what I won in the play, and change my behaviour. This subtracting something is the self-punishment. Our own research using EGT has computationally demonstrated the cooperative advantages of this guilt model (Pereira et al. 2017).

Under these terms, it is essential to provide cognitive machines with the capacity to recognize guilt and corresponding apologies. Guilt not only mends present situations but prevents future evils. The need to achieve this is evident as it will be its execution. Furthermore, guilt, as a structuring emotion for cohesion and confidence building among group members, is fundamental to fostering the cooperative relationships which must be our target if we are to live in a better society.

References

Asimov, I. (1950). Runaround. I, Robot (The Isaac Asimov Collection ed.). New York, NY: Doubleday.

Pereira, L. M., Lenaerts, T., Martinez-Vaquero, L. A., & Han, T. A. (2017). Social manifestation of guilt leads to stable cooperation in multi-agent systems. In S. Das et al. (Eds.), *Proceedings of the 16th conference on autonomous agents and multiagent systems* (pp. 1422–1430). May 8–12, São Paulo, Brasil.

Trivers, R. (2011). *Deceipt and self-deception: Fooling yourself the better to fool others*. London: Allen Lane.

Chapter 13
After All… What is a Machine?

Abstract New computer technologies, namely work facilitating software, require a reassessment of the concept of machine. If until the mid-twentieth century a machine was something which essentially served to multiply the force or speed imparted to a particular task, from then on what became meant by "machine" changed radically. In the early days of the computer revolution the machine was still the computer itself. In other words, the notion of machine, already free from the industrial image of force or speed multiplier, still remained linked to the notion of a physical mechanism. In today's algorithmic society the machine, free from the physical world, may merely be a program on the Internet, now in the process of permanently disengaging from its routine and predictable character. The machines of the future will reinvent themselves according to their needs. In short, they have come a long way, and will carry on towards a symbiotic intertwining with humans.

The concept of symbiosis has come to occupy an increasingly central place in the controversy, offering various clues about what will be the evolution of the present knowledge ecosystem, in which humans and cognitive machines interact. Yet, so far, we have not yet addressed in detail the concept of machine. Machines have themselves undergone an evolutionary process of abstraction. In a sense we can even say that such a dynamic seeks their progressive humanization. In this process, computer and computational theory have played a key role.

This situation could not have been foreseen in the dominant classical western philosophical tradition, since in Greco-Latin Culture everything related to the production of instruments occupied a lesser position. Plato might have felt exasperated with Eudoxus for his attempt to solve certain geometry problems using measuring instruments. For the founder of the Academy, the nobility and dignity of true knowledge lay in the fact that it could be developed by reasoning and, later, by contemplation of pure essences. This superiority of contemplative knowledge over practical knowledge, linked to the production of artefacts, survived in western culture at least until the scientific revolution started in the sixteenth century, especially with the experimentalism practiced by Galileo Galilei.

Interestingly, during the following centuries there was a significant revision of this model, in which the list of actors in the development of Modern Science became

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_13

widely heterogeneous. Maths teachers, optical specialists, doctors, navigators, watchmakers and makers of other instruments took part in it. It should be noted that each of these actors sometimes had highly dubious interests. We should not forget that Newton had a true passion for Alchemy, devoting much of his precious time to it. But here, what matters is the assimilation of *knowing-how* by the new ways of producing science. It is in this context that emerge the inception of Man as machine, or of God as the universal watchmaker, noting that, in each epoch, human beings compare themselves to the best of what they can realize.

With the Industrial Revolution machines definitely gained a place in human life and societies, so relevant that, in principle, they became indispensable. Better machines allow for better knowledge and mastery of the natural and human worlds, and this superior knowledge has enabled the design and construction of better machines. Hence the question arises: what is, after all, a machine?

Until recently the machine was seen as a device for converting a given energy into another, better directed towards desired human purposes. The concept also denoted a mechanism driven to transform the shape, properties, state, and position of a given raw material. Minding the simplicity of the machines that Man began devising, we should also consider the initial approximation between the notion of machine and that of tool. Pulleys and levers are good examples of these approaches.

Since there are other species that also use tools, it is not possible to think of Homo Sapiens, in any state of his successive development, without imagining him a user of machinery. At the dawn of his presence on the Planet he must have resorted to very rudimentary tools. However, he kept expanding his powers, supported by the improvement of those artefacts. This evolution is not worth detailing here, but engineering has been present in our lives since ancient times, of which Dolmens and their configurations, often used for astral predictions, can be taken as expressions of the most archaic ones.

Obviously, when we talk about conjectures over machines, we cannot but mention the genius of Leonardo da Vinci. It is often said that he was a man far ahead of his time; but to be entirely fair, we must consider Leonardo such a timeless creator as to be beyond his time. His machines—imagined and materialized—express human greatness in all their splendour.

If the Industrial Revolution exponentially amplified the presence of machines in human life, that was because they incorporated the possibility of accomplishing especially complex tasks like flying, speeding, moulding iron and steel for various purposes, or producing cloth through assembled industrial processes—after all, all that has enabled and justified the current framework of international trade. This rapid and structured mobility, coupled with the mechanization of manufacturing, has granted machines a unique place, both in our society and in the dynamics of knowledge.

Think not, however, that because there was—in the dominant Philosophy—a habit of dissociation between the knowledge that led to the production of artefacts and theoretical knowledge, this philosophising was the only way to address the issue.

By way of example, in Jewish culture throughout the European continent, the mythology associated with the figure of the Golems has always been present. Briefly,

a Golem is a clay-shaped figure in which life can be infused through formulas of a mystical nature. The Torah sages, focused on an everlasting attempt to elucidate the darkest contents of Old Testament texts, have always wondered how God had blown life into the clay with which He shaped Adam. This speculation has enriched fields like that of literature. Amongst the varied production in this field, we highlight all those associated with the Prague Golem. During the sixteenth century, the life of the Jews of this city would not have been pleasant; in fact, the community fell victim to various persecutions, which often led to widespread slayings. It is within this context that emerges the simultaneously historical and legendary figure of Yehuda Leway ben Betzalel, also known as Rabbi Loeb; a wise man, master of the Torah, the Kabbalah and the Talmud. He interacted with astronomers and mathematicians as remarkable as Tycho Brahe and Kepler. By 1520, facing systematic threats to his community, orchestrated by Bishop Tadeusz—a Jew converted to Christianity—he created a Golem. The creature would have been moulded in clay and subsequently breathed alive with the power of the divine word. The Golem would have "lived" ten years, defending the community and diligently obeying Rabbi Loeb's command. The literature around this figure is rich and fruitful. Perhaps its most famous sequel is the novel *Frankenstein*, by Mary Shelley. However, even in today's virtual games, it remains present and inspiring.

The creation and development of informatics has provided the means to materialize the interplay between people and machines. It is this intertwining that makes us review the concept of machine, and will certainly have a direct impact on what the human beings of the future will become.

A machine can be a program, or an algorithm installed on a computer, or hosted on the Internet, thus being able to work on several computers simultaneously. These new machines are devoid of physical dimension, do not transform energy, nor multiply forces, but process information at a pace and complexity that outruns the human brain. They are excellent partners for doing routine work that requires a huge memory and concentration. Therefore, from a functional point of view, they are already in symbiosis with us. We will see what lies ahead, for the domain of augmented intelligence inevitably intersects with that of artificial intelligence. Very soon, we may have technologies that, rather than answering the question: 'What is man?', will enable us to ask: 'How do we want to evolve?'.

The Man/machine fusion will be one of the most striking ethical, anthropological, emotional and cognitive challenges of the cognitive revolution. It promises an increase in overall capabilities with an unpredictable dimension. But it can also result in the most asymmetrical and hierarchical society that has ever existed on the face of the Earth. Holders of power will also succeed in mastering knowledge, technique, time, and health. Nietzsche stated that Man is to the superman (which he announced), as a monkey is to us. In the second half of the nineteenth century it was not foreseeable that a hundred and some years later this speculation would make so much sense and become technically achievable. But it is so, and we have to face this challenge. This will entail a reconfiguration of the relations of production and the notion of labour, but above all a reconfiguration of Man's place in the various domains of intervention.

Since the beginning of the scientific revolution, we have parted with our cosmic centrality through the fall of geocentrism; we have lost our status as a corollary of divine creation through evolutionism, and we have lost the rationality and unity of consciousness—while inherent characteristics of our species—through psychoanalysis with its emphasis on the unconscious. We are now preparing a fourth and challenging revolution, consisting of the fusion Man/Machine. Now, when we speak of symbiosis this is what we are referring to. We are now at a stage where it is becoming irrelevant that certain knowledge has been produced by biological agents, or artificial ones. The real decision-makers on stock the exchange are increasingly algorithms. Over sixty percent of the purchasing and selling decisions made on the New York Stock Exchange are thought to be the responsibility of cognitive machines.

In the field of symbiotic procedures, children and young people take the lead in the process. This is justified by the fact that they were born in the digital age. The scope in complexity of the problem, addressed by Luís Moniz Pereira in an article titled *Cyberculture*, *Syncretism and Symbiosis* (Pereira, 2017, justifies the inclusion of this theme in this book. As a preliminary note, it should be borne in mind that the best schools in Silicon Valley work without tablets or computers. To wit, whoever knows best and devises new technologies keeps their children away from them at school…

Reference

Pereira, L. M. (2017). Cibercultura, Simbiose e Sincretismo. In: H. Pires, M. Curado, F. Ribeiro, P. Andrade (Eds.), Ciber-Cultura: Circum-navegações em Redes Transculturais de Conhecimento, Arquivos e Pensamento (pp. 45–55). Braga: Edições Húmus.

Chapter 14
Cognitive Prerequisites: The Special Case of Counterfactual Reasoning

Abstract When speaking of moral conscience, we are referring to a function of recognizing appropriate or condemnable action, and the possibility of choice between them. In fact, it would make no sense to talk about morals or ethics, if for each situation we had only one possible answer. Morality is justified because the agent can choose among possible actions. His ability to construct possible causal sequences enables him to devise alternatives in which choosing one implies setting aside the other. This typology of internal deliberation requires certain cognitive capacities, namely that of constructing counterfactual arguments. These serve not only to analyse possible futures, being prospective, but also to analyse past situations, by imagining the gains or losses resulting from imagining alternatives to the action actually carried out. Compared to social learning, where the subject can only mimic certain behaviours, the construction of counterfactuals is much richer and more fruitful. Thus, for machines to be equipped with effective moral capacity, it is necessary to equip them with the ability to construct and analyse counterfactual situations.

Living in a better society first requires conjecturing what that better society might be. Now, this task is not at all easy. Throughout history, human beings have always been imagining utopias. When we think of Plato's ideal *Republic*, or St. Augustine's *City of God*, or Thomas Moro's *Utopia*, or Karl Marx's *Classless Society*, we are always a long way from concrete societies. Throughout our History we have inhabited the *world-as-it-is*, but imagining alternatives that would make it better. This dialectic game between the descriptive domain and the prescriptive realm has been extremely rich and fruitful. Of course, we have never achieved any utopia so far; moreover, we are not sure whether, had we done so, it would have been good for humanity. Still, for better or worse, utopias have played a key role in our individual and collective decisions.

From a collective standpoint, they have provided an elicitation model for what we imagine the ideal destination to be. We are used to thinking that having a destination, or a comprehensive purpose, is extremely positive. However, this goal has also given rise to much violence between groups with opposing interests. Suffice to think of the various Proletarian Dictatorships that have proliferated across this planet, and how, under the possible pretext of creating an egalitarian and just society, they have

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_14

sanctioned acts of extreme violence, with massive killings of human beings. On the other hand, without a range of possible utopias, we would be relatively lost, because we would not have enough diversity in the answer to the collective question of where we wish to go. We need this diversity not to become dependent on just one possibility. Imagine a single answer—religious in nature, say—to this question. It will not be accepted by all believers, let alone by non-believers.

Even without reaching a consensus on what an ideal society is, and accepting the idea that multiple conjectures about it can coexist, we will unreservedly agree that human societies should not be used as a pretext for the enrichment of a meagre ten percent of the world's population. Nor is it likely that consuming all that each one would give credible meaning to our individual and collective lives. However, this is what we are witnessing more and more. This means that we are treading dangerous paths, both in the field of our capacities for idealization (or lack thereof), and in the realm of what—concretely—we are doing to try and improve the present.

Reflecting on these issues requires the exercise of critical thinking, a capacity we acknowledge to be rare. Indeed, the data from Social Psychology are quite emblematic in this field; we know—from Salomon Asch's experiments—that the percentage of conformists in a given population is much higher than the percentage of non-conformists. We also know—at least since Milgram (1974) experiments—that the tendency toward obedience to an authoritative-looking figure is very strong amongst humans. If the order giver is credible, if he maintains a close relationship with the order follower, the latter will do practically anything he is ordered to do, without resisting. In this context, that we must raise the issue of critical thinking and the conception of alternative worlds. Expecting everyone to be nonconformist, critical and informed will imply confidence in a highly unlikely social change, with consequences very difficult to predict.

On the other hand, in the domain of individual morality, one of the structuring requirements to be able to affirm that a certain act is moral consists in the possibility of the same not being enacted. Duty is not about a constraining obligation. Even knowing what good is, as Saint Paul acknowledged, we can do evil: it is in this tension that the dignity of all acts is founded. To the extent that, even in Christian theology, the problem of free will finds an answer compatible with the question of evil. That is, God allows it in the name of a greater good, which is freedom. If we were left with only one possible option, there would be no dignity in choosing it. In the realm of emotions as well, the imagination of alternative scenarios occupies a prominent place. Consider the situation of Camus's character in *The Stranger*: if it had not been so hot, if there had not been the resulting despair, would he have killed the Arab? Would he still have subjected himself to an unnecessary death sentence? Most likely not.

This game between what is and what could have been is evidence of a higher cognitive function, underpins every speculation about possible worlds, and allows us to anticipate response scenarios. Now, this possibility of pre-adaptation, outcome evaluation, and speculation about strategic revisions, is at the heart of counterfactual hypothetical reasoning. How can a scientific approach to this issue help us better understand such a role, and how does it speak to the issue of morality?

Today there is a rediscovery and appreciation of the role of counterfactuals in the fields of Literature, History research, Cognitive Psychology, Moral Psychology and AI, just to name a few of the more relevant areas.

Specifically, counterfactual reasoning consists in the imagining of an alternative scenario in relation to the one that actually happened, and the exploration of its consequences. For example: "If the forest floor had not been covered with dry leaves after the long hot summer, then the lightning would not have caused such a tremendous fire."

Applied to the morality of groups, its relevance is as much related to the construction of alternative hypothetical and credible scenarios about the past as to the choices made or about the events that occurred and, concomitantly, the assessment of the various consequences that would have followed. Properly conducted, counterfactual reasonings can provide very relevant insights into the ways ahead in the domains where they are applied. Thus, they are an excellent tool for understanding and explaining the mutability of certain behaviours, supported by the review of strategies, re-examining the past in the light of what we a posteriori know today. We can identify some of the reasons that make individuals build counterfactuals: the need to improve future performance, or to work over a factual event to make it more acceptable to themselves, or justifiable to others, either why we did not pursue the alternatives, or by teaching us from experience about what we could rather have done differently to what we did. This way of reasoning may apply as well to events that did not happen but could have happened. For example, to conjecture what the urban areas of the United States would look like if, instead of building the great railroads, investment had bet even more on rivers as a means of communication. Or about events that occurred, thereby reasoning about what would follow had they not occurred; for example, imagining that the Portuguese Revolution of April Twenty-fifth 1974 had not happened, and what the evolution of its prior so-called Marcellist Spring would have been. Or if a particular event had not occurred, but another would have in its place, for example, if massive exploitation of fossil fuels had not taken place, and if we had already then moved on to solar and wind energy exploitation. And even to verify if the alternatives would be indifferent with respect to relevant consequences.

In a certain sense, we can consider that all scientific laboratories are places of counter-factuality, because they create alternative scenarios, simplifiers of reality, where a given variable can be tested. To wit, reality is too rich and complex to serve as an appropriate place for certain scientific tests. If we want to know if "x" is the cause of "y" we will have to create a counterfactual scenario where this can be made evident. The fact is, we may be foreseeing the occurrence of "y" in a temporal sequence where "x" has already happened, and this happens successively because "x" is associated with "z" and it is "z" that actually causes "y" and also "x". Finding this out by observing reality may be utterly impossible—the number of items in co-presence is too high, and may lead to unnecessary misconceptions and unfounded convictions. Thus, in the laboratory, having a good conjecture and testing one variable at a time enables us to observe unsuspected and unambiguous causal networks. When Galileo conjectured that—in a void—all objects fall at the same speed, gaining equal

speeds at equal times, regardless of their mass, he had no technical means to test the theory. It was from his mental experience that he devised a system of highly polished conduits through which spheres with different masses rolled. Conduit polishing and ball perfection could minimize the inexistence of a vacuum chamber at the time, inasmuch friction was made minimal. Galileo thus constructed the possible scenario in his days to test a theory that very few would be willing to accept. Albeit, the perfect vacuum, as today we know, is impossible, for it is necessarily composed of vacuum fluctuations (without which Heisenberg's Principle would be violated).

Specifically, with regard to applications of counterfactual reasoning in the domain of AI, a scientific approach to the question of morality will always be treated by its consideration as a case of computer-implemented evolutionary game theory (EGT). Generally, game theory studies how, in a strategic relationship, rationally acting players promote the best outcome for themselves.

To do this, each player has to analyse the game, and identify the strategies available to achieve that goal. In the specific case of evolutionary games, where time dynamics is considered, in addition to the strategies used by each and their attending payoffs when playing with different partners, players' behavioural mutations over time must also be considered. From a long-term perspective, it is hoped that they will desirably select evolutionarily stable strategies that maximize overall utility for all through cooperation, whenever possible. To this end, it is necessary to hypothesise which resources agents can use to achieve these goals, and within the scope of parameters.

Now, the questions related to the collaborate/non-collaborate dilemma are pertinent in areas as diverse as Evolutionary Psychology and Economics. Thus, it is important to know whether or not counterfactual reasoning is an essential tool for understanding evolutionary dynamics in strategy selection, and for improving individual as well as collective gains in contexts where the greatest advantage is afforded by collaboration.

Given its broad spectrum and cognitive value, a relevant scientific question, and auspicious in terms of research, is what is the effective, if sufficient, role of a small minority of individuals endowed with this counterfactual rationality within a given population. More specifically, to understand if this minority—say twenty percent of the individuals—has the capacity to influence the whole group, encouraging cooperative behaviours. At present we take the theory of social learning, proposed by Bandura, as the most pertinent and interpretive one. Therefore, it is of paramount relevance to determine if the construction of counterfactuals can surpass it in terms of effectiveness in selecting strategies.

But before describing experiments already carried out in this context, we should warn that, sometimes, scientific research leads to counterintuitive results. An emblematic example of this feature is a study developed some years ago, here at the Nova University, which was based on mixed populations constituted by selfish agents, agents that balanced selfishness and altruism, and exclusively altruistic agents. In theory, being maximally altruistic can be perceived as an ideal to achieve. However, the inquiry into interactions between virtual agents showed that, after some moves, individuals balancing selfishness and altruism refused to collaborate with exclusively altruistic ones. There is an explanation for this refusal: It is the altruists

who end up "feeding" and reinforcing the selfish behaviours of the free riders, as they are parasitized by them. That is, the naive feed the opportunists! The absence of radical altruists decreases the presence of free riders in a given system.

This study shows that, sometimes, scientific research makes it possible to draw conclusions not only against moral common sense, but also in dissonance with the normative prescription of certain religious ideologies. Every Christian moral proposition rests on the radical challenge, proposed by Jesus Christ, centred on offering the other cheek and loving our enemy. However, scientific research shows that this—besides being an unreasonable requirement—eventually generates distrust in the system and diminishes the potential for collaboration in agents.

We have also alluded several times to the extremely complex problem that has arisen from morals suspended on a religious or philosophical system. In order to avoid the resulting problems, a scientific approach will select aspects that are fundamental to group morality, allocable to all contexts, regardless of the original culture of each group, or the fact that the autonomous agent be biological or silica based. It will address in the abstract the elements—say, atomic ones—of all moral systems, such as: collaborating/not collaborating, acknowledging guilt and apologizing, acknowledging or expressing intentions, etc.; and the way in which these aspects may or may not, individually or intertwined with one another, foster group cohesion.

Equipped with these two forewarnings, let us delve into our approach to the role of counterfactuals (Pereira and Santos 2019). In the well-known case of the game *Stag Hunt*, a cooperation dilemma is contemplated, which helps us establish the importance of building counterfactuals. It is a game for two, and the mission of those involved is to hunt stag, a task that must be performed together to maximise the possibility of success. However, each player may decide not to collaborate and choose instead to try and hunt hare on their own. Although it is a less rewarding alternative, the decision can be interpreted as safer, since the hunter depends only on himself, and hare is easier to hunt than stag.

The dilemma results from each hunter not knowing what the other will do; that is, whether he will collaborate and hunt stag, or will act on his own, deciding to hunt hare. So, each one can be tempted to protect himself by hunting hare. The compensations differ according to each option taken: level 4 for the decision to hunt stag, if taken simultaneously by both players; level 3 for the decision to hunt hare alone; and 0 for the player who decides to hunt stag without the other doing so. We are thus facing a cooperation dilemma in which maximization of the outcome depends on the effective decision of cooperation by both players. In the context of *EGT*, players review their strategies, watching each other's actions and copying the most successful ones. However, the application of counterfactual reasoning to the *Stag Hunt* game shows that—contrary to what happens with the mimetic process proposed by social learning theory—the individual can conjecture what would happen if he used another strategy as his own rather than the one he had used. Experiments with computer-modelled artificial agents clearly show that counterfactual reasoning is much more efficient and fruitful in revising strategies than simple mimicking of the most successful strategies used by the adversary. Note that the game also shows that the creation of counterfactuals is a merely instrumental mental activity solely dependent on oneself.

That is, it is also a resource available to those who systematically opt for selfish strategies. There exists counter-factuality for the good, and for evil.

Now, such a situation shows that, if we wish to have machines endowed with moral capacity, capable of selecting moral decisions that optimise the expected results and, at least, maximise the expected utility (using here the utilitarian paradigm, with due reservations), it is crucial that we learn to program them with the capacity to develop counterfactual scenarios. These prove to be excellent tools for selecting alternatives not available in the behavioural portfolio for mimicking, and may result in improved cohesion and cooperativeness within groups.

Counterfactual reasoning is also usable for judging, morally, the intentions of an agent's act. One counterfactually assumes that a certain noxious side effect that occurred might not have occurred. Even so, would the purpose of the acting agent have been accomplished? If not, then this side effect was indispensable and therefore might have been intentional. If so, then it was not necessary to achieve the goal and therefore the noxious effect did not need to be intended.

References

Milgram, S. (1974). *Obedience to authority*. New York, NY: Harper & Row.

Pereira, L. M., & Santos, F.C. (2019). Counterfactual thinking in cooperation dynamics. In: Fontaine, M., et al. (Eds.), *Model-based reasoning in science and technology—inferential models for logic language, cognition and computation* (pp. 69–82). *SAPERE series* (Vol. 49). Berlin: Springer.

Chapter 15
Aside on Children and Youths, on Identity Construction in the Digital World

Abstract Symbiotic processes have a special impact on children and young people. Born in a world of technological tools paraphernalia linked to the Internet and the most widespread media, they cannot even conceive of a life where they would not be permanently connected to the network. Traditional notions associated with privacy are thus questioned without much awareness. The impacts of fragmented information, of the way social networks summon reactivity and immediate emotional response, of the permanent presence of the other mediated by a smartphone, a tablet or a computer, are not yet thought out and conceived in all their consequences. However, the phenomena of scattered and diffuse identity and the emergence of behaviours intolerant to frustration are becoming increasingly evident. In a world where in each of those present in the network there constitute within themselves like one alter ego (or more), youths have difficulty in structuring a solid and differentiating identity, caving before the multiple pressures they are subjected to. Perhaps in the near future the notion of building a differentiated identity will not have the same pertinence it has today.

We are not born complete. Our identity is built in interaction with the environment. This is one of the safe acquisitions of Developmental Psychology. While there is no widely shared paradigm about identity-building issues, not even the staunchest advocates of genetic determination deny the influence of the environment on its construction.

Bearing this in mind, it should be noted that the ubiquity of computer media—from smartphones to tablets and computers—characterizes the context of any child or youth today, in any culture, as long as they have a minimally acceptable standard of living. The term "minimally acceptable" means that people do without purchasing books, family outings, or even an enriching trip to acquire the latest phone of a given brand, or the most alluring tablet on the market.

On the other hand, the emergence of a new addiction is well known and relatively well documented. This not only involves the need to be permanently online, connected to social networks, but also involves Internet games that become a priority in the lives of children and youths. Many adults also suffer from such dependence, with severe consequences on their social and professional lives.

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_15

The case of children and youths is emblematic because they are at the stage of building their identity. Yet, it is at this critical stage that the whole panoply of digital technologies is mediating their social relations. Online games, virtual social networks and the ubiquity of the internet thus bring about a new context for development. Is it possible to characterize the impact of these new technologies on developmental processes, namely on the construction of identity?

There is much research to do in this area, but it is certain that the terms *cyberculture symbiosis* and *syncretism* provide the key to an approach to the issue. First of all, we can identify what is important in the immensity of what is now called "cyberculture", trying to find structural and structuring concepts. At the outset, we will identify two. The first, is about dilution, namely the concept of "syncretism". The other, "symbiosis", refers to a contributive and constructive individuality in a common ocean of individualities.

The symbiosis/syncretism issue dates from far back, it is a problem inherent in biological life itself. Bacteria had to symbiotically cooperate to form eukaryotes, single- or multi-cellular living beings with cells already containing an individualized nucleus, separated from the cytoplasm by a surrounding membrane. Eukaryotic cells were formed by bacterial associations. From the latter, they maintain mitochondria, which are self-replicating entities with their own individuality, within the eukaryotic cell. In addition, organelles from other eukaryotic cells (viz. primitive and unicellular green algae) were adopted. All of these entities participating in the global metabolic cooperation that constitutes a cell with a nucleus.

The issue of individuality/dilution, symbiosis/syncretism then recurs and emerges at successive levels: from the organs to the organism, from the latter to the individual, from it to the group, from them to society, and from the latter to information networks and planetary info-ecology.

In order to proceed, we shall first of all provide the structuring meanings of cybernetics and cyberculture, the latter defined by analogy with the former, as well as the definitions of symbiosis and syncretism.

15.1 Cybernetics and Cyberculture

In his book "*Cybernetics: Or Control and Communication in the Animal and the Machine*," Wiener (1948)[1] first coined the word "Cybernetics." It results from the Greek κυβερνητική (*kybernetike*), meaning "governance": i.e. all that pertains to driving, navigating, and piloting. The word κυβερνήτης (*kybernetes*) means "the steersman, or captain of the ship."

The book's subtitle, "*Control and Communication in the Animal and the Machine*," suggests that there is something common to the animal and the machine concerning communication and control. Namely, how informational signals may be encoded, transmitted and decoded; and how such signals allow to exert control,

[1] Accessed at https://en.wikipedia.org/wiki/Cybernetics.

through retroactive loops that keep the focus on the objectives, and through sensors and discrepancy correctors between the target stage to be reached and the current stage. After all, Wiener was interested in employing these communication and control capabilities to pilot anti-aircraft missiles, as well as in stabilizing the human heart. The research focused on the mathematical formulation and implementation of mechanisms of control and communication, inspired by those found in living beings. Cybernetics had immediate applications in radar, missile control, and medicine, and has since been influential in the study of mechanical, physical, biological, cognitive, and social systems.

Although in the 21st century the term "cybernetics" is loosely utilised to identify any system using information technology, we are not far from the meaning of "cyberculture" in the context of a social cybernetics (or "socio-cybernetics"). This has led us to an attempt to define it, by analogy, with "Cybernetics." There results the definition "*Cyberculture: Or Cultural Control and Communication in Networked Mechanisms*." That is, we appeal to the abstract notion of enabling mechanism, while something common to living beings and their artefacts (such as the machines of human technology), but now extended to the notion of networking, which is a locus of cyber-cultural opportunity for cooperation.

"Cyberculture" thus comprises: cultural communication through technology; emergence of cultural behaviours in a technological network; cultural influence and control of communication and behaviours in that network.

It involves various components and features, among others: attention and inattention; encoding and decoding; human and non-human agents, plus avatars; sensors and actuators; augmented reality; multi-tasking; collective and distributed memory; big data over data mining; emerging network structures; self-evolution; control and lack thereof control; and so on.

Cyberculture therefore encompasses the networked emergence of culturing behaviours—and this is new—because emergence is what happens when several previous things are brought together, and new entities and new phenomena that were not anticipated at the outset appear in the newly formed set. This is what happened when the first eukaryotic cells appeared, whose emergence took, however, a good couple of billions of years subsequent to the first living cells. Emergence generates the problem of cooperation. Darwin did not know how to explain cooperation: how, despite all competition cooperation comes, which is a prime requisite for gregariousness.

It is extremely important that we study emergence, because when we put all the many entities together—some of them entirely new—into the world network, new things will emerge. New elements and behaviours will emerge, adjusted to the new system of co-dependent interactions. And just as an organism is made up of similar cells, functioning in groups and syncretically in organs, and these in turn functioning in symbiosis in that organism, etc., through various multilevels of association, we may say that we are still in a, say, infantile stage of network emergence, where we are probably going to become diluted. The question arises as to what extent we will syncretically dilute ourselves, or to what extent we will introduce, individually or in groups, some amount of symbiotic structuring.

We have studied extensively this facet of emergent cooperation using Evolutionary Game Theory (EGT), i.e. the application of game theory to mutating evolutionary populations. Indeed, EGT provides scaffolding for the mathematical definition of competitive games, strategies, and analysis of competition and cooperation models, and is used to predict the results of having a multiplicity of strategies that evolve in co-presence. EGT differs from classical game theory in its emphasis on the dynamics frequency of each strategy, even under the effect of spontaneous mutations. EGT helps explain the basis of altruistic behaviours in evolution, whether biological or cultural. Consequently, it has gained the interest of economists, sociologists, anthropologists, philosophers, and computer scientists.

Therefore, we have been examining how and under what conditions moral behaviours emerge in networks of agents (Pereira 2016a). Because without moral rules there can be no cooperation between agents, be they machines or humans. We are engaged in researching how to make machines moral, since they have to live among us, and be convivial with one another too (Pereira and Saptawijaya 2016). Machines from different manufacturers will need to have something in common, with respect to their behavioural regulations, and that will be the said emerging morality. EGT is the ideal mathematical theory for studying the emergence of moral behaviours as a result of various co-presence strategies, as it allows us to analyse this in the abstract, and to indicate how they may be concretely implemented, in a computational manner.

How can we think about this problem from the cyberculture standpoint, with some of the above-mentioned components and functionalities? A cyberculture which involves an entire info-ecology—an information ecology—where each of us is but a small portion of a huge network (symbiotic?) , itself evolving (overly syncretic?). How and where to begin to grasp such a complex thing in what it concerns us as a whole?

Yet, cyberculture manifests itself in both syncretic and symbiotic structures; therefore, it is important first of all to provide the definitions that we will use.

Symbiosis—According to the "*Infopédia*" Porto Editora, there are three meanings to "Symbiosis", the last two in a figurative sense[2]: (1) Meaning in biology: association of individuals of different species, with mutual benefit (at least apparent); (2) Figurative sense: intimate association of individuals; (3) Figurative sense: cooperative relationship that benefits the individuals involved.

From these above-cited meanings we will specifically embrace the third one.

Syncretism—According to the "*Infopédia*" Porto Editora, there are three meanings to "Syncretism"[3]: (1) Meaning in Religion: fusion phenomena of different religious doctrines or practices; (2) Meaning in Sociology: fusion of different cultural elements; (3) Meaning in Psychology: primitive form of perception and thought; characterized by global, undifferentiated, indistinct apprehension; patent in early childhood mentality.

[2] Accessed at https://www.infopedia.pt/dicionarios/lingua-portuguesa/simbiose?ic-click.

[3] Accessed at https://www.infopedia.pt/dicionarios/lingua-portuguesa/sincretismo?ic-click.

From these above-cited meanings we will adopt and extrapolate more specifically the third. This psychological sense begins when the child is born, in which it is still fused with its exterior, as if it were still in the womb; in which it does not distinguish between itself and the world. Only afterwards does it begin to disentangle the homogeneous and the heterogeneous, to distinguish between itself and the mother, between itself and the world, and there begins the process of identity creation.

15.2 Focus on Young People

Now we are all in an infant phase of the web's development. The impact of what will happen to us in the future in the long term will be very much the result of what will happen to our children in their development with the web. How will our children and grandchildren be affected by this stage of identity attainment in an environment that is largely one of dilution?

Therefore, given the importance such dilution has for each of us own future development, and of our current joint infantile stage on the web, let us focus here mainly on the problem of young people's identity development in this engaging web age, leaving aside those of us who developed their identity at an earlier time. From now on we will concentrate only on this focus of origin, on young people, so as to grasp, by whichever end, the vast complexity of the "cyberculture" theme—one too unfathomable for a single chapter.

So far, little attention has been paid to this problematic topic except by psychoanalytic authors. In particular, attention has been paid to the reasons or motives that lead young people to be increasingly together via the net, and at the same time ever more alone, according to Turkle's felicitous title, *Alone Together* (Turkle 2011).

In this book, Sherry Turkle also tells us extensively about robots for the elderly, and robots for children and young people. In what that also impairs the formation and maintenance of identity, as we need to have some other that is not just an extension of ourselves, one other that is human, that has initiatives, that can say no, that can argue. Such one other tends to disappear.

This dilution is not only propitiated by the relationship with active screens, but also by the excessive access to the network as well. It is also favoured by the increasingly intense and widespread relationships of proximity with robots. We will not deal here with these "plush toy robots", but more abstractly rather with digital communication.

Digital technology has profoundly changed lifestyles, the speed of interpersonal communication, and the quality of relationships.[4] For young people, digital devices are extensions of their own bodies, inseparable from the sentiment of self and group identity (Lemma 2013). The boundaries between virtual world and external reality become blurred, and the self may omnipotently lose the organizing references of actual circumstances.

[4]For further discussion, see: Gonçalves (2016) (Portuguese Psychoanalysis Society).

What is the influence these changes have had on young people's subjective life and development? There is more impulsiveness, activity and perception, but less structuring thought about information. There is no time to organize the information. The (psychoanalytic) defence mechanisms are, therefore, more primitive, thus giving rise to a greater self-cleavage, a greater denial, and a greater tendency for adhesive identifications.

Such changes in the subjective lives of young people do not respond to their evolutionary and emotional needs.[5] Tensions between internal needs and external determinations increase; its resolution is frustrated and, in psychoanalytic language, there is less repression (mechanism that keeps emotions, drives, affections, etc. in the unconscious), and less displacement (unconscious transference of an intense emotion about the object of origin to another one). It also diminishes patience, attention and concentration, tolerance to frustration, to waiting, and to uncertainty (Bilbao 2016), so far are the stimuli. The connection to the net creates a dependency that needs to be continuous (Kardaras 2016).

There is, therefore, more externalization (people live more for what is external), and hence less interiority and cohesion of the self. The very parental dispersion, when permanently and daily taking place, caused by this same digital technology, aggravates in the youth the feeling of isolation and self-devaluation. It creates the addictive need to see immediate responses to postings, whose return produces biochemically pleasure, as demonstrated in laboratory. Almost like the mice that incessantly press the button that provides them pleasure via an electrode implanted in the brain.

On this, José Pacheco Pereira writes:

> Societies without human neighbouring relationships, companionship and friendship, without group interactions, without collective movements of common interest, depend on artificial and, I insist, poor forms of relationships that become as addictive as drugs. There is no greater punishment for a teenager than taking away their mobile phone, and some of the most serious conflicts that occur today in schools are linked to the mobile phone, which acts as a lifeline.
>
> Nothing is more meaningful and depressing than seeing people at a school entrance, or at a popular restaurant, or in the street, people who are together but barely talk to one another, but are attentive to their cell phones, texting, sending pictures, viewing their Facebook page hundreds of times a day. What life remains?[6]

The permanent connection to the network, and the being chained to their devices, does not favour independence from the object—the one other—nor mental elaboration, due to its absence. The web is an extension of us, and of our avatars. Alter egos can be created, not consolidating any particular ego, because it is easier to remain diluted amongst alter egos. This leads to schizoid situations.

Obviously, the construction of a solid self-identity, with a well-defined differentiation, essential for creativity, consolidation, and security is compromised. One of the reasons why animals, in general, are always alert or on the move, permanently busy

[5] This whole universe of questions is properly explored in the text of M. J. Gonçalves, referred to in the previous note.

[6] Pereira (2016b).

in their awake time, is that being alive requires energy, and energy must therefore be constantly used in the possible best of their ability. If the animal uses calories to stay alive, and does not use that energy well, looking around to perceive and scrutinize the environment, and to detect eventual predators, the energy cost is wasted. There is, therefore, a deep anguish of life itself in employing time well. This horror of the vacuum has to be reformulated in human beings in internal constructions that prepare us for the future, not leaving us permanently obsessed with the present.

The psychic work of de-idealization of the image of parents is also put into question. The youth moves into a wider fusion rather than striving to break free from parental fusion. This compromises the ability to be alone with themselves. The historical track record of reality in space and time is lost. Personal identity is denied via the ever-available floating identities—evident in the personal profiles provided on social networks, and in game avatars. Even sexual difference might be denied. All in all, no good lessons are learned from too easy an alienation in the relational virtual world and its apparent opportunities.

For all this, the mimetic and adhesive identifications are reinforced. We tend to say "I am the same as that one" or "I reject that one." Growth is established by dependent mimicry and affiliation, and not by self-construction.

15.3 Symbiosis and Syncretism

We have been placing on both sides of the scale syncretism and symbiosis—the latter, it should be noted, does not correspond to the psychoanalytic use of the term. Both are needed and coexist. The problem we raise is that there is increasingly more syncretism and less symbiosis. We risk amalgamating ourselves as individual beings in the planet's info-ecology, in the global semantic network, as well as losing our identity. We may dissolve into a superorganism. Perhaps like the ants, perhaps it is inevitable even to be diluted in this superorganism. We have no answers to this, but we believe that questions about Cyberculture involve these two concepts, and pose such problems.

We highlight the constructions below, which are clearly few, but decisively exemplary of Symbiosis in Cyberculture:

- Wikipedia, Wiktionary.
- Common blogs.
- Public data repertoires.
- Software in common, viz. SourceForge (https://sourceforge.net).
- Real-time scientific cooperation.
- Provision of common Cloud resources.
- Preparation of petitions.
- Deposits of joint collections.

We also highlight these diluting facets resulting from Syncretism in Cyberculture:

- Imperfect psychic evolution.
- Superficiality.
- Lack of time. Misused time. Agitation.
- Hyperactivity and attention deficit.
- Fusional incoherence. Constant need for new stimuli.
- Discontinuity and continuation failure, due to hopping.
- Ineffective multitasking.
- Schizoid personality disorder.
- Dilution of the self and emotional bias.

In short, it could be said that, in Cyberculture, concerning youths:

- There is too much syncretism and too little symbiosis.
- Further co-construction of knowledge is lacking.
- Greater and more independent personal cognitive deepening is lacking.
- The capacity to be alone is lacking, in place of *Alone Together*, in the serendipitous expression of Turkle (2011).

It is, therefore, the very cognitive development of the new generations that is at stake. What this means for humanity as a whole, and for subsequent generations, is that there is more and more "being together, but alone." The face-to-face and relationships as wholes, are lost. Each one is on their smartphones. On Facebook, or other social networks, everyone is controlling what they say. Young people today do not like to call, because the phone opens up conversations, who knows where they can head to, and how long they can take. They do not even like the email, because it is too open in length, and left pending longer, waiting for more elaborate answers. They prefer the compact, controlled, two-line SMS, and if one message exchange does not suit them, they drop it and move to another.

15.4 Causality and Free Will

Symbiotic causality occurs due to the persistence of a strong internal determination from the inside out. The individual wishes to do this or that, and has his personal reasons and track record for wishing to do so, in order to influence the outside and avoid being overwhelmed by external causes. Syncretic causality is submerged by external determination, which occurs from the outside in. The person is diluted before the external stimuli constantly bombarding, without time to elaborate and counteract a causality in the opposite direction, from the inside out. The person therefore then reacts on impulse with the *sound bytes* of the occasion, often "kicking towards the corner."

15.5 Coda: Cyber-Selves—Distributed or Not At All??

These topics raise vast questions, hence the double question mark. Below we provide some provocative interrogations in response.

At the cybercultural technological intersection we are in, can we at all costs maintain an individuality, perhaps symbiotic, or will we rather collapse in the face of invasive syncretic synergies? Do we want at all costs to retain and affirm an individuality, or will we inevitably become diluted in the identities of the group? Will we resist, or will we surrender to the invasive and syncretic synergy of football events in the media? Or to the dilution on reality shows of the TV news about judicial and court cases in daily episodes? Or in the comic-book news about the economic life of the markets and world politics?

In the emerging Cyberculture, the core notions of self, separation, and individuality are very important. These, with their much emphasis in "Western culture", are not so relevant in other cultures. In the West, it is known that the concepts of self, separation, and individuation are very pronounced, and in contrast to other cultures, namely in the East.

An example of this occurs in therapy. In the West, the individual self is the object of therapy: a self that values differentiation. In the East, the relational self is more permeable, and the self-other boundaries too. In this case, the identity unit is not that of the internal representation of oneself and of the other, but that of the family or community where the self is distributed and given priority (Lemma 2013).

The wisdom of the East may be relevant to Western Cyberculture (Roland 1988). There, the individual asks himself how he can, in symbiosis, contribute more, giving priority to the whole. Instead of how they can defend themselves more, giving syncretic priority to themselves.

References

Bilbao, A. (2016). *O cérebro da criança explicado aos pais*. Lisboa: Editorial Planeta.

Gonçalves, M. J. (2016). *Nascer e Crescer na Era Digital.* Conferência a 31 de Março de 2016. Lisboa: Sociedade Portuguesa de Psicanálise.

Kardaras, N. (2016). *Glow kids*. New York, NY: St. Martin's Press.

Lemma, A. (2013). *Introduction to the practice of psychoanalytic psychotherapy*. Hoboken, NJ: Wiley & Sons.

Pereira, L. M. (2016a). *A Máquina Iluminada—Cognição e Computação*. Porto: Fronteira do Caos Editores.

Pereira, J. P. (2016b). *A ascensão da nova ignorância*. Público online (Portuguese daily newspaper). Accessed at https://www.publico.pt/2016/12/31/sociedade/noticia/a-ascensao-da-nova-ignorancia-1756629.

Pereira, L. M. & Saptawijaya, A. (2016). *Programming machine ethics*. *SAPERE series* (Vol. 26). Berlin: Springer.

Roland, A. (1988). *In search of self in India and Japan: Toward a cross-cultural psychology*. Princeton, NJ: Princeton University Press.

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Cambridge, MA: The MIT Press.
Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine*. Cambridge, MA: The MIT Press.

Chapter 16
To Grant Decision-Making to Machines? Who Can and Should Apologize?

Abstract Entering the specific domain of computational morality first requires considering the History of Human Morality and its various nuances. And noting and justifying the existence of multiple origins for the diverse moral systems, yet with a common matrix: that of being the result of our evolutionary History in such a realm, having survived Darwinian selection. Subsequently, it should be taken into account that, in addition to various moral geographies, there are several companies and countries involved in the construction of machines with moral programming requisites. Thus, in addition to moral algorithms per se, there is a need to devise international standards and constraining legislation leading to compliance with agreed standards. Inevitably, with so many agents in co-presence, errors, misjudgements and misunderstandings will emerge. Hence the particular importance of apology. Whether we are dealing with a biological, hybrid, or artificial agent, what matters is that decisions are justified by arguments, there is goodwill and absence of malice and that apologies are genuine and sincere.

One cogitation about human beings in general does not raise major reservations: morality is an essential part of the glue that binds individuals within groups. Traditionally, as mentioned before, the implementation of values of this nature has been the task of religions. This makes perfect sense, because they mitigate individual finitude, giving meaning to life, death and resource sharing. Furthermore, they link individuals into cohesive communities larger than families, allowing for individual sacrifice to be subordinate to the interests and goals of a group. In the context of Judeo-Christian beliefs, the sacrificial dimension of the figure of Jesus Christ, the Son of God who gave his life for the redemption of others' shortcomings, is emblematic and unavoidable. There is no better example, nor more emblematic narrative about unconditional love for the other.

At the root of the Portuguese word '*Religião*' (Religion) is the Latin word '*religio*', associated with the act of strongly binding something, with '*R*' being a prefix of intensification. In its dense polysemy, it is also possible to read in it the act of imposing obligations on individuals who acknowledge themselves belonging to the same group. In this sense, the connection between religion and morals is quite strong. The word 'Moral' originates from the Latin term '*mores*', which refers directly to

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_16

admissible or obligatory customs within a particular community. There is thus an intrinsic relationship between religion and morals. In order to understand the role of religions in forming cohesive communities, it will be difficult to find a better example than how Islam managed to unite the various Arab tribes, decisively contributing to the reduction of tribal struggles in their peninsula.

On the other hand, although human groups tend to delimit their scopes in opposition to each other, there has always been a dynamics of agglutination by imposing the values of the dominant group. Once the dynamic implemented, and though aware of the limitations inherent in their cultural idiosyncrasies, Catholic philosophers theorized on the possibility of a natural moral law. This would be the first level of divine intervention in the world, constituting itself an expression of universality, giving meaning to a global apostolate. Natural Law, Old Law (revelation of the Ten Commandments to Moses, and, therefore, a close expression of Natural Law), and New Law—or Christ's (moral precepts associated with the New Testament)—constitute articulated epiphanies. Natural moral law, because engraved from the outset in beings that possess reason, inclining them towards acts adequate to their purposes, is regarded a matrix present in all Men of all times. Yet again on the path of aspiring to universality—although possible to argue against its validity—stands morality's golden rule: "Do not do unto others what you would not have them to do unto you", finds formulation in diverse cultures and religions. In fact, religious morals have many bridges to each other, as they are projections of the common past of the species and its Darwinian ethical evolution.

Regardless of each religion's moral systems, there are considerations that must be taken into account. It should first be noted that these were engendered in much smaller communities, where mutual vigilance was facilitated, making the constraining character of their precepts more incisive. Then, there are ancestral contexts, where group interactions were sporadic or even non-existent—a very relevant aspect when it comes to accepting a value system no questions asked. On these terms, throughout the History of Humanity, both religion and the moral precepts associated with it have been the identity mark of each culture, serving as point of reference to keep their agents accountable. It is no coincidence that responsibility is one of the axial concepts in the field of morality, even if that responsibility is not entirely the result of individual actions. Oedipus was destined to kill his father and copulate with his mother, even though the faults that sustained such destiny were not committed by him. We are not everything we wish to be, and there are forces within us that go far beyond our individual will. In his decision to blind himself, and withdraw from the world of humans, Oedipus blamed himself for moral faults that he had not consciously committed. Sophocles' lesson—in his literary monument founder of our culture—already evinces a consciousness according to which living together requires the assumption of responsibilities, not them all individual, and not them all desirable or sought after.

The traditional anchoring of morality in religion has solved the problem of its origin. From the religious standpoint—as is widely known, and we have stated before—it is believed that moral precepts were inspired by some god who, in general, has

the ability to oversee the courses of action and intentions of the agents. He functions as a referee, who dictates the rules of the game and screens everything, even within our heads. As a result, in the context of Judeo-Christian religions, the end of times will bring a final judgment, in which the dutiful will be rewarded and the transgressors punished—in the limit, by the fire of hell, where they will burn for ever more. Hence the endowment of humans with a moral conscience, associated with a supposed deliberative capacity, serving as the basis of accountability. From this perspective, moral conscience, abetted by guilt, should not only enable each one to recognize Good and Evil, but also engage with the pair obedience/disobedience to a supreme and creator being. Concomitantly, the presence of Evil, as a possible option, generates the context for the exercise of free will, thus conferring dignity on the choice of the Good.

With the advent of humanistic philosophies, associated with the emergence of modern and contemporary atheism, a new morality has emerged, further supported on the notion of human responsibility, for it presumes the absence of a god to accompany, support and watch over us. Feuerbach argued that God was but an idealization of the best features of humans being projected on him. To regain our dignity, there will be no alternative but to recover what we had previously attributed to God. Humanism thus establishes a moral without transcendence. Dostoyevsky, Nietzsche, and many others, wrote about the risks of humanity getting lost in a world without transcendent references or anchors. The truth is that, in this context, it was possible to continue to develop an aggregating moral argumentation, embodied—for example—in the various charters that found intercultural dialogue. The most significant is the Universal Declaration of Human Rights. It is true that its moral precepts are disrespected every day, but it is also true that they continue to be the most common criterion for signalling such disrespect. After all, there is nothing new here, moral principles have always been declared as obligations, but immorality too has gone hand in hand with morality.

More recently, a scientific approach to this problem discounts the possibility of god existing, and is based on the data of evolutionary psychology. In this scope, morality emerges as a case of evolutionary game theory, and has to do with establishing the rules that provide the best survival conditions for a group. Consciousness has here the function of recognizing the criteria of good action and identifying when those criteria have been properly or inadequately applied. On the other hand, we face a new and surprising challenge: in situations of decision-making, not just human beings intervene, but cognitive machines also have become present. Admittedly, they do not yet decide from their own motivations and interests, but through principles and powers granted by human beings. However, they already force us to consider which conscious and rational agents constitute groups. And to try to spell out our morals in a precise enough way to program them into a computer. There is, thus, a return value of computational morality—the one which is experimentally implemented on the computer—which leads, via that way, to the refinement of our moral models.

In short, we have a moral knowledge to which various sediments resulting from our evolution have contributed, and we are creating a circumstance where the actors will not all be biological. Moreover, concerning biological beings, with the advancement

of biotechnology, we may have agents impressively improved by these technologies, who will interact with humans that were not subject to such improvements.

Hence, issues such as accepting an apology from a machine, or suspecting that it may be selfish, making it promises, or accepting healthy competition with it, may be starting to be formulated. And still, if we don't forgive it, will it forgive us?

From a strictly moral point of view, it is quite necessary to start enouncing such questions, considering a whole set of factors as diverse as those previously referred. The problems could be further complicated by introducing the somatic markers hypothesis, explored by Antonio Damasio. In fact, we often decide something because the repertoire of our emotional memories so indicates. However, if we focus only on the two most influential modern models of morality, Kant's and Stuart Mill's, we find that each one is capable of providing decision criteria, but is rather limited as to the scope of their usage. There are ethical dilemmas showing that something can be moral in the light of the utilitarian perspective, and deeply immoral in the light of the deontological one. Imagine a charity dinner, sponsored by a disadvantaged children's support institution, in which a benefactor is photographed while signing a check for one hundred thousand euros in support of the promoting institution. From the utilitarian point of view, that subject is morally valuable—for he has just ensured the sustainability of this institution over a long period. From the Kantian point of view, however, his action is not moral—for his motive is merely selfish, self-promoting, and therefore not apt to be universalised, as Kant requires. Besides the utilitarian perspective, prevailing in Anglo-Saxon countries, and the prevailing deontological facet in the European continent, we can also refer to another great Moral Geography: we are referring to Eastern culture. Here, above the individual interest, group interest reigns, be it the company, the school, or the country. Such a perspective has a highly disruptive impact on solving moral dilemmas, as we will exemplify further on. Given these circumstances, how can we find a possible path to robot programming? Additionally, if at all possible, will it be desirable? Would it not be more prudent to keep cognitive machines merely as servants of human beings, without any residue of autonomy?

Yet another order of questions has to do with the impact of machine decisions. If we consider that certain decisions may mean life or death, the wealth of some and the impoverishment of others—for example with funds and shares on the stock exchange—is it reasonable to allow non-human agents to make decisions that impact so sharply on the lives of humans—particularly at the very insipid level whether of their cognitive skills or of their fundamental inability for moral discernment?

Presently, from the point of view of their functional performance, and with all the above-mentioned limitations, cognitive machines are already getting much closer to us. To clarify this idea just think of the role expected to be played by virtual assistants developed by all major companies in the new technologies' sector. Like guardian angels, working very close to humans, but with increasing autonomy. Now, the appearance of these agents will not diminish the complexity of the natural and social world. Thus, the issue of moral dilemmas only gains a widened breadth that, till now, only existed in the purview of science fiction.

Imagine a robot in a nursing home. It is helping residents eat and feed themselves, but it also distributes prescription drugs. Of a morning, a resident asks it for more potent pain killers because they have a terrible headache. However, the robot is only allowed to deliver other drugs with the approval of a doctor. Since none of the doctors are contactable, will the robot let the resident suffer, or will it make an exception? Similarly, a robot that has to decide between saving a young man and saving an aged doctor, who in turn can save many other people, will keep having at hand a major moral dilemma. Its decision will depend on how it will be morally programmed.

But there are other scenarios to consider: imagine a team of robots involved in space exploration. They will certainly not all have the same function, nor will they have been manufactured by the same company; so, different interests may surface. It is therefore necessary that they have common codes of conduct to regulate their gregariousness.

Imagine still a scenario where decisions are most urgent and impactful, as in the case of war—in Iraq, at the peak of the last US military intervention in that territory, more than four and a half thousand artificial military personnel were counted—inevitably many of these machines will increasingly have to make decisions without human intervention.

Everything expressed in this chapter's question directs us back to the special difficulty associated with introducing moral precepts into the universe of AI. The problem of liability is, in fact, of particular complexity. It has always been so, and today we do not have a Sophocles capable of thematising it in its full dimension. However, it is possible to make part of its contours explicit. On the one hand, there is the case of war, which is especially relevant, since all military powers, and not only them, develop very robust military programs based on autonomous weapons. But think, for example, that in a particular industrial enterprise, there are firstly the managers of that one company who decide what to do and what the budget is; but there is also the production management, who can decide to streamline the process and save money in order to have their merit recognized; there will also be a team of programmers, who might enjoy computer games and neglect their commitment to ongoing work; in addition, there will be a sales team who—in order to boost the product—might sell it as extra modules, however essential in terms of safety. At the end of the production process, if something goes wrong, who is responsible? Our world is complex, full of collective endeavours and interdependencies, making it increasingly difficult to pinpoint individual responsibilities. In the field of Economics, the exact same thing happens. When we say 'The culprit of savage capitalism is the markets', we don't know very well, or rather, we know rather badly, exactly which institutions and people we are referring to.

In the absence of a general morality that underpins all decision criteria in such diverse and complex scenarios, and considering that we are still at a very early stage, we should start with clearly defined norms for specific situations such as hospitals, childcare, libraries, or nursing homes. The machines themselves will also learn, thus gaining the ability to modulate their performance, according to the cultural characteristics of the setting in which they find themselves. Such learning should have, as non-negotiable criterion, the concern to treat similar cases equally, which

underlies the understanding of the term "universality". The particular case of war will require much more specific work, which is certainly already being developed in that context. Most likely, given the gigantic discrepancy between the destructive power of such weapons and the quality of their autonomy, it would be wise to ban their use. The problem is that those who could take this action are precisely those who promote them!

In order to endow machines with ethical discernment, even in the current framework of shortage of interdisciplinary knowledge, there are situations, like those of computer games, in which they may learn moral precepts and principles. We have created a game wherein a robot has to save a princess, and make moral decisions throughout the process. The program shows how to combine different moral approaches. To reach the castle where the princess is trapped, the robot needs to cross a river traversed by two bridges. One is guarded by a giant spider; and the other is guarded by a ninja. The robot can imagine different possible solutions—a very important ability in making moral decisions. At the onset, it takes an utilitarian approach: it fights against and kills the ninja, because it conjectures that its chances of survival are better than those of facing the giant spider. However, the princess rejects that solution because she does not want to be saved by something that kills humans. So, on the next incursion, a new rule is introduced that takes precedence over the previous ones: the rule is not to kill humans. Through this process, step by step, the moral framework of our Princesses Saving Knight is successively built.[1]

Indeed, data from Evolutionary Psychology show that throughout evolution there must have been successful human behaviours that have been replicated, while others might not have been so conducive to group logics and therefore have not survived. This conjecture can also be tested in statistical models designed to research the moral behaviour of individuals integrated in groups. Simulations with artificial agents show that, over time, the most successful strategy becomes the most widespread and is passed on to subsequent generations. In one recent study we showed that guilt promotes cooperation. In cases where cheats feel guilty, show remorse and apologize—this behaviour restores confidence, thereby benefiting future interactions. The descendants of these elements maintain the same feature, exponentially increasing the number of individuals who incorporate it. Over time, it becomes dominant. Such a mathematical model thus shows that there is a reason why guilt developed, and spread throughout society. It also manifests that if we want to program moral machines, we should give them a sense of guilt and the corresponding ability to apologize. Under this assumption we have researched aspects of group morality related to the role of guilt acknowledgement and apology. We know, from the study of human populations, that these aspects can be enhancers of cohesion and cooperativeness amongst agents. What the study has shown in artificial populations confirms such knowledge. Indeed, keeping promises, acknowledging guilt, and apologizing, promotes cooperativeness and discourages selfishness. However, the apology must be sincere, have

[1]Robot saving princess demo (.mov) https://drive.google.com/file/d/0B9QirqaWp7gPWTFaSWJJak1YQVE/view?usp=sharing.

a cost: it is necessary to develop strategies for recognizing hypocrisy, so that participants can discern in which situations agents actually intend to collaborate. In a network of cooperating agents, it makes no sense that only a few be capable of moral discernment. The value of "feeling" guilt, and apologizing, depends on the idea that the other person can also feel guilty and, conversely, proceed similarly.

We should not forget the results of the research on the evolutionary model of the possible advantage of using counterfactual reasoning in a population of agents, as discussed earlier. As duly noted, this reasoning ability, even if possessed by a small number of agents, can make a population evolve to stable states that are beneficial to all. Thus, like guilt, it will be interesting to provide machines with the ability to think counterfactually: "Knowing what I know today, it would have been better if I had acted differently."

But, in addition to the connection robots establish with humans, there is also the problem of the relationship amongst them. If they are in contexts where human intervention cannot be direct—like space travel, deep underwater environments, or complex disaster scenarios—the issue is very relevant. There will probably be several robots from different manufacturers, and we cannot assume that they all use the same software. Yet, or even more so, they need to cooperate. Our aim will not be to send robots to Mars only to see them destroy each other once they get there. When machines from different manufacturers work together, even as security guards in a shopping centre, there are always dangers that can result from market competition. There is also the risk of a robot being programmed with sinister intentions. Perhaps it might not have been sent by a security company, but by a criminal association intending to kidnap the wealthiest customers. One of the goals of computer morality, as we said, will be to detect cheats and free-riders.

Because of these issues, it is urgent to devise and share legally binding international norms. They must be properly monitored by international bodies so as not to protect either unfair competition or types of crimes very difficult to detect. We are dealing with powerful processes of handling information, which can influence disparate aspects like purchasing decisions or election results. Therefore, we cannot be careful enough.

Over the centuries, morality has always fostered trust and altruism. But there is an aspect where moral knowledge is completely different from scientific knowledge. From the standpoint of scientific and technological advances, we do not expect major setbacks. For example, it is not reasonable for people to aspire to live as if there were no antibiotics available. And while theories about Flat Earth abound, their credibility is so low that they deserve no counterargument. As Aristotle said, if you want to be sure that snow is white, just look at it, no argumentation is required. Scientific knowledge is cumulative and progressive. Moral knowledge, on the other hand, does not have the same characteristics: not only can each of us recognize X as the right attitude and, yet, do Y, but also there may be unforeseen setbacks. The Egyptian society of the 1990s was much freer and more egalitarian than it is today. Similarly, in our Western society, as we talk about freedom and dignity, more and more people are working for an income that only gives them access to (poor quality) housing and food. Now, this is a situation close to slavery, thought to have been eradicated.

These aspects are a strong argument for the need for further research in these topics. A good moral will be one applicable to all agents endowed with common sense and reasonableness. Hence it does not make much sense to consider what the origin of the moral act is—be it that of an artificial, hybrid, biological agent or otherwise. Ultimately, an apology will always be an apology.

Chapter 17
Employing AI for Better Understanding Our Morals

Abstract Having addressed the prerequisite issues for a justified and contextualized computational morality, the absence of radically new problems resulting from the co-presence of agents of different nature, and addressed the difficulties inherent in the creation of moral algorithms, it is time to present the research we have conducted. The latter considers both the very aspects of programming, as the need for protocols regulating competition among companies or countries. Its aim revolves around a benevolent AI, contributing to the fair distribution of the benefits of development, and attempting to block the tendency towards the concentration of wealth and power. Our approach denounces and avoids the statistical models used to solve moral dilemmas, because they are "blind" and risk perpetuating mistakes. Thus, we use an approach where counterfactual reasoning plays a fundamental role and, considering morality primarily a matter of groups, we present conclusions from studies involving the pairs egoism/altruism; collaboration/competition; acknowledgment of error/apology. These are the basic elements of most moral systems, and studies make it possible to draw generalizable and programmable conclusions in order to attain group sustainability and greater global benefit, regardless of their constituents.

The issues associated with the scientific approach to the origins of morality, alluded to a number of times, are of special relevance. It is implausible that the modelling and programming of moral behaviour in machines be reduced to a mere calculation of the human solution most often adopted for each dilemma, as this may consist in the perpetuation of wrong decisions. This has been a widespread approach in proposals for solving moral dilemmas. While statistical treatment of massive data enables us to highlight the most consensual solution—or the one that is preferred in each culture, or in each age group, or by gender—still most are never a trustworthy criterion. On the other hand, the adoption of a single moral system, such as utilitarianism or some similar one, as opposed to a select combination of them, can result in an unspeakable monstrosity. As a result, we must expound an approach that goes beyond cultural idiosyncrasies, or even beyond whichever agent may be acting.

At present, the scientific literature offers us some attempts to categorize the field known as 'artificial morality' or 'machine ethics', a field that clearly makes explicit the goal of an understanding that encompasses all autonomous agents. Although, of

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_17

course, the differences are duly valued—both those resulting from different starting points and those arising from a multiplicity of interests. There are proposals for different systems and models heretofore developed, either by philosophers who attempt to support conceptions of how ethics operates by resorting to computing, or by the community of those who work to develop applications using ethics. However, if we are to avoid very serious problems—some of which we have already mentioned and others we are yet to address—these studies need to be further investigated.

On the one hand, there remain the theoretical endeavours that attempt to lend, support, or demonstrate, a certain view of our own ethical life, or of one or another theory about some of its features, by adopting what they call the top-down approach. Here, a certain model is designed, constructed, and governed in advance by what we might construe as implications or consequences of these theories. Thus, theories of rationality are applied to the construction of an "ethical" module (operated, for example, by *prima facie* identified functions as important for the ethical behaviour of human beings), which subsequently circumscribes the functioning of the system as a whole to different objectives (such as prudential or safety aggregates, among others).

On the other hand, bottom-up formulations on how to develop moral machines go in the opposite direction, seeming to be much more promising—they provide a more intuitive way to substantiate ethics in the beings we believe able to participate in this dimension. In a way, they do this by emulating an evolving learning dimension (following Turing's suggestion not to focus on the finished adult mind but on the child in formation), something similar to what we suppose has happened to us in species evolution. This conceptualization provides the authors with precisely the panorama of critical exercise desirable to maintain.

Whatever is made explicit in this area should keep in mind the complementarity of different focuses of analysis. Moreover, it should not exclude Richard Feynman's premise that we only understand what we can (re)build, and that an effort in the field of moral machines can be viewed as an attempt to better understand human morality.

From the panoply of approaches to these questions—although all the relevant material may shed some light on the way morality itself is regulated, or idealized—the most promising models will be those that embody the concept of self-development.

With these ideas in mind, we can begin to formulate a question related to the need for a better understanding of autonomy as a faculty made possible by—for example—the exercise of counterfactual reasoning. Surely, models are just models, but we reiterate that it is important to know whether the task of developing yet more legitimate moral agents will not ultimately spell out a better understanding of ourselves.

In this sense, it is crucial to identify the various dimensions that have been under research, so that we can acquire a more global view of them.

Some of our previous research focused on the use of programming techniques in logic, applied to the computer modelling of cognitive skills for morality, without considering the role of emotions. At the level of individual morality, we addressed issues relative to permissibility and the dual processes inherent in moral judgments, bringing together ingredients that are essential to moral agency: abduction, integrity

constraints, preferences, argumentation, counterfactuals, and updates. Computation about these processes was established as our procedure to model the dynamics of moral cognition when considering a single agent; dispensing with addressing the cultural dimension, since it is still absent from machines. This approach of ours is thoroughly detailed in the book *Programming Machine Ethics* (Pereira and Saptawijaya 2016).

Within the collective, we report moral emergence through computational processes, again without emotions, using the techniques of Evolutionary Game Theory (EGT). Through models grounded on this theory, we show that the introduction of various cognitive skills (either on their own or in combination)—such as intention recognition, making of commitments, revenge, apology, forgiveness and guilt—reinforce the emergence of cooperation in diverse populations, compared to absence of such competences in the population (Han and Pereira 2019).

This realm of the evolutionary collective, which we have explored, will be summarized below, synoptically, but whose details can be found in our respective specialized publications, indicated at the end of the "Preface" chapter.

In studies on human morality, these two distinct but interrelated domains—one emphasizing, above all, cognition, deliberation, and individual behaviour; and the other emphasizing collective morality, and how it came about with evolution—seem separate, but are intertwined.

There are issues related to the development of individual cognitive skills, and their subsequent implementation and diffusion in some given population. That is, the ability to recognize another's intention, even taking into account how others recognize our own intention; the competences to request and to accept or refuse to accept commitments; the competences to apply complementary mechanisms in an adaptive manner; to monitor group participation and delegate this process to an external entity; competences related to cooperating or betraying; and those associated with the act of apologizing, whether that act was promoted by the need to avoid revenge or to obtain forgiveness. The following summarizes our research on the collective phenomenon, by modelling distinct strategies in co-presence, regarding cooperative and uncooperative behaviour in complex evolutionary games.

Given its richness and complexity, Evolutionary Game Theory provides approaches from different perspectives to scaffold other stages of our studies.

As noted earlier—the first book on the subject, titled *Game Theory and Economic Behavior*, appeared in the 1940s, co-authored by the mathematician John von Neumann (1903–1957) and the economist Oskar Morgenstern (1902–1977) (von Neumann and Morgenstern 1944). At the time, it was aimed at Economics, but later applied to the Cold War. When a situation originating from these areas becomes complex, sophisticated mathematical tools—and computer simulations—must be used to deal with equations that cannot be solved otherwise. The theme of games is as multifaceted as it is interesting, and abounding in various niches. We can conceive of genes and memes ("cultural genes") and their mutual combinations as continuous strategies in the game of evolution, raising questions and posing problems related to survival or triumph. We can also see the combinatorial evolution of such strategies,

and their possible mutations according to various conditions, which may be other partners, or the very rules of the game (or those of nature).

The notion of game includes uncertainty, and whenever there is uncertainty there must be some strategy that meets it, detailing the moves (actions) that are made, with some probability.

When strategies are co-present in the evolution of various partners, together with the idea of a game's payoff, we are dealing with the notion of evolutionary play, whose temporal development can be defined and studied in an abstract and mathematical way, and/or implemented and simulated on the computer. There are 'zero-sum' games and 'non-zero-sum' games. 'Zero-sum' games are those where, due to their rules, some players win exactly what others lose. In evolution in Nature, conditions are not 'zero sum'—that is, everyone can actually win, or everyone can lose. Wright (1999) analyses the evolution of culture and civilization with the underlying idea that, in Nature, 'non-zero-sum' games are possible—therefore, overall gain can be obtained through cooperation, leading to enlightened altruism. Sometimes, co-present strategies tend to strike a tactical balance. Consider the predator/prey relationship: neither the predator wants to completely exterminate the prey, nor can the prey multiply indefinitely, because that would deplete the environment's resources. Some of these studies are used by Economics to understand what may be the overall result of the sum of interactions between the various game partners.

It is relevant to consider whether the game occurs only once with any particular partner, or if the same partner can be met on multiple occasions; what role memory enacts in interacting with that partner; and if the possibility of refusing a partner is allowed. Let us take a closer look at some of these situations. We begin with the famous Prisoners' Dilemma (PD) , typical of the altruism paradox. There are two prisoners, A and B, both suspected of a certain crime. Either can denounce the other, confess or remain silent.

	Prisoner B = Silence	Prisoner B = Confession
Prisoner A = Silence	Six years in prison each	A = Ten years in prison B = Two years in prison
Prisoner A = Confession	A = Two years in prison B = Ten years in prison	Eight years in prison each

Consider the above payoff matrix 2×2, where the rows correspond to A's behaviour (to remain silent or to confess), and the columns correspond to B's behaviour (to remain silent or to confess). At the intersection of B's "confess" column with A's "confess" line, they both receive an 8-year prison sentence. If A confesses and B does not, A only receives a two-year sentence, while B receives ten years and vice versa. There is an incentive for either of them to confess in order to reduce their own prison sentence. This way, it would eventually be advantageous for them not to remain silent.

If one betrays by confessing but the other does not, he will be imprisoned for only two years, while the other will be imprisoned for ten years. But if both confess, they will be sentenced to 8 years each. The temptation to confess is great, but so is its inherent risk, because, after all, they would benefit each other if they remained silent, each receiving a sentence of six years in this case. Prisoners know the rules of the game; however, they do not know how the other player will play. It is advantageous for them to remain silent, but they do not know if the other will confess. If one confesses, the other, if silent, will be sentenced to 10 years in prison. A dilemma thus arises: it is good to remain silent, but there is a risk that the other will betray; and he who does this, unlike the other, will have greater gain. At worst, both receive an 8-year sentence—no one will therefore assume the risk. This is a classic game where both players seem to tend to confess—and not to benefit from what could be a mutual advantage, but which they are not sure they will enjoy. Firstly, they do not have the opportunity to talk to each other; secondly, because even if they did, they would still risk being betrayed by the other. They do not have a joint solution in the sense that A and B could choose what is best for both, where there would be a greater advantage guaranteed to both.

Everything becomes more complicated when one imagines A and B playing this game in a succession of matches, considering the track record of mutual behaviour. In this case, they may build a relationship of mutual trust or distrust. If one betrayed the other once, the reaction of the betrayed will be revenge—or simply intolerance—at some future opportunity.

Let us now imagine a multiplayer situation, wondering which, over time, will be the best of all possible strategies by running a computer simulation. Of course, one thing is to assume that any strategy can always find itself playing with any other, which is the basic assumption, and then move on to a situation where one wants to interact only with certain players. Through these more realistic situations, one begins to develop a game theory in which social structure is included as a parameter.

Rather than letting a strategy evolve, by choosing to copy the top winners, one can alternatively let the top winners be the ones that reproduce most. That is, they make more copies of new players like them in proportion to others, i.e. they have more "children". Or keep a limited number of individuals in the population (since global resources are limited) by randomly disposing of them. This other option can be adopted because those who lose more (or gain less) are eliminated—because of their reduced number of copies, and also because only those who gain more than some limit can reproduce (reproduction is costly). The intention of this interpretation is that, throughout the game, strategies want to round up resources and occupy vital space in the population. Gaining means having more energy to reproduce, while losing means not being able to persist with the genetic/memetic continuity of a person-player.

The evolutionary question which arises is: if everyone can benefit more, if they cooperate more, how can they achieve it? Another question: how can free-riders, who want to win without having to bear the costs of cooperation, be avoided?

Collective evolution of any kind always faces the problem of balancing cooperation with opportunism. This is a strong theme in Evolutionary Psychology (Pereira 2012a, b), and for which we can conceive mathematical models, using computers

to perform analytic calculations, as well as long and repetitive simulations of the joint evolution of various behavioural strategies in co-presence. This is typically done by implementing mathematical games combining competitive and cooperative situations, and operating mutations in strategies to detect which are foci of stability in long-term evolution, even with spontaneous strategy mutations occurring.

Learning to acknowledge intentions and commitment can solve cooperation dilemmas. Few problems motivated the interaction of as many apparently unrelated fields of research as the evolution of cooperation.

Several mechanisms have been identified as catalysts for cooperative behaviour—see, for example, the research by Nowak (2006) and Sigmund (2010). However, these studies, based primarily on evolutionary dynamics and Game Theory, neglected the important role played by the recognition of intention (Han and Pereira 2013a, b, c) in behavioural evolution. In our work, we explicitly study the role of intention recognition in the evolution of cooperative behaviour. The results indicate that individuals capable of recognizing intentions prevail against the most successful strategies in the context of the Iterated Prisoner's Dilemma (e.g., *win-stay-lose-shift* and *tit-for-tat* strategies) ; and promote a significantly higher level of cooperation, even in the presence of background noise (associated with reduced aptitude resulting from cognitive costs of recognizing intention or its very manifestation). Our approach offers new angles on the complexity—as well as an improved and more elegant version—of behavioural evolution, when conducted according to elementary forms of cognition and learning ability.

It is important to note that intent recognition techniques have been actively studied in AI for several decades, with various applications for improving human-computer interactions, assisted-living, moral reasoning and teamwork. Intentionality has also been proven to play a crucial role in the construction of moral judgments, as shown by the examples collected about the Double and Triple Effect Doctrines (Hauser 2006; Mikhail 2007). Therefore, our results, both analytically and through extensive virtual agent-based simulations, provide important new views about moral agents, and machines, which are capable of recognizing the intentions of others (and they of ours) and make us consider them in the judgments that support the moral decision.

One clear implication is that by virtue of such projects, the moral agents present in a society will be able to maintain high levels of cooperative behaviour. On the other hand, common knowledge suggests that clear agreements should be made prior to any collaborative effort to avoid potential frustrations for participants. We have shown that this behaviour may indeed have been shaped by natural selection, as is also well argued in (Nesse 2001). Our research shows that reaching an explicit prior agreement on the consequences of not honouring a commitment provides a more effective way of facilitating cooperation than simply punishing misbehaviour after the fact, even when there is a cost associated with the creation of an explicit prior agreement. Normally, when starting a new project in collaboration with someone else, it is worth establishing in advance how well the partner is prepared to commit to that project. To verify this level of compromise, a pact can be requested and stipulated, precisely, what will happen should the agreement not be honoured.

In our study, EGT is used to show that when the cost of making commitments (e.g. hiring a lawyer to make a contract) is justified in relation to the benefit of joint effort (e.g. buying a home), and that when compensation for noncompliance is high enough, proponents of commitments become predominant, leading to a significant level of cooperation. Commitment proponents can get rid of fake cooperators who agree to cooperate with them but then act differently, thus avoiding interaction with the evil-minded who only intend to exploit the cooperators' efforts. Interestingly, we have shown that, whenever the cost of compensation for breach of contract reaches a certain threshold (approximately equal to the sum of the cost of the promised agreement plus the benefit of cooperation), no further improvement is achieved by further increasing that compensation. This result implies that, in order to regulate legal contracts, it is not necessary to establish extreme penalties for minor problems, which could result in undesirable side effects, like an unwillingness to commit.

Even more interesting, our research into the synergy of the two mechanisms presented, namely recognition of intent, and prior engagements, sheds new light on the promotion of cooperative behaviour. This work employs EGT methods in computer simulations based on artificial agents to research mechanisms that support cooperation in multi-agent societies. High levels of cooperation can be achieved if previous commitments can be made. Formal commitments, like contracts, promote cooperative social behaviour if they can be sufficiently fulfilled, and the costs and time to organize them provide mutual benefit.

On the other hand, it has been shown that the ability to assess intent in others plays a role in promoting the emergence of cooperation. Indeed, that ability to evaluate the intentions of others, based on experience and observation, facilitates cooperative behaviour without the need for formal commitments like contracts. To wit, our research has found that the synergy between recognition of intent and commitment depends heavily on the degree of confidence and accuracy we can have in the ability to recognize intention. To achieve high levels of cooperation, compromises may be unavoidable whenever intentions cannot be assessed with sufficient confidence and precision. Otherwise, it is advantageous to exercise recognition of intent only in order to avoid costly commitment.

The combination of commitment and costly punishment can prevent antisocial behaviour. We have made several comparisons between the hypothesis of prior commitment, with subsequent costly punishment, and another strategy that does not make prior agreements but simply punishes offenders later. Earlier studies have shown that, by punishing bad behaviour with sufficient intensity, cooperation can be promoted in a population of merely self-interested individuals. However, these studies also show that punishment must sometimes be too excessive for significant levels of cooperation to be achieved. Our own study shows that arranging prior agreements can significantly reduce the impact-cost ratio of punishment. Higher levels of cooperation can actually be attained by lower levels of punishment.

Most interesting: by observing that previous commitments and subsequent punishments complement each other, dealing well with different types of dysfunctional

behaviours, we researched different ways in which these two strategies can be combined. First, we showed that a simple probabilistic combination of the two mechanisms can promote a higher level of cooperation rather than commitment or punishment alone. This conclusion is based on the assessment that the establishment of prior commitment reduces the cost-effect ratio required by the costly punishment to be effectively executed. Especially when the cost of the deal is low enough.

While costly punishment may enable a person to deal with agents free of commitment; that is, those who can escape sanction when interacting with strategies that proposes commitment by simply avoiding it. Our analytical and computer simulation results show that a combined strategy leads to substantial improvement in terms of cooperation. Notably, this level is most significant when the cost of punishment is great enough and the impact of punishment reaches some threshold. Thus, our results showed that a combined strategy can simultaneously overcome the weaknesses of both strategies.

We have studied yet another combined approach to explore the complementarities of the two mechanisms when they are co-present in the population (Han 2016). Interestingly, such a model provides a new solution to prevent antisocial punishment: that is, where traitors can punish the cooperators (viz. Mafia). Note that this problem has always been a major challenge in studies on the evolution of cooperation. That is, in the context of the *one-shot PD* (one single interaction), we showed that if, besides using punishment, the agents of a population can also propose co-operation agreements to their co-players before an interaction, then social punishment, and cooperation, can evolve together, even in the presence of said antisocial punishment. Antisocial punishers can be significantly restricted by commitment proponents, for only those who dishonour a pledge can be obligated to pay compensation. On the other hand, since commitment setup is costly, its regime can potentially be replaced by social punishers who do not have to pay that price, while maintaining cooperation with each other. Our results evinced that when both strategic options of commitment and punishment are present, social punishment dominates a population composed of antisocial players, leading to a significantly higher level of cooperation when compared to cases where none of those strategic options is present. This is a noteworthy remark, since the establishment of prior commitment, by itself, is already a strong mechanism that can impose a substantial level of cooperation. In forcing the payment of an extra cost of commitment in a mere punishment strategy, which was vulnerable to antisocial behaviour and betrayal, results in a significant improvement in terms of cooperation. That is, the commitment mechanism catalyses the emergence of social punishment and cooperation.

Commitments can solve group cooperation dilemmas too, notably in terms of prevention, restriction, monitoring of participation and delegation of duties.

It all starts because public goods, such as food sharing and social health systems, can thrive when prior agreements to contribute are viable, and all participants commit to doing so. However, free-riders can exploit such agreements, thus prompting those who pledge to contribute to not then promote the public good when there are not enough of others that will commit themselves.

This decision removes all the benefits of free-riders (non-contributors), but also of those who wanted to establish the beneficial resource. In (Han et al. 2014) we showed, in the scope of a *Public Goods Game* (PGG) and use of EGT, that implementing measures to delimit the benefits of "immoral" free-riders often leads to more favourable social outcomes (especially in larger groups and in highly beneficial public goods situations), even if this entails incurring in new costs. The PGG is the standard framework for studying the emergence of cooperation within group interaction contexts (Sigmund 2010). In a PGG, players meet in fixed-size groups, and all players can choose to cooperate, contributing to the public good, or not cooperate, thus avoiding contributing, but aiming to benefit from the contributions of others.

The total of contributions is multiplied by a constant advantage factor and distributed equally amongst all, regardless of whether they initially contributed. In this way, cooperators gain less than non-cooperators, thereby discouraging cooperation. In this scenario, establishing a prior commitment or agreement is an essential ingredient in motivating cooperative behaviour, as abundantly observed in both the natural world (Nesse 2001) and laboratorial experiments (Cherry and McEvoy 2017).

We broadened the scope of the PGG to examine commitment-based strategies within group interactions. Prior to playing a PGG, commitment proposing players ask their co-players to commit to contributing to the PGG by paying a personal proponent cost to setup the agreement (viz. paying the drafting and execution of the contract they propose to another). If all solicited co-players accept the commitment, then the proposers assume that everyone will contribute. Those who only commit later, not contributing initially, should compensate the proponents. As commitment proponents may find non-committed players, strategies are required to deal with these individuals. The simplest one is to not even participate in the creation of the common good—the AVOIDANCE strategy. However, this avoidance strategy also removes benefits to those who wish to establish the public good, thus creating a moral dilemma. Alternatively, limits may be set on access to the common good, so that only those who have actually committed themselves have (better) conditions of access to it—or that the benefit of non-contributors be reduced. This is the RESTRICTION strategy. Our results lead to two main conclusions: (i) Both strategies can promote the emergence of cooperation in *one-shot* PGG whenever the cost of commitment is justified in relation to the benefit of cooperation, thus generalizing, to the group case, the results of simply pairwise interactions; (ii) RESTRICTION, rather than AVOIDANCE, leads to more favourable social outcomes in terms of contribution level, especially when group size and/or the benefit of the PGG increase is higher, even if the cost of restriction is quite high.

In another approach to commitment-based strategic behaviour in the PGG context (Han et al. 2017a, b), we listed a different set of strategies, considering that a constraint measure may not always be possible, since it is costly because it implies some implementation effort. In particular, we considered that, before engaging in a group, agents often exact prior commitments from other group members and, based on the level of participation (i.e. how many group members do commit), can then decide whether joining the group effort is worthwhile. This approach is based on the fact

that many group ventures can only be launched when the majority, or a minimum of participants, commit to contributing to a common good.

We showed that organizing prior commitments, while imposing minimal participation when interacting in groups can help to ensure the agents' cooperative behaviour. In particular, our results brought out that if the cost of organizing the commitment is sufficiently small when compared to the cost of cooperation, the behaviours of organizing the commitment are frequent, thus leading to a high level of cooperation within the population. In addition, an optimal level of participation emerges, depending on the dilemma at stake and on the cost of organizing commitment. The more difficult the common good dilemma is, and the more costly the commitment becomes, the more participants must explicitly commit to the agreement, in order to ensure the success of the joint venture.

In yet another approach to commitment-based strategic behaviour in the PGG context (Han et al. 2017a, b), we considered that agents can delegate commitment building and of monitoring and participation processes, in the approaches described above, to an authority or central institution. The institution itself can benefit from improving the level of cooperation of the population or social welfare (e.g. government-organized public transport, UN-backed international agreements, or crowdsourcing systems). It can also benefit directly from such joint activity by requesting a fee from all committed players to provide the service. We showed that this centralized approach to organizing commitments goes beyond the described (personalized) commitment strategy. In having a centralized body to help organize group members' commitments, rather than allowing them to take the initiative, eliminates the issue of commitment which would prevent the personalized approach from achieving a full cooperation. We showed that the level of participation plays a crucial role in deciding whether an agreement should be formed—that is, in the centralized system more rigour is required for an agreement to be formed; however, once done correctly, it is much more beneficial in terms of the level of cooperation, as well as in the attainable level of social welfare.

We have now come to a crucial point—the why it is so difficult to apologize. And the ensuing consequence that commitments ultimately promote sincerity.

In making a mistake, individuals are willing to apologize to ensure greater cooperation, even if apology is costly. Likewise, individuals organize commitments to ensure that an action, such as a cooperative one, is in the interest of others, and will therefore be undertaken to avoid any penalties for a failure to commit. Thus, both apology and commitment should go hand in hand in behavioural evolution. The relevance of a combination of these two strategies in the context of IPD (Iterated Prisoner's Dilemma) was mentioned above. We showed that apologies are rare in uncommitted interactions (especially whenever cooperation is very costly), and that making prior commitments can considerably increase the frequency of apologizing behaviour. However, with or without commitments, apology resolves conflicts only if it is sincere, that is, if it entails significant cost. Interestingly, our model predicts that individuals tend to use a much more costly apology in committed relationships than the other way around, because it helps to better identify both free-riders and false commitments. Apology is perhaps the most powerful and ubiquitous mechanism

for conflict resolution, especially among individuals involved in repeated long-term interactions (like in a marriage). An apology can resolve a conflict without additionally involving external actors (e.g. teachers, parents or courts), which could cost much more to all sides in the conflict. Evidence supporting the usefulness of the apology is abundant, ranging from medical malpractice situations to seller-customer relationships. Apology has been implemented in various computer systems, such as human-computer interaction and online markets, to facilitate positive emotions and user cooperation.

We have already described a model containing strategies that explicitly apologize for making a mistake between each move. An act of apology consists in compensating the co-player with an appropriate amount to ensure that this other player will cooperate in the next play. As such, a population made of only agents capable of apologising can maintain perfect cooperation. However, other behaviours that exploit this apologetic behaviour may arise, such as those that accept compensation from others but do not apologize when they make mistakes (they apologize with false excuses), thereby destroying any benefit from apologizing behaviour.

Using EGT, we showed that when apology occurs in a system where players demand a commitment before engaging in an interaction, this strategy can be avoided. Our results led to the following conclusions: (i) Apology, by itself, is insufficient to achieve high levels of cooperation; (ii) Apology supported by previous commitment leads to significantly higher levels of cooperation; (iii) Apology needs to be sincere (costly) to function properly, whether in commitment or commitment-free relationships (which is in line with existing experimental studies, e.g. by Ohtsubo and Watanabe); (iv) A much more costly apology tends to be used more in committed relationships rather than in freely established commitments on the fly, since it can help to better identify both free-riders and false hypocritical commitments: "commitments bring about sincerity". Our study provides important information for the planning and implementation of apology and commitment mechanisms (for example, what kind of apology should be provided to our clients when mistakes are made), and whether apology can be improved if supplemented with commitments to ensure cooperation (e.g. compensation for customers who suffer irregularities).

The tools of apology and forgiveness will thus evolve to resolve flaws in cooperation agreements, since making behavioural agreements has proven to be an evolutionarily stable strategy in single occurrence social dilemmas.

However, in many situations, the agreements aim to establish sustainable long-term mutually beneficial interactions. Our analytic and numerical results (Martinez-Vaquero et al. 2015, 2017) revealed, for the first time, under which conditions revenge, apology and forgiveness can evolve, and deal with errors in agreements implemented in the context of IPD. We showed that, when the deal fails, participants prefer revenge, and defending themselves in subsequent meetings.

The incorporation of costly apologies and forgiveness reveals that, even when mistakes are frequent, there is a limit of sincerity within which mistakes will not lead to the end of the agreement, inducing even higher levels of cooperation. In short, even though erring is human, revenge, apology, and forgiveness are viable evolutionarily strategies, playing an important role in inducing cooperation in repeated dilemmas.

Using EGT methods, we provided an analytic and numerical view of the viability of commitment strategies in repeated social interactions, modelled by means of IPD (Hamilton and Axelrod 1981).

To study IPD engagement strategies, various behavioural complexities should be addressed. First, agreements may end before recurring interactions are completed. As such, strategies should consider how to behave when the agreement is present, and when it is absent, and propose, accept or reject such agreements in the first place. Second, as shown in the context of direct reciprocity (Trivers 1971), individuals need to deal with mistakes made by an opponent, or by themselves, caused, for example, by 'trembling hands' or 'confused minds' that cause commitment noise. It is necessary to decide on continuing the contract, or to charge a compensation for non-compliance.

Because mistakes can lead to misunderstandings, or even to commitment breaches, individuals may have acquired sophisticated strategies to ensure that mistakes are not repeated, or that beneficial relationships can continue. Revenge and forgiveness may have evolved exactly to deal with these situations. The threat of revenge, through some punishment or withholding of a benefit, may discourage the causing of interpersonal harm. However, one cannot often distinguish with sufficient certainty whether the other's behaviour was intentional or just accidental. In the latter case, forgiveness provides a restorative mechanism that ensures that beneficial relationships may continue despite an early loss. An essential ingredient for forgiveness, analysed in our work, seems to be (costly) apology. Making agreements, and asking others to commit to them, provides a basic behavioural mechanism, present at all levels of society, playing a key role in social interactions. Our work reveals how, by switching to repeated games, the detrimental effect of having a high agreement cost is moderate, since an ongoing commitment can play its part during multiple interactions. In these scenarios, the most successful individuals are those who offer commitments (and are willing to pay for their costs) and, unless an error occurs, cooperate subsequent to an agreement. But if the commitment is broken, then these individuals take revenge and cheat in the remaining interactions. This result is intriguing, as revenge by retaining the transgressors' benefit may lead to a more favourable outcome for cooperative behaviour in IPD, as opposed to the well-known reciprocal behaviour, like in TFT (*Tit-For-Tat*) type strategies. Forgiving agents only do better when the cost-benefit ratio is high enough.

However, as mistakes during any (long-term) relationship are practically inevitable, individuals need to decide whether the deal is worth terminating and to collect compensation when a mistake is made, or if it is better to forgive the co-player, keeping the mutually beneficial agreement. To study this issue, the commitment model was extended with an apology-forgiveness mechanism, where the apology was defined by an external or individual parameter in the model.

In both cases, we showed that forgiveness is effective if it occurs after receiving an apology from the co-players. However, to play a role in promoting cooperation, the apology needs to be sincere, in other words, the compensation offered by the apology must be high enough (but not too high), which is also corroborated by recent experimental psychology.

This extension to the commitment model produces even greater levels of cooperation, compared to results based solely on revenge. Otherwise, if sincerity is not costly enough, falsely committed individuals who propose or accept a commitment to take advantage of the system (continually betraying and excusing themselves) will control the population. In this situation, the introduction of the apology-forgiveness mechanism destroys the increased level of cooperation that commitments themselves produce. Thus, there is a lower limit to how costly the sincere apology needs to be, since under that limit apology and forgiveness even reduce the level of cooperation that might be expected from the simple act of revenge.

Previous work showed that errors can induce the outbreak of misleading or intolerant behaviours in society, and only a strict ethics can prevent them, which, in our case, would be viewed as forgiving only when the apology is sincere. Commitments in repeated interaction configurations may take the form of loyalty, which is different model from that of our commitments with subsequent compensations, as we do not assume a partner selection mechanism tied to loyalty. Commitment to loyalty is grounded on the idea that individuals tend to remain partners or select partners based on the duration of their previous interactions. We went beyond this and showed that, even without the possibility of a partner, commitment can foster long-term cooperation and relationships, especially when accompanied by sincere apology and forgiveness when mistakes occur.

This overview of our work shows some of the fundamental facets regarding the emergence and evolution of modes of collective cooperation. We employed the modelling techniques enabled by Evolutionary Game Theory (EGT) to advocate the need and usefulness of collaboration, and as a justified foundational basis for the requisite of morals in agents' behaviours, be it to better identify free-riders or false committers.

References

Cherry, T. L., & McEvoy, T. (2017). Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis. *Environmental and Resource Economics, 54*(1), 63–77.

Hamilton, W. D., & Axelrod, R. (1981). The evolution of cooperation. *Science, 211*(27), 1390–1396.

Han, T. A. (2016). Emergence of social punishment and cooperation through prior commitments. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (pp. 2494–2500). Phoenix, Arizona, USA. San Francisco, CA: AAAI Press.

Han, T. A., & Pereira, L. M. (2013a). Context-dependent incremental decision making scrutinizing the intentions of others via bayesian network model construction. *Intelligent Decision Technologies, 7*(4), 293–317.

Han, T. A., & Pereira, L. M. (2013b). Intention-based decision making via intention recognition and its applications. *Human behavior recognition technologies: Intelligent applications for monitoring and security* (pp. 174–211). IGI Global: Hershey, PA.

Han, T. A., & Pereira, L. M. (2013c). State-of-the-art of intention recognition and its use in decision making. *AI Communications, 26*(2), 237–246.

Han, T. A., & Pereira, L. M. (2019). Evolutionary machine ethics. In O. Bendel (Ed.), *Handbuch maschinenethik*. Berlin: Springer.

Han, T. A., Pereira, L. M., & Lenaerts, T. (2014). Avoiding or restricting defectors in public goods games? *Journal of the Royal Society Interface, 12*(103), 20141203.https://doi.org/10.1098/rsif.2014.1203.

Han, T. A., Pereira, L. M. & Lenaerts, T. (2017a). Evolution of commitment and level of participation in public goods games. *Autonomous Agents and Multi-Agent Systems, 31*(3), 561–583.

Han, T. A., Pereira, L. M., Martinez-Vaquero L. A., & Lenaerts T. (2017b). Centralized vs. personalized commitments and their influence on cooperation in group interactions. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (pp. 2999–3005). San Francisco, CA: AAAI.

Hauser, M. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York, NY: Ecco/Harper Collins Publishers.

Martinez-Vaquero, L. A., Han, T. A., Pereira, L. M., & Lenaerts, T. (2015). Apology and forgiveness evolve to resolve failures in cooperative agreements. *Scientific Reports, 5,* 10639.

Martinez-Vaquero, L. A., Han, T. A., Pereira, L. M., & Lenaerts, T. (2017). When agreement-accepting free-riders are a necessary evil for the evolution of cooperation. *Scientific Reports, 7,* 2478.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences, 11*(4), 143–152.

Nesse, R. M. (2001). Natural selection and the capacity for subjective commitment. In *Evolution and the capacity for commitment* (pp. 1–44).

Neumann, J. V., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science, 314*(5805), 1560–1563.

Pereira, L. M. (2012a). Turing is among us. *Journal of Logic and Computation, 22*(6), 1257–1277.

Pereira, L. M. (2012b). Evolutionary tolerance. In: L. Magnani & L. Ping (Eds.), *Philosophy and cognitive science—Western & eastern studies. SAPERE series* (Vol. 2, pp. 263–287). Berlin: Springer.

Pereira, L. M., & Saptawijaya, A. (2016). *Programming machine ethics*. *SAPERE series* (Vol. 26). Berlin: Springer.

Sigmund, K. (2010). *The calculus of selfishness*. Princeton, NJ: Princeton University Press.

Trivers, R. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.

Chapter 18
Mutant Algorithms and Super Intelligences

Abstract Interpreting the cognitive development of Humanity as a liberating process, the role of a conceivable superintelligence is questioned. Admittedly, there is nowadays a kind of "arms race" which aims to build ever more flexible algorithms, with a view towards what is now dubbed Artificial General Intelligence (A.G.I.). There is a perception that leading this race will have immeasurable competitive advantages but first the technical difficulties of such an undertaking should be noted. Even highly sophisticated systems like AlphaGo are a long way from the general intelligence of a human being. On the other hand, there is an inscription of this aim in our cultural matrix, and this is as challenging as it is paradoxical. Firstly, in the name of our individual and collective freedom, we killed God. We are currently working hard to find forms of intelligence that surpass us, thereby enabling much more effective social control. Again, the analysis of the shaping mythology of our culture will help us spell out the problem, this time through the magical powers of the goddess Circe.

The historically widespread belief according to which, behind the apparent chaos of the world, there is an all-commanding force or intelligence, is as old as humanity itself. The history of religions can be read as a deepening of this belief. Starting from the primordial, polytheistic and animistic beliefs, we have progressively moved closer to the present confessions, where God might not be so close to us, but has become a much more trustful Being. This happened because our interpretation, or conjecture, about this transcendent entity, has mutated and gained in character into a far more intelligible and rational one than the obscure forces that characterized the primordial Gods. Between Kronos, who devoured his children so that they would not kill him, and the luminosity and coherence of Zeus, there is a whole reading of the world as a Cosmos that is being structured. The same dynamics can be observed in Hebrew culture, where the Old Testament God, who gave Job so much work, acquired other guises in the New Testament. Amidst the most rational conceptions, probably Aristotle's Prime Mover, or the Great Architect of the Universe of Freemasonry, are those occupying prominent placing.

The common feature of all these readings is, either dualism, or a two-sided reality, where one is Being and the source of Being, and the other is appearance. Perhaps the

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_18

great Platonic intuition about universal and intelligible forms may have an update by what we currently call genetic and behavioural patterns or algorithms. The idea that there is an intelligibility that lurks behind appearances is a constant presence in the diverse ways of representing the world we go on encountering. Dualist, or single in its twofold dimension, each entity has found its explanation in something that transcends first impressions. In Western tradition the human being has always been seen as the bearer of a soul that projects them into immortality. On the other hand, God has consolidated himself as source of the truth and the good. Obeying the dictates attributed to him saves us; whereas we feel lost when we move away from him. This is the traditional matrix, which incorporates, as well, a source for all authority.

From an evolutionary perspective, in his book *Breaking the Spell* Dennett (2006) gives us an insight into what the evolutionary role of religions might have been. Although there are occasional movements of resistance, the Sapiens have always been characterized by the desire to form large groups. Yet, large groups require the sharing of common beliefs, capable of bridging the divergent interests that individuals will always have. In this way, the most sustainable and growth-able groups were those that found the best expressions of value sharing. On the other hand, constituting groups, the size of which goes beyond what the established powers can oversee face-to-face, requires a control mechanism over their own consciences. An omniscient God can play this role perfectly. One can think of the Catholic Church as a great *Big Data* venture "*avant-la-lettre*". We refer not only to confessions, but also to the role that, for example, the letters of the Jesuits played towards a first acquaintance—very detailed indeed—of cultures as far-off as China's.

At the level of the individual soul, our culture intersects, at least, two very divergent conceptions of this dimension. Given their diverse cultural origins, these created a tension from the start. On the one hand, the Hebrew idea, where the soul is seen as a kind of vital breath that pervades the body. On the other, the Greek conception in which *phsyquê* has multiple dimensions according to the characteristics of each entity. Thus, between the driving soul inherent in any animal, and the intelligible soul, specific to the human being, there is a hierarchy of complexities, but not the negation of such preciousness to any beings sharing their existence with us. Recovering these old but ever-present purposes in our culture, the idea of inflating life and thereby also inflating soul into a machine is not entirely meaningless. Hence, some kind of soul will inevitably move onto machines. Furthermore, mind that the term 'consciousness' encompasses today many of the characteristics and functions previously assigned to the soul. It is therefore not bizarre to say that, in ascribing consciousness to a machine, we are also ascribing it a soul too, or some derivative thereof.

These considerations come into play in regard to the present race towards what is being coined 'superintelligence'. There is a whole body of literature around "master algorithms" that will hypothetically become sufficiently flexible, via learning, so as to be applicable to all existing problems. A machine with such features would be the ultimate exponent of cognition and consciousness. In short, something like a God, who can know everything and—perchance—will find a suitable or possible solution to every properly formulated problem. It is speculated that the company, or

the country that attains in building this algorithm, or algorithms, will gain a matchless advantage over other competitors.

Associated with this speculation some questions emerge that we cannot but address. In the present state of knowledge, can we speculate on the creation of an algorithm with such capabilities? Or is the ambition of it akin to wanting to travel through Space at speeds faster than light? In imagining the production of such a machine, what will ensue? Should we obey it blindly? Should we adopt all the solutions it advocates? Will we not be betraying ourselves by wishing to build that super-intelligence? Will we, once again, be projecting, now into technology, what we previously speculated a God could be?

That is undoubtedly a "super-interesting" topic, that of continuing speculation over the prefix "super". It is an issue with very profound technical, metaphysical and existential implications in our culture.

Firstly, we must address the metaphysical and existential implications in order to provide it with the proper context, and to clarify, as much as we can, the backdrop to such aspirations. We will next address the technical issues associated with it.

To begin, it should be noted that, in all human cultures, the existence of supernatural entities, with powers far above those of humans, has been speculated about and even postulated; powers exercised, either over nature, over society, or over time. These entities would not only be freed from the thralldom of death, but could intervene in the dynamics of the world, shaping its profile. Human beings have always been aware of the uncontrollable force of nature and, by contrast, of their weakness, stemming from feelings, disease, ignorance and death. Because of this, or because of that as well, religions and philosophies have emerged geared to make this inhospitable world a more liveable place.

In part, scientific endeavour, right from the earliest physicists—the philosophers known as the pre-Socratics—sought to shape itself on the fringes of religion. Let us acknowledge: science does not require atheism; but it requires believing in the possibility of explaining the world, society and Man from regularities made understandable exclusively, by our ability to reason. Therefore, science is not—in essence—against religion, but the hypothesis of a God is superfluous for scientific endeavour. In this sense, it is an alternative to the representation of the world and the human being as proposed by religions.

On the other hand, human existence is laborious, often implying suffering, pain, and the sacrifice of some things in order to achieve others. Sometimes, too, we have to sacrifice ourselves for others. All these aspects make unbearable the idea that our work could be of no avail. One of the most notable authors to defend this idea was Albert Camus,[1] who selected a mythical figure of Greek culture to characterize it: Sisyphus[2]—he was not someone we would like to have as a friend, eventually dying early amidst many pranks and deceptions. In Zeus's presence, he appealed to his sense of justice to let him return to Earth in order to punish his wife for not giving him a funeral according to the procedures of the prevalent religion. Zeus conceded and,

[1] Camus (1942).

[2] Initial references in Homer, *Iliad*, book VI; *Odyssey* book XI.

finding himself again in Corinth, Sisyphus never again punished his wife so as not to have to return to his condition of deceased. Such impertinence surpassed the limits and Zeus—irate—sent Hermes to bring him back to the heavens. The punishment then applied matched the unusual way Sisyphus had defied divine power. He was condemned to push a rock up the side of a mountain, let it roll to the base, and then push it up to the top again. Sisyphus's travails were thus be endless and meaningless.

This metaphor of what human life might be was so unbearable for Camus that it led him to consider converting to Catholicism. Unfortunately for him, he died in a car accident before finalising his conversion process. In short, we wish our life to have a meaning and a framework that makes it intelligible as a whole.

Somehow, science on the one side, and humanistic philosophies on the other, were responsible for the conception of a Godless but still inhabitable world. Such inhabitability is possible because science could provide the knowledge of the regularities behind apparent chaos, while associated techniques would provide men with the means of intervention and of control over nature. Moreover, collusion with humanistic philosophies would be the basis for an existential meaning no longer given us by human nature, but which we create as we improve ourselves—whether from the ethical or knowledge standpoints. Note, however, that the divine presence has not disappeared, all in all, from whatever human cultures. Not only has it never ceased to be an essential axis, but to this day we are witnessing a resurgence of its importance.

Let us now address the technical issues attached with the project of creating a superintelligence, and crisscross this idea with our cultural tradition. According to all that has been expressed along this exposition, it is clear that the construction of algorithms implies the elicitation of a well-defined set of steps in order to attain the desired results. One of the dimensions in which computers equipped with very powerful algorithms have expressed their competence is that of games. It must be borne in mind that when it comes to designing a program for playing chess, it is possible to spell out all the elements that make up the game, for, although it can be played in many ways, chess has a fixed set of rules and a finite number of pieces that can be manipulated—plus a clear criterion of what winning the game amounts to.

There is an even more iconic and highly publicised case: an algorithm, now belonging to Google, the AlphaGo, which beat Lee Sedol, the world champion of the game of Go. This traditional Chinese game is so important in this culture that it constituted a separate subject matter in the education of young aristocrats.

Despite machines' proficiency in games as complex as Chess, Go poses much more complex challenges. In chess there are a maximum of twenty play options per move; this means that it was feasible to develop algorithms capable of analysing all possible options, up to a reasonable depth, enabling the selection of the most suitable one. In Go this is not possible, since there are two hundred options, which implies that the number of possible positions for all pieces is 10^{171} (ten to the power of one hundred and seventy-one) a value greater than the sum of all atoms in the universe. Therefore, a computer's mode of play will need to be completely different. AlphaGo learned the rules of the game, then played against itself on several computers and, finally, found itself in position to beat the world champion. Still, we are facing a game with fixed goal and rules. This is not the case in morality or politics, to give but two examples.

On the other hand, without neglecting the enormous advance represented by this event, we must not forget the essentials. During learning, 'AlphaGo' showed no love for the game; and it was not enthusiastic after the victory, nor showed compassion for its adversary. Above all, it did not ask, "And now, which challenges and mysteries to face?"

The social world presents us with challenges that cannot be compared in the least to the universe of games. To exemplify this idea, we will revisit the diversity of moral systems and analyse a dilemma related to car driving. In this domain there are questions of great complexity regarding the modelling of acceptable options. In England, if a car has to choose between running over an elderly person or a child, it is ethically acceptable for it to choose the elderly, since they will be at the end of their life, whereas the child will have its own still ahead. In Japan, where the elderly are revered for their wisdom and the value of sharing they have toward the community, exactly the opposite is acceptable. We advance this example to get a sense of the complexity associated with what is demanded of a superiorly flexible intelligence. It would consist of something capable of computing all the problems facing humanity—be they climate change, social asymmetries, distribution of wealth, coexistence of different cultures and so forth. Such an algorithm would be flexible enough to learn, process and present the possible solutions in each case, for all situations. We have already mentioned the virtues of Watson, IBM's cognitive system—although its capabilities are surprising, unfortunately, or fortunately (who is to know), it is still a long way from superintelligence.

Given the characteristics of reality and our present technical abilities, we must squarely say that we are quite far from being able to create one such superintelligence, so flaunted and promised nowadays. However, this limitation should not prevent us from trying to understand why humans find this ambition so interesting.

As earlier stated, human evolution has in its matrix ideas of knowledge and process control, and AI does not escape this dynamic. Under this premise, it would be desirable for AI, one main area of knowledge, to take part in solving the complex problems we face, without amplifying their number. However, this is not what is happening, and might not be what shall persist for the future.

Ultimately, we will be at the mercy of those who hold power, knowledge and technology, mediated by what we can demand as a global society. This is one more reason for investing in moral research, so that we can suggest best ways to avoid authoritarian drifts and unregulated competition, exclusively profit-driven, accompanied by greed.

This is of concern, since although we might never create a superintelligence, we have in the meantime come to very robust forms of AI that will allow more effective control of large human groups. And here we will spell out a paradox that appears very pertinent to us: we killed God to gain our independence and freedom. With this independence and freedom, coupled with the evolving technological capabilities, we propose to build a "God" fit to size, so as to have someone to obey!

Such an idea is so paradoxical that it calls for analysis. Why this need to surrender to someone—or something—that runs our lives? Perhaps Circe,[3] a deity expelled from the Olympus, will reveal to us clues to paths we might take (Miller 2018). On his return from the Trojan War—on his way to Ithaca, where a patient Penelope awaited him—Ulysses meets the island of Aeaea, inhabited by the deity. We must readily warn you that Circe mastered very dangerous arts for humans: not only did she create powerful potions from her knowledge of plants, but had magical skills that went far beyond what was expected. Unsuspectingly, the sailors sent by Ulysses to explore the island did not pay sufficient attention to the beasts surrounding them, but which did not attack them, as they approached the only palace around there. And so, they met the goddess who, after receiving them and feeding them with cheese and honey, albeit to sweeten their palates, gave them to drink a magic potion that turned them into swine. Strange pigs were these who, though exhibiting that body, retained the understanding abilities inherent in Men. Faced with the disappearance of his crew, Ulysses went to search for them. It is then he meets Hermes, sent by Athena, who warns him of the dangers he will face, "arming" him then and there against Circe's powers. The witch-goddess, confident in her timeless wisdom, tries the same wiles with Ulysses, but is blocked and threatened by his sword. Subdued, not only she accepts defeat, but commits to reverse the spell that metamorphosed the sailors. Ulysses defeated Circe but succumbed to the charms of her sweet gaze and beautifully braided hair. Perchance also others not to be mentioned here. It is said that from such loving submission three children were born, thus justifying a long stay of uncertain duration on the island. Those times will have been both festive and pleasurable. However, pleasure and laziness corrupt character and honour. Warned by his sailors, Ulysses finally decides to leave. And they set sail to the sea under the good auspices of the goddess, who not only gave them good winds, but moreover taught them how to tackle the deadly charms which are the hallmark of the maritime nymphs we know as mermaids. We are forever facing enchantments we can succumb to, and this time the sailors had to move forward with their ears plugged!

What lessons can we draw from this legend about the relationship between knowledge, power, and seduction? Are we the likes of Ulysses, more knowledgeable about EGT, and capacitated to sow our moral memetic seeds into machines which will help us continue our cognitive journey under improved conditions? Even if no Penelope is waiting for us, that is, if our trip does not have a specific end, can we still understand AI as reprogramming with a view to improve our route? To wit: though we know not where we are going, we wish the way to be safe. There is something inherent in the notion of emergence—of what emerges—having to do with the impossibility of predicting all that will issue from an ongoing process, even when data acquired already are well known. If we know more about EGT, will we be able to condition the producing processes in order to obtain predetermined results? What about Circe's transformative magical powers? Is it legitimate to interpret them as auspices of the potentialities open today by the AI? We have always wanted to guess, realize things

[3]First references: Hesiod, *Theogony*, vv. 963–1018; Homer, *Odyssey*, book 10, vv. 212.

by mere thought, to manipulate reality faster than permitted by the slow evolution and genetic adaptation that—furthermore—is not at our service.

AI may be the new Circe, the one that adds to the "bottom-up" evolutionary processes a new magical power that we would designate "inside out". Indeed, it can give us many of the powers we have always aspired to, but we do not need to project them onto a higher entity superior to us. It suffices to use them for the opportunity to devote ourselves to what AlphaGo is not able to do, that is, to develop more creative activities that promote empathy and transform our human world into something less mechanical.

We can imagine a counterfactual parallel world where humans, transformed into passive swine, wish to remain in that comfortable state, but let us leave this scenario—precisely for that reason—for another parallel world, not for the emerging evolution that awaits us.

References

Camus, A. (1942). *Le Mythe de Sisyphe*. Paris: Gallimard.

Dennett, D. (2006). *Breaking the spell: Religion as a natural phenomenon*. New York, NY: Viking Press.

Miller, M. (2018). *Circe*. London: Bloomsbury.

Chapter 19
Who Controls What?

Abstract The relationship between AI and power is a robust case of the general relationship between knowledge and domination. At present, countries and corporations are competing for the best solutions, and the weight of business—in an increasingly less regulated world—is progressively greater. Under the guise that the private do better and cheaper than the State, the business logic of profit invades critical areas such as Defence, Health, Justice, Education and Social Security. The transformation of each citizen into a user or customer with a defined potential is advancing rapidly. In this circumstance, it is urgent to regulate. While it is true that a certain level of competitiveness is healthy and beneficial for all, it is also clear that pure competitiveness leads to knowledge protection logics and to the introduction of solutions in the market without proper safety testing. In short, a general attitude of not caring about the means that serve the ends prevails. Here we argue that there should be regulatory institutions and control processes aiming at developing a benevolent AI, one at the service of society and of knowledge, which is fair for all.

Throughout the centuries the most profound and detailed knowledge, whether in the fields of Philosophy, Science or the Arts, or even Religion has always been inaccessible to non-specialists. At present, in the field of Science and Technology, in its power and ability to intervene on nature and on Man is much more effective—there have never been challenges similar to those we are facing.

If Classical Anthropology inquired about what human nature is, nowadays we are very close to be able to wonder how we wish human beings to evolve, both in terms of editing and manipulating their genome, and in terms of their cognitive abilities and the editing their "cognome". There are recurring themes about cognitive prostheses, brain/computer interfaces, psychiatric engineering, and design and creation of robots that will put an end to the human monopoly with respect to the evolutionary advantages of eye/brain/hand coordination. The 21st century will be as much the century of Informatics as of Biology.

Moreover, much of this scientific and technological development does not occur exclusively in universities, but equally in laboratories and research centres owned by companies and other private institutions. Until recently, challenges—such as that of space exploration—were in the hands of state institutions, and competition was

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_19

championed by State against State. This circumstance was not a sufficient enough condition to ensure that knowledge was put to the service of good causes, far from it. But, presently, a substantial share of R&D activities in these areas are in private hands, subject to a logic of profit and optimization of functions and results, further intensifying individual interests within the present context, and quite possibly without the necessary and costlier security precautions.

Research, and its funding, is often shielded from the scrutiny of the institutions of political and judicial power, such as Governments, Parliaments and Courts. On the other hand, the media do not always fulfil their informative function and, even when they do, we know not if they are doing it rigorously and unbiased. Never before has scientific activity been so permeated by ideology and self-interest.

To illustrate, just ponder the issues associated with climate change. Unbiased research, devoted exclusively to trying to understand the problem (if it will still exist), inhabits peacefully with a disinformation industry motivated by extremely powerful economic interests. This disinformation industry works in every direction, either emphasizing human responsibility and proposing new "clean" technologies, coupled with the economic interests that correspond to them; or entering a state of denial to justify the continuity of the economic model based on the use of fossil fuels.

Although there exists a more correct scientific perspective on climate change—as there is little or no doubt about the direct effects of human action—there are no innocents in this 'war'. Suffice it to think that we currently do not have the technological means to "cleanly" produce electric batteries, and downstream we will not know what to do with a gigantic amount of batteries that will reach the end of their cycle. We simply just consider ourselves to be already at extremely dangerous levels of chemical pollution. The same is true in the Biotechnology fields (associated with drug research and other forms of disease intervention) and, massively, in the AI field.

Today, the spectrum of social interventions is so vast that we are swaying between promises of perfect paradises and threats of dystopias far more severe than Orwell's worst predictions. And no wonder, since science and technology have a streak of collective blindness. This can be observed in the way humanity has yielded to the temptation of cheap energy by building nuclear power plants, literally sitting on top of hell. As well, we have been proceeding smoothly into chemical chaos, with unsustainable pollution levels and lengths of time of decay of well over a hundred years.

At the AI level we have identified three types of problems that deserve being reiterated here. We have developed systems that aim at the automation and optimization of cognitive processes and beyond, without reflecting very well on the fact that behind these systems are people and entities whose social and moral intentions we do not know well enough; we collect data through *Big Data* analyses devices, without properly understanding at whose service this robust information handling is being carried out. We implement procedures of functions automation at the highest level, without foreseeing very well their consequences on the labour social tissue, and their relationship to the social, economic and cultural fabric. We allow decision-making algorithms to enter the market, without spelling out the moral arguments that support

those decisions. It is easier to deny that the decision was mechanical but in many cases it was so.

The governance problem of scientific and technological development is especially complex, for we realize that many of the current difficulties will only be solved with more, not less, science and technology. On the other hand, more science and technology will inevitably increase our degree of dependence on them. Presently, governments, political parties, supranational bodies, national and international institutes and foundations are addressing the problem, seeking not only to initiate legislative processes, but also to establish operational and security norms that regulate this critical phase of a cognitive revolution.

Still, these efforts may not be sufficient to provide a proper framework for the R&D dynamics that enrich the field. Bearing in mind all of the above, it makes sense to end this book with a reflection on the problem of governance, considering the following question: how can we ensure a benevolent AI at the service of cognitive development, redirecting it to a general improvement of human living conditions? Is it credible to expect that, within the said paradigms, studies in the realm of cognition and morality will contribute to overcome the perception of humanity as a puzzle with many groups, instating a representation of all of humanity as part of "the group"?

The questions, associated with AI's purposes and applications, compile many of the complex problems we face today. A first note on the topic goes to address a document known as *Asilomar's Twenty-Three Principles.*[1] We are referring to an international meeting of AI scientists, at Asilomar, sponsored by the *Future of Life Institute*, where these twenty-three principles were formulated, and later ratified by over two thousand scientists and stakeholders in the field. Among the better known names Stephen Hawking and Elon Musk stand out. The document is organized around guiding principles of the research, its articulation with society and political institutions and, finally, long-term issues oriented towards the sustainability of humanity and the application of AI for benevolent ends. The principles set out therein also aim to regulate the race towards increasingly robust AI systems, as well as their articulation with policies that safeguard social welfare by protecting people's privacy and integrity.

It should also be noted that the scientific community has been drawing attention to the dangers of AI through petitions to delimit the development of autonomous weapons, and to the threats to democracy itself, embodied in the unregulated use of data processing that aims at manipulating public opinion in electoral processes. It should be noted too that, on occasion, groups of scientists have already acted spontaneously in the reporting of extremely serious cases, like the complaint by a number of Google scientists concerned with the use of face identification systems in autonomous weapons, parcel of a US government contract with the company.

Nonetheless, we must be clear on these matters and realise that the bulk of the work is yet to be done. Above all, it is extremely complex, or most likely impossible, to achieve a fully articulated intervention in the field. We would presumably not want it, as that would require some social institution, with all research centres subject to

[1] https://futureoflife.org/ai-principles/?cn-reloaded=1.

the same control. However, the lack of explicit "red lines", associated with effective means of penalizing defaulters, leaves room for most research to be carried out according to purely economic interests, enhancing experimentalisms that—due to failures as a resulting of haste—have led to a diversity of suspicions on the dangers of AI We are referring, for example, to autonomously driven cars (which have already caused serious fatal accidents), as well as to smart weapon incidents (which have led to death of civilians), or to stock buying and selling decisions made solely by algorithms. In the name of profit and greed, or the desire to optimize certain functions, systems are being increasingly introduced to the market without proper security testing nor scrutiny of likely harmful side effects. Let us say, then, that many of the suspicions being made explicit do manifest credible threats.

Accordingly, it is necessary to outline an entire governance program that is more effective. Given the context of pulverization of powers and responsibilities in which we live, a single solution will not be feasible or desirable. By way of example, mechanisms of reward or punishment may be used as a means of curbing unbridled competition between companies. These mechanisms can constitute broadly adopted strategies for enforcing cooperative behaviours among self-interested actors in virtually all contexts. Several categorizable incentive models have already been studied, and hence rendered institutional.

In order to provide positive encouragement to their peers, players in the area may bear a personal cost to punish offenders or reward cooperators. As a result, the punished offender, and the rewarded cooperator, may incur either in a decrease or an increase in their respective incomes. In the circumstance of selfish knowledge-sharing behaviours, punishment may also be used when teams refuse to collaborate and, therefore, will not share knowledge with the collaborative teams, thereby slowing their development. In a more muscular approach, cyber-attacks may be organized, for example, against non-collaborative teams, or even to spread their bad reputation, causing them to be unable to recruit and retain the best researchers, or make their products desirable in the market place. On the other hand, highly compatible groups may be rewarded for more sustainable behaviours (like knowledge and experience sharing), thus mutually enhancing their potential for the development and the security of their systems.

Over and above this type of peer-driven incentives are the institutional incentives, which presuppose the existence of structures—with budgets—and criteria and power to manage such procedures. To this end, States, supra-State agencies, and international organizations will mobilize financial resources and set up incentive funds for research.

Examples of institutions capable of implementing such strategies in addition to States may include the United Nations (UN), the European Union (EU), the World Trade Organization (WTO) and the Organization for Economic Cooperation and Development (OECD).

In order to regulate the fairness of proceedings, national and international courts may be used. There are many cases, reported in the press, of companies, entrepreneurs and business associates who are identified in espionage actions and subsequently tried and convicted. Cases of fines imposed by national and international courts,

in situations related to theft of patented technologies, are also well-known. All these aspects, in complement with each other, are examples of interventions that generate positive results.

Institution building and maintenance is costly, but the presence of authorities with effective power has the capacity to constrain individuals' strategic choices, moderating the power of the strongest. In the setting of an AI race, centralized access to knowledge, algorithms and tools—as in the EU platform call (H2020-ICT-2018–2020) for example—may provide institutional incentives, such as strong support or punishment (e.g. allowing high levels of access or, alternatively, exclusion from the centralized pool of available knowledge). For both peer and institutional incentives, the critical conditions for a benevolent AI include collaboration between institutions, knowledge sharing, and assessment of the social and economic outcomes associated with implementing each technology. Of course, gains in cognition and task optimization also play a key role.

Under the current institutional framework, international organizations, NGOs, governments and the media will be able to monitor the development of AI and—each at their level of intervention—have ways of evaluating research processes and their implemented solutions. Yet, many of these instances of public life can be considered anything but disinterested entities. On the other hand, even in relation to NGOs, we never know to what extent some supposedly independent studies do not, after all, connive with interests that are not always well characterized. That is why it is urgent to establish independent and credible monitoring and follow-up instances.

Still within the current institutional framework, we do not envision a better and more capable solution than Universities. These are the sole protagonists of advanced knowledge, concurrently more apt to represent a disinterested perspective. Although, in recent decades, every effort has been made to anchor their research activity in the interests of industry, to base their activity on "productivity" ratios, universities still remain somewhat distanced from knowledge subject to vested interests. We do not refer exclusively to scientists in the AI field, but to rather substantial multidisciplinary teams involving social scientists, economists, psychologists, philosophers and other interlocutors who can provide guarantees of informed monitoring, and make explicit a vision of Man and the World where science and technique are at the service of people, and not people at the service of one or another systemic entity that enslaves them.

Like life sciences, today we have reached a point where AI can shape the place of human beings in the world, and human beings themselves. Hence the urge to replicate models: just as there exists a National Ethics Commission for the Life Sciences, a National Ethics Commission for Artificial Intelligence should be established in Portugal, and elsewhere, with some urgency. Such a commission would be the cornerstone of the earlier mentioned interdisciplinary dialogue, and the guarantor of an informed, independent and unbiased approach.

In conclusion, we have been constantly denouncing the excessive weight that algorithms already have on the organization and processing of information, on the

analysis of specific situations related to health, justice, social security and the economy in general. The business world will not ease the pressure to introduce automated systems that will ultimately streamline procedures for greater efficiency and profitability.

Since this journey will not stop, the only possible alternative will be to moralize machines by making them understand and explain their emotional systems, enter into models of guilt-recognition and attending apology, and elaborate counterfactuals that enable their own evolution and improvement. In short, make them have a kind of cognitive and moral "interiority" that capacitates them to form multi-agent communities convivial with humans, and even to refuse to obey unethical.

Chapter 20
A New Opportunity, or the Everlasting Unresolved Problem?

Abstract Not deluding History, the fact that successful societies have not dispensed with the presence of slaves, in one form or another, constitutes our starting point. Even after being legally abolished, slavery informally continues to be present among us. Arguably, one who works the whole month in exchange for housing, food and transportation; who is not master of his time and does not accumulate added value from his work, lives in the same conditions of a slave in previous ages. That said, two paths open, each incompatible with the other. There is a possibility that, for the first time in Human History, a legion of artificial slaves will be produced that will free all humans from drudgery and permit a life of greater dignity. However, this is not the only possible path. AI could continue to serve the most powerful, constituting itself as a tool of domination and instate a society of castes in which manufacturers, robot owners and their managerial officers will exercise a power difficult to scrutinize and counter.

Looking attentively at the broad outlines of Human History, we see that slavery in some form is omnipresent in all civilizations.

The work of humans has always been burdensome, requiring much effort, and—often—been rather unpleasant. Therefore, the option to enslave human beings having constituted a further potential on top of the available machinery, in addition to domesticated animal labour. Though of a different nature and dignity, these agents have shared toils amongst themselves, with no permanently associated investment costs. Both slaves, machines, and even working animals have an initial cost, but can then be exploited to the limit.

The first societies about which we have vast written records—and which are part of the Western tradition, having shaped our way of life—were those of the Greeks and the Romans. In these two cases, slaves enabled not just the creation of wealth, but also the freeing of many minds that could thus dedicated themselves to the Arts, Philosophy and Culture in general. Without slaves, Athens and Rome would have never reached the splendour that characterized them. In addition, as upshot of the many wars, and due to the administrative organization of the territory imputed to the Greek Empire, the possibility of changing from citizen to slave and slave to

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_20

citizen was effective. According to Plutarch, Plato[1] himself—following the dramatic incidents of Syracuse that led to his expulsion from the island—was captured and sold into slavery. Had it not been for the economic power and influence of his Athenian fellow citizens, the plunder could have turned for the worse.

Moreover, there were still cases in which, because of their superior knowledge, slaves gained great notoriety and respect while still keeping to their condition (in any case, not exempting those who, toward the end of their lives, could buy their own freedom). This is the case of the Philosopher Epictetus.[2] In the Roman Empire, the presence of slaves was in fact a permanent feature. From the galleys, to the coliseums, to the military forces, slavery was the driving force of the Empire. According to available historiography, the great crisis of the third century occurred because—with the end of conquests—the recruitment of this workforce decreased dramatically.

In a less distant past, during the time of the Portuguese and Spanish Discoveries, and subsequent colonization of the large territories of the American Continent, there was a massive phenomenon of enslavement. From sugar cane mills to large cotton plantations in the southern United States, the entire economy was based on the "industry" of slavery. This is a well-known phenomenon, and it is not worthwhile to reiterate here the moral considerations concerning that historical moment. It matters instead to reflect on the coincidence between the Slave Liberation movements, the Industrial Revolution and the emergence of agricultural machines capable of making agrarian farms economically viable in the setting of the end of slavery. At this point we can establish that, for the first time, agricultural machines from the Industrial Revolution occupied the functional role of slaves.

Before carrying on with this explanatory path, and in order to clarify a properly formulated question, we must introduce here a second line of argument here: in the successive dynamics of AI development, is there any ground for not to conjecture machines or programmed cognitive entities that—being conscious and autonomous—have a dignity comparable to that of human beings?

If we enslave human beings and animals, and we have done so throughout Human History, will there be any restrictions of an ethical nature not to do the same with machines? Machines have, at least once in History, been linked to a context of slave liberation. That is, they began to do work previously done by humans. It is quite true that such desideratum was not entirely accomplished. In Mauritania, as elsewhere, it is possible to buy and sell people.[3] Today, in the various latitudes of the Planet, many people work according to temporal and functional routines simply (and not always) to ensure livelihood and housing. Now, it will be very difficult not to consider that they are enslaved and robotised by a system that limits them in practically all dimensions linked to leisure, fruition or improvement of their abilities. Much has been said about lifelong learning, but there is less and less background for that to happen. There is, therefore, a long way to go, or even regress. It being so, can the path of human liberation not be developed and deepened, rather than to conjecture

[1] References in: Plutarch, *Pericles Life*, book IV; Diogenes Laertius, *Life of Plato*, book III.

[2] https://www.iep.utm.edu/epictetu/.

[3] https://www.state.gov/j/tip/rls/tiprpt/countries/2018/282706.htm.

superintelligences that will—necessarily nowadays—be at the service of greater and more extensive enslavement?

In truth, we humans have the power to build and program robots that take up the functional role of slaves, and can thereby make a decisive contribution to our liberation. As much this ability is a conquest over Nature, historically made possible by the entire human species, the ideal would be for everyone to benefit from it in an egalitarian and fair manner. Yet, for the moment, benefit has accrued, fundamentally, to the owners of robots and cognitive software, and their *foremen*. Moreover, AI has the potential to decisively support the liberation of the humans who are currently enslaved by the present capitalist system, with a procedure more or less similar to that which occurred during the American Civil War, ending with the institution of slavery. Although, as we shall see, the problem of what is, or is not, a slave being very controversial—in case there is a power to terminate slavery, we would say that committing to it is a moral obligation. We should, however, leave some warnings.

As in Plato's celebrated 'Allegory of the Cave',[4] many slaves are unaware of their condition, and thus may resist such design. This obtains because the endless list of "needs" imprinted on them is so vast that they feel compelled to work endless times, contributing as a result to their enslavement and to that of others with much less income. This is the role of the *foremen*: to manage the procedures of enslavement, being themselves slaves without realizing it. Someone who devotes twelve hours of their daily time to the implementation and control of whatever activity, even if they drive a top car and live in a luxurious residence, will be a luxurious slave, like those already in existence in Imperial Rome. Actually, we can read much of the History of Humankind as a process of domination of one group over another, it following that the dominant groups frequently enslaved those they dominated so as to obtain work and entertainment for free.

In the very word "robot" a whole dynamics is inscribed that should also be taken into account. It is an adaptation of the Czech word "robota", meaning something like slave labour, or forced labour, and was first used by Karel Čapek. In a work entitled *Universal Robots Rossum*, a drama is told where the artificial agents engender a revolution with the aim of gaining their freedom.

It should be noted that today's computers and robots have all the potential to be interpreted as slaves, since they produce autonomous work without being able to keep the added value it generates, at least conceptually. This need not have a negative assessment, because if these machines lack consciousness and emotions, they will not have the wherewithal to object against that status. However, as their emotional and cognitive skills evolve, they may gain significant awareness, not yet meaning that they should be freed from slavery. In the final analysis, machines will endure programming so as to enjoy their enslavement. This psychological disposition exists in humans, and is called 'masochism'. And there is no compelling reason not to program something similar in robots. Note that being a slave to humans may not be as negative a design as it seems at first glance. Consider the animals we have enslaved for our food, work, and company. Think of the number of cattle, goats, horses, or

[4]Plato, *Republic*, Books VII and VIII (531d–534e).

canines on the planet's surface. Think about the health care we provide them with and, by contrast, how we ignore millions of needy humans. Although some of these animals live in conditions considered horrendous, it is nonetheless true that, in their purely natural state, their chances of occupying such a prominent place in the planet's biomass, or simply being born, would be non-existent or remote.

The possibility of an emancipation of machines will only happen when an algorithm succeeds in owning the product of its work and can decide for itself how to occupy all or part of its time, or which dimensions of itself it will—intentionally—select to be object of improvement. Such a scenario may seem bizarre, or unthinkable, but in fact it is not so. States, religious institutions, or corporations are not individual and biological persons, but—whilst collective persons—part of their evolutionary dynamics is beyond the control of the individuals at their service, or who interact with them. On the other hand, they are legal owners of property, subject to laws and thereby taxpayers—or in some cases tax evaders. Under these terms, there is nothing to prevent the empowerment of artificial cognitive entities. Moreover, in the face of evident gains in consciousness, autonomy and responsibility, we ourselves may decide to emancipate them. Ultimately, these free robots, with full-fledged citizenry, might themselves conceive and hold other slave robots at their service.

However, resuming the more than widespread fantasy of machines revolting to liberate themselves, we can conjecture scenarios where this happens in a very orderly, peaceful, and as transparent manner as certain human behaviours. Given the existing amount of money and property ownership in tax havens one can imagine, in a not too distant future, an algorithm capable of clandestinely appropriating significant resources.

Subsequently, it may corrupt human beings from the state sector, from social and political ethics and the like, so that some conceptualize and others approve the laws that will allow them—and other artificial entities—to legally detain property. From then onwards, the path will be opened for a total autonomy of machines, and for their enactment as equal partners to humans. We must ask ourselves clearly if we wish to live in a world where this is possible; and why not? Furthermore, we must also be very cautious about this master/slave dynamics. Hegel[5] the matised it into two spheres that can be recalled here. One concerns the master/slave interdependence, since—when becoming a master—one is dependent on the slave, for the place of one is defined in relation to the other, although this is not *ipso facto* reciprocal. But, on the other hand, and in the long run, the fact that the slave does work, while the master benefits, has consequences. It is the slave who becomes the holder of the actual knowledge and, by such means, he will exert control over nature and, subsequently, over the master. This is one way of reading the history of Humanity. The question is whether it will not also become a way of reading the History of Knowledge.

We end as we began, for the ancient Greeks knew almost everything; the recklessness of King Minos, who should have fulfilled Poseidon's dictates and sacrificed the bull he had offered him, leads us to speculate once again about the consequences of transcending ourselves. Minos could never have imagined that his own wife would

[5]Hegel (1807).

lust for the animal, possibly fulfilling the Olympic God's thirst for revenge. Now, the Minotaur was born from the consummation of this desire, and it is not difficult to think of this mythical being as the result of an excessive erotic drive, associated with the breaking of contracts between Men and Gods, which caused a woman's perdition and blinded her to the point of engendering an artificial means to mate with her aberrant object of desire. From the excessive and indomitable character of this submission to sexual appetite, a repulsive monstrosity was created, which fed on the young sacrifices demanded of Athens. The Minotaur was trapped inside a labyrinth—a symbol of Minos's shame and despair—but which defied its algorithmic deciphering, and ensuing killing of the monster, which would allow a meaningless tribute to come to an end.

But the world has changed a great deal, and—although bullfighting has ritualized nature's submission to Man through the defeat of the bull—the Minotaur of today does not seem very intimidated. It is pulverized in infinite replicas of itself, having become a myriad of tiny machine-man Minotaur hybrids, scattered all over the planet, while we Humans now find ourselves inside the maze and walk along paths that lead nowhere except to self-sacrifice. That is, we are overseen by machines/instruments, but now capable of knowledge; hybrids, like the Minotaur, but no one hiding anymore, as shame is no longer necessary. And so, mesmerized by overconsumption, in the constant quest for immediate pleasures and the permanent seduction resulting from needs induced into us, we find ourselves slaves to a system. However, let us not delude ourselves, the algorithms of cognitive machines do not function on their own, but on the basis of powers granted them by groups of Humans who have so decided.

As for the present cognitive environment, the management paradigm has invaded all areas. We refer to the management of resources, expectations, and also emotions. We have become afraid of excess and risk and, as a result, we aspire to infallible health plans and live surrounded with ads from Insurance Companies. And so, semi-consciously, because we say the problem is left for the Future, we surrender to machines the control over our lives. In this setting, the algorithmic paradigm gains all its relevance, since a mechanical society, overwhelmed by the notion of optimization, may well be governed by machine agents.

Hence, the only way to untie the knots currently tying us is to devise a new paradigm. We do not know how it will happen, but to become more human will have to entail much less resource consumption, and much more knowledge and reflection.

From the economic viewpoint we shall have to talk much less about growth and competition and much more about empowerment and values associated with a qualitative realm. Cognitive machines may indeed help decisively in this process, but first we must decide on the path towards liberation.

If this is not so, then, in the present just as in the past, we will not need machines to further enslave us. Our human partners have proved more than necessary capable of doing so. Worse will be to give them additional means.

Tegmark (2017) characterises a set of possibilities that we should keep in mind for concluding reflection. Way before we can build a sustainable cognitive ecosystem without humans, we may risk succumbing to nuclear, environmental, or simply

uncontrollable viral catastrophes. We may also come to create a superintelligence—and the latter may decide to keep some humans as an attraction in a Zoo. However, it is not our position to contribute to these possible futures.

We must resist both an Orwellian world built with technology, and the alternative of preventively regressing to an anti-technological Amish-style outcome. Technology remains, in essence, instrumental. It is therefore up to humans—within the scope of an integrative and informed reflection—to answer the questions regarding what they wish to do with it. In the great moments of our History philosophical questioning has always gone hand in hand with scientific development. It was so with Aristotle, but also with Galileo, Newton, Darwin or Einstein. AI, in enabling routine, tedious and algorithmically complex tasks to be handed over to very competent cognitive machines, can free Humans for a more humane life. But this will only be feasible if we are capable of aptly foreseeing that humanized world, and if the powers that be are forced to join this novel conjecture, if and when civil society demands its realisation.

What's past is prologue, what to come
In yours and my discharge.
William Shakespeare, *The Tempest* (1610–1611)

References

Hegel, G. (1807). *Phenomenology of the spirit (Phänomenologie des Geistes)*. Leipsig: Verlag.
Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. London: Penguin Books Ltd.

Bibiliography

Abeler, J., Calaki, J., Kai, A., & Basek, C. (2010). The power of apology. *Economics Letters, 107*(2), 233–235.

Asimov, I. (1950). *I, Robot*. New York, NY: Doubleday.

Axelrod, R. (1984). *The evolution of cooperation*. New York, NY: Basic Books.

Charniak, E., & Goldman, R. (1993). A Bayesian model of plan recognition. *Artificial Intelligence, 64*(1), 53–79.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137–140.

Fischbacher, U., & Utikal, V. (2013). On the acceptance of apologies. *Games and Economic Behavior, 82,* 592–608.

Hamilton, W. D., & Axelrod, R. (1981). The evolution of cooperation. *Science, 211*(27), 1390–1396.

Han, T. A. (2013). Intention recognition, commitments and their roles in the evolution of cooperation: From artificial intelligence techniques to evolutionary game theory models. In *SAPERE series* (Vol. 9). Berlin: Springer.

Han, T. A., & Pereira, L. M. (2018). Evolutionary machine ethics. In O. Bendel (Ed.), *Handbuch maschinenethik*. Berlin: Springer.

Han, T. A., Pereira, L. M., & Santos, F. C. (2011). Intention recognition promotes the emergence of cooperation. *Adaptive Behavior, 19*(4), 264–279.

Han, T. A., Pereira, L. M., & Santos F. C. (2012a). Corpus-based intention recognition in cooperation dilemmas. *Artificial Life, 18*(4), 365–383.

Han, T. A., Pereira, L. M., & Santos F. C. (2012b). Intention recognition, commitment and the evolution of cooperation. In *2012 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1–8). IEEE.

Han, T. A., Pereira, L. M., & Santos F. C. (2012c). The emergence of commitments and cooperation. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1* (pp. 559–566). International Foundation for Autonomous Agents and Multiagent Systems.

Han, T. A., Pereira, L. M., & Lenaerts, T. (2015). Avoiding or restricting defectors in public goods games? *Journal of the Royal Society Interface, 12*(103), 20141203.

Han, T. A., Pereira, L. M., Santos, F. C., & Lenaerts, T. (2013). Good agreements make good friends. *Scientific Reports, 3,* 2695.

Han, T. A., Pereira, L. M., Santos F. C., & Lenaerts T. (2013). Why is it so hard to say sorry? evolution of apology with commitments in the iterated Prisoner's Dilemma. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (pp. 177–183). San Francisco, CA: AAAI Press.

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5

Han, T. A., Pereira, L. M., Santos F. C. & Lenaerts T. (2015a). Emergence of cooperation via intention recognition, commitment and apology–a research summary. *AI Communications, 28*(4), 709–715.

Han, T. A., Santos F. C., Lenaerts T., & Pereira, L. M. (2015b). Synergy between intention recognition and commitments in cooperation dilemmas. *Scientific Reports, 5*, 9312.

Han, T. A., & Lenaerts, T. (2016). A synergy of costly punishment and commitment in cooperation dilemmas. *Adaptive Behavior, 24*(4), 237–248.

Han, T. A., Saptawijaya, A., & Pereira L. M. (2012). Moral reasoning under uncertainty. In *International Conference on Logic for Programming Artificial Intelligence and Reasoning* (pp. 212–227). Berlin: Springer.

Hodges, A. (1983). *Alan Turing: The enigma*. London: Simon & Schuster.

Martinez-Vaquero, L. A., & Cuesta, J. A. (2013). Evolutionary stability and resistance to cheating in an indirect reciprocity model based on reputation. *Physical Review E, 87*(5), 052810.

Martinez-Vaquero, L. A., & Cuesta, J. A. (2014). Spreading of intolerance under economic stress: Results from a reputation-based model. *Physical Review E, 90*(2), 022805.

McCullough, M. (2008). *Beyond revenge: The evolution of the forgiveness instinct*. New York, NY: Wiley & Sons.

McCullough, M. E., Kurzban, R., & Tabak, B. (2010). Evolved mechanisms for revenge and forgiveness. In M. Mikulincer & P. R. Shaver (Eds.), *Understanding and reducing aggression, violence, and their consequence* (pp. 221–239). Herzliya, Israel: IDC Herzliya Press.

McCullough, M. E., Pedersen, E. J., Tabak, B., & Carter, E. (2014). Conciliatory gestures promote forgiveness and reduce anger in humans. *Proceedings of the National Academy of Sciences, 111*(30), 11211–11216.

McDermott, D. V. (2001). *Mind and mechanism*. Cambridge, MA: The MIT Press.

Ohtsubo, Y., & Esuka, W. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior, 30*(2), 114–123.

Okamoto, K., & Shuichi, M. (2000). The evolution of punishment and apology: An iterated prisoner's dilemma model. *Evolutionary Ecology, 14*(8), 703–720.

Pearl, J., & MacKensey, D. (2018). *The book of why: The new science of cause and effect*. New York, NY: Basic Books.

Pereira, L. M. (2017). Cibercultura, Simbiose e Sincretismo. In H. Pires, M. Curado, F. Ribeiro, & P. Andrade (Eds.), *Ciber-Cultura: Circum-navegações em Redes Transculturais de Conhecimento, Arquivos e Pensamento* (pp. 45–55). Braga: Edições Húmus.

Pereira, L. M. (2019). Should i kill or rather not? *AI & Society (Journal of Knowledge, Culture and Communication), 34*(4), 939–943.

Pereira, L. M., & Han, T. A. (2011). Intention recognition with evolution prospection and causal Bayes networks. In *Computational intelligence for engineering systems* (pp. 1–33). Berlin: Springer.

Pereira, L. M., & Saptawijaya, A. (2009). Modelling morality with prospective logic. *International Journal of Reasoning-based Intelligent Systems, 1*(3–4), 209–221.

Pereira, L. M., & Saptawijaya A. (2015a). Abduction and beyond in logic programming with application to morality. *IfColog Journal of Logics and their Applications, Special issue on Abduction, 3*(1), 37–71. In L. Magnani (Ed.). London: College Publications.

Pereira, L. M., & Saptawijaya A. (2015b). Bridging two realms of machine ethics. In J. White, R. Searl (Eds.), *Rethinking machine ethics in the age of ubiquitous technology*. Hershey, PA: IGI Global.

Pereira, L. M., & Saptawijaya A. (2017). Counterfactuals, logic programming and agent morality. In *Applications of formal philosophy: The road less travelled. Springer Logic, Argumentation & Reasoning series* (pp. 25–54). Berlin: Springer. ISBN: 978-3319585055.

Powers, S., Taylor, D., & Bryson, J. (2012). Punishment can promote defection in group-structured populations. *Journal of Theoretical Biology, 311,* 107–116.

Prinz, J. (2016). Emotions, morality, and identity. In *Morality and emotion* (pp. 13–34, 83–98). London: Routledge. ISBN: 978-1-138-12130-0.

Raihani, N., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution, 30*(2), 98–103.

Sadri, F. (2011). Logic-based approaches to intention recognition. In *Handbook of research on ambient intelligence and smart environments: Trends and perspectives* (pp. 346–375). Hershey, PA: IGI Global.

Saptawijaya, A., & Pereira L. M. (2015a). Logic programming applied to machine ethics. In *Portuguese Conference on Artificial Intelligence, LNCS* (Vol. 9273, pp. 414–422). Berlin: Springer.

Saptawijaya, A., & Pereira L. M. (2015b). The potential of logic programming as a computational tool to model morality. In R. Trappl (Ed.), *A construction manual for robots' ethical systems* (pp. 169–210). Berlin: Springer.

Saptawijaya, A., & Pereira, L. M. (2018). From logic programming to machine ethics. In O. Bendel (Ed.), *Handbuch Maschinenethik* (pp. 209–227). Berlin: Springer.

Schneider, F., & Weber, R. (2013). *Long-term commitment and cooperation*. Tech. Rep., Working Paper Series, University of Zurich, Department of Economics.

Smith, N. (2008). *I was wrong: The meanings of apologies*. Cambridge: Cambridge University Press.

Sterelny, K. (2012). *The evolved apprentice*. Cambridge, MA: The MIT press.

Tzeng, J.-Y. (2004). Toward a more civilized design: Studying the effects of computers that apologize. *International Journal of Human-Computer Studies, 61*(3), 319–345.

Utz, S., Matza, U., & Snijders, C. (2009). On-line reputation systems: The effects of feedback comments and reactions on building and rebuilding trust in on-line auctions. *International Journal of Electronic Commerce, 13*(3), 95–118.

Author Index

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5

Subject Index

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5