REVIEW ESSAY

Machine Ethics: From Machine Morals to the Machinery of Morality by Luis Moniz Pereira and António Barata Lopes (2020) xxi +164pp., £100 (hardback) Springer, Cham, Switzerland, ISBN13 9783030396299

Introduction

Readers of *Prometheus* with its focus on innovation enabling change – 'open innovation' – may be drawn to Luis Pereira's newest book with Antonio Lopes for a number of reasons. For example, someone interested in machine ethics, and a policymaker interested in the potential for evolutionary game theory applied to large-scale social coordination problems modelled in computer simulations over generational timescales, may both find the text rewarding yet come to it from different perspectives. The former may be most interested in Pereira's pioneering work in logic programming in the late 1970s and how this grounds his thinking about human morality now. The latter may be most interested in his more recent (also pioneering) work modelling social dynamics, including intention recognition, apology and guilt, and thereby demonstrating the positive effects that such capacities and practices have in the constitution of the resulting social system as a whole. For this reason alone (although the text leaves the reader with a rather stark dilemma and can be seen as a single argument for choosing one horn), composing a direct summary of such a central argument would seem to do a disservice to the scope of the book. As well, it would spoil the ending. So, instead, the present review essay begins with a strong focus on the context in which this book emerges as a product, as established by the invited prefaces and authors' introduction to the work. Then it pauses over some of the early chapters to relate some aspects of these to later developments in the text. Finally, this essay concludes by locating this new book (and us with it) in the context of the history of ideas that it surveys.

An appended review: context

Google scholar tells us that Luis Moniz Pereira has either written or contributed to publications garnering more than 8000 citations, including more than four per day, every day, since 2015. This newest entry is a book in the highly regarded SAPERE series of studies under the editorship of Lorenzo Magnani. Those familiar with Magnani's style may recognize a similar taste for erudition in the current work, a mode much less technical than the hard-core computer science on which Luis built his early career, and more conversational than his recent work on why agreement-accepting free-riders are a necessary evil in the evolution of cooperation (e.g., Martinez-Vaquero *et al.*, 2017). In this way, this short book with such broad discussion is more suited to philosophically oriented readers. And the subjects that it addresses – rule by algorithm, artificial emotions, so-called 'superintelligence' – suit this discursive style. These are popular subjects and should appeal to a general audience. Indeed, reaching this audience with lessons drawn from Pereira and colleagues' more technical studies seems to have been one purpose motivating the collaboration (as described on p.xv).

The book begins with a preface from Helena Barbas of the faculty of humanities and social sciences at NOVA in Lisbon. She places this work in the context of ongoing global resource wars, revolt and oppression, and points to the raw nerve piquing interest in machine ethics today. Given that morals are evolved customs, what are these to tell us to do when confronted with self-driving cars, for example? Where are we to find direction when inherited rules prove to be incomplete in the face of

rapid technological change, 'unable to supervise the new' (p.ix)? Questions such as these have stirred (recorded, Western) philosophy since Heraclitus. How the different chapters of the book funnel towards possible solutions to these age-old problems is a question addressed as this review closes.

Following this first preface is a second, written by scientist and author João Caraça, of the Gulbenkian Foundation in Lisbon, whose interest in social issues is most evident in for example a collaborative effort from 2018, which argues that fundamental changes in social institutions, corresponding customs and morals are necessary if political economies are ever to satisfy long-term human needs (cf. Jacobs *et al.*, 2018). By contributing this preface, Caraça further establishes the context within which the effort behind this book should be appreciated. He, for one, is not taken in by the hype of super-intelligent machines ruling the world, for good or evil. He reminds us that, behind every machine, there is a person or a group of people who built it for a purpose, to do a job. Rather than fear machines, it is the creator we should fear. From this concern, it is crucial that we deliberate together and openly about how we might proceed where evolved moral routines leave off. Interdisciplinary knowledge is required for this task and (in the estimation of Pereira and Lopes) is sadly lacking.

Caraça works from what should be an obvious fact, but one that is easily neglected. This is that the future of a society depends on its technological support structures and on the knowledge required to develop these in ways that support the highest aspirations of that society. At the same time, he derides the prevailing political economy for its inequality, wastefulness and exploitation of the natural systems on which we all depend. Most deeply, he is critical of the slavery that emerges in the separation of the person from her or his productive life, a process exacerbated by what his father, Bento Caraça, considered the 'automatism of man', now realized through human replacement by increasingly intelligent machines.

From this critical standpoint, Caraça places this book within the tumult of the contemporary world as does Barbas. Here, he locates the work in the middle of a revolutionary digitalization of social infrastructure which, through repetitive and daily interaction, supports the social organization that emerges through that continuous interaction, and that can be currently characterized by increasing injustice. Poignantly, however, he does not blame the technology. Drawing inspiration from his own father's pioneering work, he holds that we must look behind the machine, at the human beings responsible for the vision of society towards which such technologies are developed:

> 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. (quoting Bento Caraça from 1939, p.xiii)

This is also to put a fine point on a central theme around which the book itself revolves, and towards which it builds throughout.

Again resonating with Barbas, Caraça locates the reader in the midst of unprecedented change and feels called upon – as a civic duty – to develop an evolutionary overview of this process, to get a handle on essential dynamics and, thus empowered, to change the way that things turn out in the end. Here is the promise of Pereira and colleagues' research, to help provide such an overview so that society might extricate itself from the current situation. Society has accepted an historical and cultural evolution and corresponding practices relatively uncritically, an evolution largely directed by forces beyond human anticipation if not understanding. It is for this reason that Caraça suggests that study of Pereira and colleagues' work is a civic duty.

Next is the authors' preface explaining the purpose of the book, describing how it came to be in its current form, and setting out who has been responsible for what. The heart of the text has been drawn from manuscripts that Luis had been amassing, which were then revised in collaboration. And this process is evident in the way the text reads. Here, we find a strong voice speaking from a very recognizable position. Throughout the book, this position is developed in familiar ways.

> Evolutionary psychology . . . 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. (p.xvi)

This is a principle on which Pereira and colleagues' more technical work is ultimately based. Looking ahead in the text, the authors can be seen to draw this line from genetic evolution to founding Western mythology to discussion of moral life in the contemporary context (as established by Barbas and Caraça previously). Genes serve as vehicles for memes which programme persons with routines that serve the interests of the group: 'We are a discard package for both. . . . The educational system is just a meme production system, right inside our heads' (p.65). On this account, memes are 'cultural genes' (p.123), including inherited religious rules that represent (mal)adaptive strategies at the level of a group, and that are selected for their potential to enable coordination towards common goods that might otherwise have been inconceivable. Trouble arises, again, when they outlive their usefulness and render a society too ossified to adapt.

Trouble also arises when the process of meme replacement and revision in these individuals is somehow faulty. Later in this text, Pereira and Lopes confront the reader with the fact that erstwhile adaptive tendencies to synthesize and collaborate are being diluted by contemporary cyberculture, resulting in youth unable to integrate across disciplines and domains, unable to focus on solving complex problems, disinclined to collaborate and thus maladapted to the challenges facing civilization in this revolutionary era (chapter 15). Having lost our religions, we find ourselves with nothing to replace them.

Working against this trend, this text applies some of the successes of Pereira and colleagues' computational models in clarifying contemporary challenges so that we might face them head on. From the beginning, we read that intelligence – work requiring intelligent operations, including speculation – may be simulated in computers, thereby helping us to overcome biological limitations. Here, think about artificial intelligence (AI) as a sort of telescope, showing us what might happen if X or Y were the case. In this way, the foundational research supporting the arguments of this book help us both to understand how we got to where we are today, and also to predict in which sorts of situations a group may find itself if its members act according to certain rules reinforced by certain institutions. Social policy may be informed by this overview. With corresponding institutions so ordered, AI developed for such a purpose may afford a handle on the way the world turns out after all. In the end, however, the success of any such initiative depends on human beings and their capacities to make sense of things, to find such developments meaningful. This explains the authors' recurring emphasis on the interdisciplinary knowledge-base necessary to realize this potential, both now, conceptually, and through future developments, practically. Finally, the structure of the book is set out and this authors' preface ends with a rather extensive list for further reading, including links to PDFs in (almost) every case.

The body of the book consists of 20 short chapters. The next section briefly comments on some of these, pausing for discussion on notes taken during the initial read of the book. This review then concludes with brief discussion before leaving readers to discover the authors' final recommendations on their own.

An appended review: content

Though this work is grounded in decades of computer modelling, there is surprisingly little mention of these programs in this book. Rather, the work accepts the results of Pereira and colleagues' research, extensively reviewed in the preface, and suggests how this work may inform our understanding of contemporary and anticipated social problems as well as help us formulate possible solutions.

The first chapter is the most important. It argues that problems facing humanity today are of two types. One concerns what type of society we are to realize through our concerted technological development. The other is how we may understand human morality well enough to engineer

moral machines. We are confronted by machines that liberate us from effort. However, to respect the value of the human beings who found purpose in corresponding ways of life, 'a new social contract is indispensable', a contract that re-establishes what is expected of people given the robot revolution currently under way. Human beings are constituents of social systems, and live and act as integral members of society on which they depend and to which they contribute. Without a new social contract recognizing this dynamic, the authors forecast an emerging neo-feudalism, with one caste controlling the means of production and another alienated from the determination of production and yet dependent on the eventual form of this system of automation. What is left is an image of social support structures without a people to support so much as a set of programs to keep them running. Already 'The vast majority do not live but fulfill pre-established algorithms' (p.50). And, with this dystopic view in mind, the urgency with which the issues of this book must be met becomes comes clear.

In short, the authors' research focuses on understanding what promotes moral cooperation in populations of logic-programmed computer agents so that similar dynamics in human populations can be understood. We may consider that a computer program 'is a set of strategies defined by rules' that tells a given agent what to do in a given situation, just as religious rules may tell a follower what to do. A program may be populated with different agents representing distinct strategies, themselves represented differently in the lines of code that tell them what to do in given situations. Agents can also learn from each other, through social learning, which 'consists of any given player imitating the strategy of another, whose results indicate that they have been more successful' (p.5).

'There is no fixed, frozen morality' (p.10). 'All life is an evolutionary stage, where replication, reproduction, and genetic recombination have been testing solutions for an increasingly improved cognition and action' (p.16). Improved cognition provides the potential for further adaptation, communication and the mixture of moral practices throughout populations wherein as individuals follow each other, innovate in the face of novel situations, or free ride, in order to produce more offspring in the sense of representing winning strategies as they appear more frequently in the next generation. In their more technical work, and as introduced in the authors' preface, this is all cashed out in terms of evolutionary game theory (EGT), 'which consists of seeing how, in a given game with well-defined rules, a population evolves through social learning'. The question for Pereira and colleagues, then, becomes: 'Once certain rules are defined, how does the social game evolve?' (p.5).

This general approach is developed in various ways throughout the book as the authors meet challenges arising in different contexts. For instance, skipping ahead to chapter 17, 'Employing AI for better understanding our morals', the authors review research on intention recognition, and introduce a principle resulting from this research balancing costs of prefiguring and enforcing cooperative arrangements for mutual benefit while minimizing free riders:

> 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. (p.127)

With such an example, the potential for such fundamental research in computational modelling to help policymakers understand how to serve public interests during periods of rapid social change should be clear.

One important question introduced early on in the text concerns the roles of autonomy and free will in the evolution of morality, and how the contributions of individual expressions of freedom to an eventual social organization may be evaluated. For any given agent within a population to be considered moral, it must have options from which to select. Morals themselves form as strategies are adopted within a population in response to contextual and informational change, when some new or different way of doing things results in a better situation overall. It is necessary that an agent deliberate over possible strategies and their combinations, with options to act

in one way or another in order to be free to exercise autonomy towards some self-determined optimal end. These ends are treated as hypotheses which an agent selects to test through action. This is basically how Pereira and colleagues' programs work. Moral agency depends on counterfactual reasoning exercised in the deliberation over possible ends, and in the selection of the most desirable, given their consequences and side effects. Social agents are able to leverage this ability in the consideration of the possibilities available to other agents, and to surmise their likely intentions in order to coordinate action. They are also able to communicate this reasoning in a similar form, explaining why one course of action is preferable to another. As it turns out, groups as wholes do better with a certain admixture of strategies, with some constituents more gregarious than others, for example. Too many following overly selfish or overly optimistic strategies? Sub-optimal situations result.

Pereira and Lopes do not buy the hype surrounding the notion of super intelligent killer robots. Contemporary AIs remain relatively simple programs. But, because even these relatively simple programs can replace human beings in performing certain tasks, they are oversold as panaceas for social problems while proponents neglect anticipated fallout (e.g., worker displacement, loss of productive roles in society, diminished tax revenues). We are in urgent need of a new social contract with the full impact of such automation made clear. Finally, we need to understand our own human morality better, at least in part because morality is concerned with how to do the right things (e.g., produce the greatest good for the greatest number). The authors emphasize that this study should take place in universities as places in which reasoned discourse can drive inquiry into sensitive areas. Universities must respond with urgency. To do less – given the relationship between adaptation to such radical change and morality that grounds this text – would be nothing less than immoral. Indeed, to further this study and the solutions that may come from it is a civic duty, full stop.

The second chapter reveals more about the authors' view of morality. Here, they explain that moral methodology is essentially top-down, and that moral machines must be able to explicitly account for their behaviours (i.e., give and respond to reasons). The substantial third chapter builds from this thesis, offering a sectioned account of AI and emphasizing that the capacity for computer hardware to run any given software is responsible for progress in AI research:

> 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. (p.31)

This chapter extends the thesis that morality requires, and is ultimately realized through, symbolic reasoning. Human symbolization represents evolution at work, culminating in statements of human morality including universal moral rules. An example might be the externalization of these symbols into artificial (moral) intelligence and the progress towards an 'engineered platform for cognition [which] might be interpreted as "just" another evolutionary leap' (p.25). In other words, one need not be surprised by moral machines, and should see them as a next step in a natural course of human development. The history of AI briefly articulated in this context is interesting. The authors note that its progress has steadily brought computational intelligence closer to human-like intelligence, evident in the development of intuitive graphical user interfaces on the one hand, and advances in human robot interaction and social robotics on the other. The authors also emphasize that the goal of AI research in its purest form is to understand intelligence in a general way, such that intelligent artifacts, including autonomous machines, may be built by engineers, just as musical instruments are created by artisans and compositions by music composers.

Chapter 4 begins by recognizing the difficulties in designing autonomous machines. Noting that there is nothing in principle preventing autonomous machines, the chapter concludes that they are possible. How might the pinnacle of evolution – evolved human morality – be captured in a computer? Borrowing from Daniel Dennett, the authors argue that, though the world is

more complex than any explanatory model conceived to account for it, all this complexity may arise from simple processes. Thus, though complex in appearance, 'our future is closed, we just don't know how'.

> This thesis is of crucial importance for understanding the work that 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.[1]

To this, one might object: If everything is determined, where is the room for the freedom required for moral agency? The authors answer: 'In this scenario, freewill probably emerges from the interaction between the various items that constitute a context' and that, through such interaction, 'the entire evolutionary process can be traced as a selection of well-adapted algorithms' (see p.35). Finally, given that 'what matters is the agent's ability to represent itself in action, and to generate and analyze *possible futures* by virtue of their internal models of reality', the authors answer 'Yes: we can build autonomous machines' (p.37).

Discussion and conclusion

The rest of the text becomes increasingly critical and indeed controversial in its assessment of contemporary problems and their origins (for instance, in discussion of the Minotaur in chapter 20). At every turn, the authors emphasize the potential for fundamental research (in AI, and also social psychology, philosophy, evolutionary biology and other fields) to help resolve these problems. With every choice, we must ask ourselves what is really important. To come to an answer, 'it is becoming increasingly urgent to have critically informed citizens who are not anesthetized with football and soap operas'. Instead, the authors recommend directing public attention to the paths forward illuminated by new technologies and innovative scientific research, such as that discussed in this book.

> 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 [to ignore]. (p.66)

At the same time, the authors recognize that, given contemporary social pressures affecting self-development in so many counter-productive ways, the requisite degree of critical information may be increasingly difficult for us, individually and collectively, to realize (see, again, chapter 15). This is to say that the current maladaptive state of Western culture seems not to be preparing humanity for a successful transition into anything other than dystopia, though the authors purposefully set this likelihood aside in order to focus on the positive potential of AI and related technologies to pave our way in the ongoing odyssey of human evolution (cf. p.141). It is on this adventurous note that the text leads the reader to its conclusion, at the window frame of the future and with a telescope in hand (or at least the sketch of such a device and what it can reveal) in hand to show the way.

Ultimately, it is this ability to present possible futures in a clear and accessible form that is of lasting personal interest in Pereira's research for the present reviewer. Can computational models – psychologically realistic computational models – help us to see our way through necessary social transitions in the self-directed, open and cooperative movement from here, now, to a collectively better future (cf. White, 2020)? These transitions may have to take place over the course of many generations. Can these and similar technologies help us to understand how these transitions may be effected (cf. White, 2016)? In the past, such intergenerational guiding frameworks were religious.

---

[1] Those interested in the kernels of these ideas in the context of Pereira's foundational work in logic programming should see Warren *et al.*, 1977, p.113).

People were born into ongoing religious narratives, oriented to good and bad, with happy and unhappy endings to life stories. These learned values have been reinforced with native mechanisms experienced as guilt, or shame, trained and enacted through apology or revenge. In this book, for example, the authors often discuss the role of guilt in the Catholic religious tradition, representing the metaphysical space of value that characterized daily life in late 1970s Portugal. The point here is that these constructs, these grand religious cosmologies, held and still hold people together. Many may have outlasted their usefulness. Memes may fail to be adaptive. At the same time, there has been, it is fair to say, nothing short of a war on religion fuelled by technological developments. Consider the impact of applications such as magnetic resonance imaging in neurological contexts, and cognitive neurorobots demonstrating musical improvisation, on notions that consciousness is a divine light and that intelligence is unique to human beings in all of Creation. With the former we may correlate subjective phenomena with mechanical transformations, and in the latter we confirm that artifacts can act as if they are living even though we know that they are not alive. Against such a backdrop, what is the role of Catholic guilt? If guilt plays a necessary role, do we need a God to assign it?

Recalling the prefaces to this work and the contemporary context they establish, the question that we are presently and collectively facing is what to do now that we can recognize so clearly that we have to change direction. If guilt is useful, but inherited institutions no longer represent this usefulness, how are we to shape new ones? One point seems sorely missed in all of this discussion. Religions themselves are technologies. Kant is clear on this. Religion – at root a relationship with God – is a device invented for the furtherance of morality. Religion is innovative. We make it to do a job. Religion does the job of tacitly directing members of a population toward a commonly recognized good, keeping them going in the same general direction from birth to death. And from this understanding, we may ask why our religions don't work to improve adaptability to change, including the rapid change that we are witnessing today. Have we been using them incorrectly?

Finally, for all of the authors' discussion of different myths and their reinterpretation in the light of contemporary AI and related technologies, upon reflection there is one reference worth adding. This current era, as recognized in the concerns motivating Pereira and Lopes to craft this text, returns us to the third book of Plato's *Republic*. Here, we meet with discussion about which stories to recount and which songs to sing, which virtues to extol and which ways of life to champion, should we aspire to anything like an ideal society.[2] Pereira and colleagues' fundamental research can help us resolve such a complicated problem. This book challenges us to remake the myths into which we have all been born, in terms of which we have all been educated and currently live – if only in the mode of rejection, struggling to free ourselves. The great promise of this work is that it can help us refashion our city without the extermination of older generations in order to remove resistance to change.

The authors have presented their text as a sort of discursive alternation between specific insights and their potential to inform future group-level change. In my opinion, this is the great benefit of this and other applications of AI. However, without a plan, left for instance to faith in the markets, in the state, or in gods instead, history teaches us that we are doomed. Indeed, some religious texts (from Hinduism's fourth turning, to Judaeo-Christianity's end times) seem to forecast exactly that – and teach us to expect it. How are we to undo such an education without the tools to replace these dead-end memes and their old religious vehicles? How might we replace these pre-structured moral traditions that no longer fit our times with something of our own composition with the aid of advancing technologies? This is the context of the present text, and it is the problem that the text resolves. The problem, though immensely complex, may be accessible to relatively simple solutions, after all.

---

[2] See Plato, 1997, *Laws*, books I, II (for instance 659d–660e) and VII (for instance 796e3–800b1).

References

Jacobs, G., Caraça, J., Fiorini, R., Hoedl, E., Nagan, W., Reuter, T. and Zucconi, A. (2018) 'The future of democracy: challenges and prospects', *Cadmus*, 3, 4, pp.7–31.

Martinez-Vaquero, L., Han, T., Pereira, L. and Lenaerts, T. (2017) 'When agreement-accepting free-riders are a necessary evil for the evolution of cooperation', *Nature Scientific Reports*, 7, 2478, pp.1–9.

Plato (1997) *Complete Works*, Hackett Publishing, Indianapolis IN.

Warren, D., Pereira, L. and Pereira, F. (1977) 'PROLOG – the language and its implementation compared with LISP', *ACM SIGART Bulletin*, 64, pp.109–15.

White, J. (2016) 'Simulation, self-extinction, and philosophy in the service of human civilization', *AI & Society*, 31, 2, pp.171–90.

White, J. (2020) 'The role of robotics and AI in technologically mediated human evolution: a constructive proposal', *AI & Society*, 35, 1, pp.177–85.

Jeffrey White
Cognitive neurorobotics - Tani unit
Okinawa Institute of Science and Technology, Onna-son
Japan
jeffreywhitephd@gmail.com