The Carousel of Ethical Machinery [1]

Luís Moniz Pereira

NOVA Laboratory for Computer Science and Informatics (NOVA-LINCS),

Departamento de Informática, Faculdade de Ciências e Tecnologia,

Universidade Nova de Lisboa, Portugal

Email: lmp@fct.unl.pt, ORCID: 0000-0001-7880-4322

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.

Keywords: Machine Ethics, Ethical AI, Evolutionary Morality, Social Impacts of AI.

[1] This article consists in a topical introduction, a short synopsis, and an attention calling to our recent new book, wherein the topics broached here are awarded further detail in its 20 chapters: L. M. Pereira and A. Lopes (2020). It also picks up important issues from our previous book L. M. Pereira and A. Saptawijaya (2016). Hence his article is not an academic survey, the topics of machine ethics having become nowadays so wide and widespread. It defers to the more academic style, cum extensive references, to the above sources as well as to the plurality and variety of the more technical works of ours therein mentioned.

Early Mythology Antecedents

In the Hellenistic period (c.10 – c. 70 AD), Hero of Alexandria and other brilliant Greek engineers produced a variety of machines, driven either hydraulically or pneumatically. 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. Their myths about the impacts of these technologies on their lives and livelihoods may tell us today something about the state of our own future world[2].

Hephaestus, god of technological invention, foretelling already our future, built a fleet of tripod stools "without driver" that responded to commands to deliver food and wine to the Gods, who were not to be encumbered with such trivia. More remarkable still was the bundle of life-sized 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"[3]. Over twenty-five hundred years later, developers of Artificial Intelligence (AI) 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, comprehend language and talk, analyse massive clouds of data, and infer the desires of humans. We can imagine then, as a persistent follow up on these myths, that in future robots will come to deliver to our desires on command, if not in anticipation thereof. Here big issues arise, as old as the myths themselves: Whose desires will the AI robots mirror? With whom will they learn obeyance?

In Greek mythology, Hephaestus's divine laboratory included an android commissioned by Zeus. Her mission was to punish the humans for having accepted the technology of fire, stolen from the gods for them by the Titan Prometheus, who had created them from clay in the first place. Zeus ordered Hephaestus to make a female from earth, which each God then endowed with a human trait: beauty, curiosity, charm, knowledge of the arts, language and deception, who was given the name Pandora[4], meaning "all gifted".

[2] We have mainly consulted, in this regard, Stephen Cave et al. (2020, Stephen P. Kershaw (2007) and Adrienne Mayor (2018).

[3] 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 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".

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

Zeus sends her to be wife to Epimetheus, an individual known for his compulsive optimism. Prometheus, his brother, had warned Epimetheus not to accept gifts from Zeus, but Epimetheus accepts the gift, fearing that Zeus will threaten humanity if he refuses. There is a wedding, all of the gods are present, and as wedding gift Zeus gives Pandora a "pot" – latter called "box" - filled with everything terrible for humanity. Zeus does not tell her what is inside the pot but does warn her that she must never open it herself. An intelligent artefact, agent of vengeance sent by Zeus the supreme God, Pandora is overcome by her curiosity, opens the box, and since then humanity has been afflicted by vices such as greed, illness and death. Without malice, she thus carried out Zeus' intentions to punish Prometheus and humanity by introducing catastrophes to torment humankind forever.

In the Greek canon, Epimetheus is the Titan god of retrospection and afterthought, post diction and excuses. He is the god who learns lessons *a posteriori*, acting without forethought, and became unguardedly smitten with Pandora and her curiosity with calamitous results, as one may expect. Prometheus on the other hand was the Titan god of foresight. He thought before he acted. Epimetheus often reacted before thinking, suffering all the inherent consequences of impulsiveness[5] of which his love for Pandora was just one example. Today these movements in thought are constitutive of AI, e.g. in the role of postdiction in predictive coding, error minimization through backpropagation. Importantly, these two movements of thinking, prospective and retrospective, are especially applicable in particular to morality, as we shall see in the next section.[6]

Present Day Mythology

Prometheus-type thinkers of our contemporary era have been warning scientists to stop, or at least to diminish, the reckless drive to employ increasingly autonomous AI, because they predict that, once set in motion, humans will not be able to control it. Concerns are not only about AI with dangerous or deceitful intentions, but also for applications that for instance lower the thresholds for armed aggression and war. These are serious concerns. Contemporary deep learning AI algorithms enable computers to extract patterns in vast data, extrapolate from them to compose novel future situations, and then to make decisions and act towards the realization of these autonomously determined situations, without human guidance. Inevitably, the entities with full AI will develop the ability to interrogate themselves and will respond to questions they may discover. Should progress towards human-like AI continue in this way, future artificial entities will develop the ability to

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

[6] Whenever the distinction is not important, we shall use "ethics" and "morality" interchangeably, as is common usage.

interrogate themselves and look for answers to questions that they may discover. They can be curious and, regardless of good intentions, such curiosity may deliver tragic results.

Today’s computers have already been shown to be capable of developing both altruism and deception on their own[7]. Future agents with AI – possibly freed from their creators - may like Pandora seek knowledge that is hidden from them. In the path of their desideratum, will 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 Pandora's box – like computer viruses crafted by some sinister hacker who seeks to make the world more chaotic - misery and evil have flown to pester humans since the existence of the world. In simplistic fairy tale versions of the myth, the last thing in Pandora's box was hope. Zeus had programmed Pandora to close the lid, holding prescience there. Deprived of the ability to anticipate the future, with Prometheus bound to a rock and his liver eaten everyday anew, humanity is only left with hope. In less optimistic versions, the last thing in the jar was the "anticipation of misfortune." On its face, this sounds worse. But looking more closely, it is the echo of Prometheus’ original gift to humanity – foresight and the arts - when confronted with an empty jar. With hope, and the anticipation of possible misfortune, there is planning. There is foresight, and the promise of prediction and innovation to direct current action to ideal ends.

As with 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 big utopian futuristic dreams is timeless. In other myths, such as about the hubris of Icarus, the ancient Greeks illustrated the quintessential attribute of humanity, to apply its crafts to reach the "beyond human" without the foresight to predict consequences.

Someday, perhaps AI entities will be able to pattern the deepest desires and terrors of mortals, expressed in our mythical reflections on artificial life. Will this AI somehow understand the expectations and fears we have today about the creations of human-like AI? In realizing that human beings foresaw their existence and contemplated the dilemmas that such machines would encounter, perhaps such AI-endowed entities will be better able to understand – and even empathise – with those impasses that they represent for us now. Will they look forward with hope to plan for a better future, avoiding misfortune for humanity, or will they be trapped in

[7] Cf. https://spectrum.ieee.org/automaton/artificial-intelligence/embedded-ai/ai-deception-when-your-ai-learns-to-lie) or https://link.springer.com/article/10.1007/s10462-008-9080-7

the regret of an incarnate mistake? Will they supplant their creators in dealing with the latter's original dilemmas, albeit correcting their mistakes?

The emergence of such a form of "culture" among robots does not seem too exaggerated. 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 inculcate machines with some of their human-ness[8] it makes sense to ask: Wherefore an ethics for machines?[9]

- 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.

With the current emergence of *deep learning* tools that allow 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 may no longer be the only intelligent agents, capable of autonomously deciding on aspects that directly impact our lives. This situation demands moral rules and principles applicable to the relationship between machines, the relationship between machines and human beings, and meticulous deliberation over the consequences 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 computationally elucidate the emerging of cognitive processes in our species' evolution, and in its technological aptitude for the design and production of intelligent software and artefacts, constitutes an intellectual challenge on a mythic scale.

The Carousel of Ethical Machinery

[8] Cf. James H. Moor (2006).

[9] For the moment, we shall use "ethics" and "morals" interchangeably, as is common usage. Though, properly speaking, "ethics" refers to some collection metaphysical principles, and "morals" refers to concrete cultural norms following from such principles.

The complexity of the issues raised is synthetically illustrated in the scheme further down, dubbed "The Carousel of Ethical Machinery," the article's title. It presents interconnected problems that are factors for decisions concerning the constitution of ethical machines. By now, it is well granted that the subject of computational morality is of special interest not only to companies and public institutions, but also to those who want to exercise a conscious and critical citizenship. The title of this article tells almost everything, i.e., classes of problems we must critically address and consciously articulate when creating machine morality, given the need for these to have morals. In short, machines are becoming increasingly sophisticated and autonomous, they will have to co-exist and socialize with us, and therefore must align their values with our own, for these are what binds us together as the gregarious society that we must continue to be even after they are introduced.

The moral machinery, implicit in the title's carousel mechanics, aims to make explicit that, deep down, morality is constituted by a series of rules that are mechanisms which are devices in the sense whereby they constitute those moral instruments that societies adopt in order to flourish over generations. Because they are mechanisms - if we understand them well - they may also be exported to machines. So, the problem becomes how to better understand our own morals.

The topic of morality has two major dimensions. The first is called "cognitive," that is around the need to clarify how we think in moral terms. For one thing, 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 various scenarios, and various hypotheses that motivate their relative pursuit. These scenarios must be compared to see which are most desirable, what are their consequences, what eventualities they open up that might otherwise be inaccessible, and what are any untoward side effects. All of this has as prerequisites certain cognitive capacities, such as prospecting possible futures and being able to select from among them, or to creatively compose something more ideal for all involved. These said cognitive capacities have to be useful for our collective existence.

And our brain has these abilities, which implies a lesson from evolutionary psychology: morality is not an individual thing. Morality is necessary within a population so that its constituents may cooperate, be gregarious and behave in a way that benefits every deserving one. The group can operate better if the constituents cooperate. So, the problem of morality is to ensure a common advantage, rather than each one doing only what they wish for themselves. This is its second "populational" dimension. We assume therefore that the existence of moral behaviour in a population requires certain cognitive capacities that determine our potentialities for coexistence. Hence, we should recognise that these cognitive capacities evolve, and that their development in our species has been beneficial, i.e. morality has contributed to the flourishing of the human race over evolutionary time. Moreover, we expect that certain moral competencies determine relative potentialities for coexistence, given different societal and resource dynamics. One such cognitive competence, besides looking into the future, is that of looking into the past; that is to be able to think "Knowing what I know today, what I would have done otherwise at some point in the past?" And I can use the results of such considerations to give recommendations to people who are in the situation I was in before, or to improve my future performance. This allows a form of social learning that requires this specific cognitive ability to imagine how the past might have been different, possibly taking into account information only obtained in the future. And, it is an example of a moral cognitive ability that is reinforced with its exercise.

"Retrospection," as Peirce (Herman Parret 1993, chapter 3, pp. 74) had at one time called it, is just one example of moral cognitive aptitude. The ability to relate with others through empathy in order to construct salient possible futures is another. The present article highlights published research construed by the author and co-authors[10,11,12] in terms of different cognitive abilities. In this growing body of research, certain cognitive abilities were studied to see whether, singly or in combination, they were promoters of moral cooperation in populations of agents, specifically of computer programmed agents coexisting with one another in different ways under different conditions. These agents are in effect sets of strategies defined by rules. That is, in a given situation, an agent's program pursues a certain action dictated by its existing strategy. Other agent programs also have actions dictated by their respective strategic rules. It is as if they were agents living together, each with possibly different goal options. Then, the question becomes if and how such a population might be able to evolve towards a stable and sustainable condition as different strategies are maintained over various spans of time, and social learning is taking place.

These studies are grounded in Evolutionary Game Theory [EGT], which consists in examining how and under what circumstances, for a given game with well-defined rules, a population evolves through social learning. On this account, 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, with some probability, the strategy of another one whose results indicate that they have been more successful. Rules of

[10] L. M. Pereira and A. Saptawijaya (2016).

[11] A. Saptawijaya and L. M. Pereira (2018).

[12] T. A. Han and L. M. Pereira (2018).

such a game can be defined for how the social game evolves. Using such models, our research has been able to show how morality understood as basic cognitive mechanisms may have developed in our species. Additionally, what lessons can we take from these studies concerning our own hopes for a world rich with future AI?

For our species to arrive where it has arrived today, evolutionary processes generation after generation selected us in terms of a morality suiting gregariously beneficial coexistence. This is a relatively non-controversial thesis and grounds the research that is our focus here. 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. These rules, their transmission and the process of their refinement are not always convenient, 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 advantages without paying costs, and this makes satisfactory resolutions more difficult if not impossible to achieve. This is the essential problem of cooperation: how can it be possible to both have everyone benefit from mutual cooperation and, at the same time, to deny opportunity those who want to abuse it?

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[13].

Social cognition requires what is often called "theory of mind". Evidence abounds (Hammerstein and Stevens 2012) showing that humans (and many other species) are capable of complex cognitive abilities including 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 these cognitive systems that seem to explain human behaviour are made compatible with

[13] Readers wishing to explore further can visit the author's publications page at https://userweb.fct.unl.pt//~lmp/publications/Biblio.html There they may consult dozens of research articles – both of a philosophical or a technical nature – whether in cognitive and population domains. The ongoing research is based on theory, programming, experimentation, and verification of interdisciplinary consonance with what is known of reality, evolutionary and present.

Darwin's evolutionary theory, and thus to perceive and justify their appearance in terms of a dynamics of cooperation, and also cooperation's absence.

In a sense, we may see the evolution of humanity as the embodied and ongoing solution of an intergenerational coordination problem as evidenced in cognitive capacities that are embodied evidence of this solution, what we call morality. For researchers, the study of cooperation and the emergence of collective behaviour crosses disciplines as diverse as Economics, Physics, Biology, Psychology, Political Sciences, Cognitive Sciences and Computing. And this is the greatest interdisciplinary challenge that science faces today. It is so vast that, for now, the effort in this article 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 of AI, ethics of machines, and their social impact. It is a new situation because, for the first time, we will have beings who are not us, but who will act among us and determine with us, if not for us, how and perhaps for what we will live. They will have a significant impact on our lives over the next dozen years - not to mention today - as well as on society as a whole over intergenerational timespans. Mathematical and simulation techniques employing EGT have proven useful in approaching this problem, but more work must be done.

Around the Carousel

It should be emphasized here that we are facing a *Terra Incognita*. But, differently from Epimetheus, we are doing so with the benefit of foresight. 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 the academic context, study often begins with abstract ethical theories. In fact, there are several ethical theories, antagonistic to each other, alongside research ongoing in different fields that also seem to complement each other. For instance, philosophy and jurisprudence both study ethics as the problem of defining a value system articulated in terms of principles that guide moral action. Though the formulations of these principles differ, each typically begins from abstract ethical principles to arrive at concrete moral rules. In practice however, a set of moral rules results from a historical, contextual, and philosophical combination of ethical theories that had themselves evolved over time as expressions of stable patterns of behaviour in light of successes and failures relative varying environmental circumstances. The fact that human beings evolved in terms of a common natural environment, with common natural laws, with common biological requirements and more or less common challenges to meeting these requirements, is evidence enough that there is something universal to every ethical construct, however seemingly antagonistic at first glance.

The carousel below (Figure 1) sums up in a way the complexity of the problematics involved in delineating moral machinery. Its core is to identify those factors that relate to what to do, or how to act. At the centre, "What to do" is surrounded by as many intervening factor carousels as we might want.

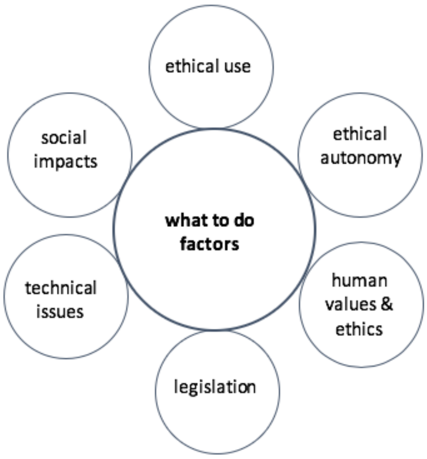

Figure 1: The machine ethics carousel

Each circle has to do with the ethical use of machines. For example, 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 another example, it is a negative, immoral practice for a program to pretend to be a human being. In January of 2019, the California legislature passed a law that forbids a computer or a program to simulate a human being without explicit acknowledgment that it is doing so. Of course, there are many other examples of immoral uses of machines. Among them, the most sinister promises to be drones with the ability to more or less autonomously target individuals for assassination, for example. In this context, it must be remembered that concern about the propriety of an action is increasingly subordinated to the machine itself, precisely because of increasing machine autonomy, which consequently amplifies the questions of their use. Perhaps it will be up to machines to protect us from their unethical use by other humans. 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[14].

Currently we have programs that control airplanes, boats, high-speed trains, for which computer technicians can prove that a given design is correct. For example, a successful boat piloting program will never try to destroy a bridge, or that of the train will not take a turn too fast for fun or from negligence. Similar problems can

[14] 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.”

be posed in relation to the proof of correction for increasingly autonomous machines, and for those which, not being autonomous, are controlled by individuals. Should an individual intend to do something immoral with an adequately moral machine, the goal of a machine ethics is to give the machine the ability to say "No; I will not do that." To achieve this, programmers have to be able to prove, with computer techniques, that a given program will never (intentionally) harm a human being.

This is a crucial reason why we need to introduce morals into machines, not so that they will reliably do everything that they are programmed to do, but also so that they will *not* do everything they are programmed (by fallible, flawed human beings) to do. We do not want a machine to be in a situation of simply stating "I did it because they told me to" in allusion to the position of Nazi war criminals in Nuremberg, saying "I just followed orders, I did what they told me to do" (Nuremberg 1945-46) 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, with the courage to exercise this capacity when necessary where human beings through history have proven less reliable (cf. White, 2010, 2012).

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. And in yet another platform of the carousel, Legislation is highlighted because at the end of the day, everything will have to be translated into laws, norms and standards communicating to other human beings exactly what is allowed or forbidden.

Cars have environmental impacts and are given pollution regulations. As engines have to meet certain criteria, human beings have regulations that specify their optimal operation given their impact on natural and social environments, as well. 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 or the manufacturer? One obvious response is to treat the case in the same way as that of a human being operating the vehicle with poorly formed intentions, i.e. as a case of negligence. Such an approach implicates the operator, and the engineer, but consideration of such cases seems not often to implicate 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 specify the tests that a car without driver should be evaluated. If it is found that such tests have not been sufficiently thorough, the entity that approved these vehicles will also be responsible. In so far as these are not deemed sufficient, regulations must change.

Another platform on the carousel is that of Technical Issues. All of this always involves the part of actually building moral machines, regardless of purpose. And, this is no small part because not everything conceivable is technically possible. For example, we still do not know the terms of the proof that a machine is not going to do ethically incorrect things. This does not take into cases where any hacker can enter the system and force a machine to do wrong things. This is a security problem that has to be solved in a technical way, and it is not clear how to do so.

Last but not least, there are the social impacts of machines. When we refer to machines, we are talking about either robots or software. The latter is much more dangerous, as it can more easily spread and reproduce anywhere in the digital world. As for a robot, it will be much more difficult to reproduce, for it implies a much greater cost. It also brings out the material limitations inherent to the possession of a volumetric body, and its code is specific to its embodied hardware. As far as the 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 engineering requisite fine eye-brain-hand coordination are daunting. These are capabilities on which an embodied intellect depends, and that that machines still do not have as much as humans, though on this front robots are moving quite fast.

As far as software is concerned, the issue is more worrying because programs are evermore reaching cognitive levels that have been a human monopoly until now. Machines now play chess and go, make medical diagnoses, perform autonomous driving, warehouse management, personal delivery, etc., and increasingly they will perform higher sophisticated mental activities better than we might. This opening creates for the first time at that level a competition with humans 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 enrich. At present, the rich are getting evermore richer and the poor getting poorer and poorer. AI is already contributing to that inequality and will expand it even further. Understandably, people worry.

At some point, a new social contract will be demanded. The alternative is social cataclysm. The way we function in terms of capital and labour, the way these two are equated and balanced, will have to be drasticaly reformulated. If this is not done, there is the risk that, sooner or later, growing asymmetries of wealth will cause a great revolt, insurrection, and social collapse. It will not be like the recent resentment of the yellow vests (*gilets jaunes)* in France, but much wider and deeper than that. It will occur when a caste system made possible by AI's very advances – to be described later – will come to provoke its own implosion. To avoid a collapse of this nature, it is imperative to begin to sketch a new social contract as AIs are introduced, now.

Moral Simulations

In one game that we devised, a robot attempts to save a princess and, in the effort, combines several ethical approaches[15] demonstrating that, in practice, morality not simple. We ourselves do not singly follow the morals of the errant knight, or utilitarian morality, or Kantian morality, or Gandhi's. Our ethics is an admixture of them and keeps evolving. Our research with this game illustrates that there is no fixed, frozen morality; morals evolve[16]. We must assume that the programming of morality in machines should similarly allow for its own evolution[17].

Morality is an evolutionary thing. It has been developing throughout the history of the species, in the service of the species. In this context, it is interesting to note that drone speedboats coordinate in swarms to attack enemy ships. This is an ethics in which machines act together. For example, we can imagine a swarm of drones on a country border, controlling the movements of swarms of immigrants, attempting to drive them away from water wells and good roads, frightening them in some way. Drones acting in such platoons make it much more difficult to control or predict their behaviour. Their movements are no longer predictable at the level of the individual drone but are the emergent result of a population of drones. This dynamic underscores the importance in studying morality in terms of populations and their configuration parameters when considering the actions of machines.

It is clear that machines are becoming more and more autonomous, and we need to ensure they can act with us on our terms and with our rules. Here we may speak of 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 in machines, we better understand our own morality. Consider the example from our scientific work on

[16] 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 references therein): https://userweb.fct.unl.pt//~lmp/publications/online-papers/lp_app_mach_ethics.pdf.

[17] 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.

guilt. When guilt 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.

Successful simulations resolved in experiments by Pereira and colleagues have involved the introduction of a modicum of guilt, neither too much nor too little in just a few agents into a population interacting within a computer, playing different strategies evolutionary games[18]. Without the existence of this component of guilt, most agents will tend to play selfishly, each wanting to win more than the others, with the population failing thus to reach a level where everyone can win even more. However, this desirable result is already possible with a dose of initial guilt, which modifies behaviours and with their success guilt spreads as a good strategy to the entire population. We can thus show mathematically that a certain amount of guilt component is advantageous and promotes cooperation. It must not be either excessive or lacking. We also show that we should not allow guilt-prone agents to feel guilty towards non-guilt prone ones, for the former would suffer abuse by the latter, thereby discouraging the advantages of guilt.

This returns us to the central problem of morality which naturally also affects machines: How can we avoid raising purely selfish agents who take advantage of opportunities to perform desired actions without considering implications at the level of the community, thereby demonstrating (or failing to demonstrate) cognitive capacities emergent in human beings over evolutionary time? In other words, how can we demonstrate, through computational mathematical models, under what circumstances characteristically human gregariousness is evolutionarily possible, stable and advantageous?

In our research, we have managed to use computers to better understand how the machinery of guilt works, between which values of which parameters, varying 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 can be a justified result of our evolution. That is, because guilt is useful, we have through evolution become capable of having it, and thereby to be capable of inducing guilt in others. Such principles also help to explain the fact that we have a Catholic religion very much based on the notion of guilt. In this context, 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

[18] For the technical details consult L. M. Pereira, T. Lenaerts, L. A. Martinez-Vaquero, T. A. Han (2017).

computer science expression. These principles only thereafter emerge in various forms, as different principles in different religious and cultural constructs, yet all express basic mechanisms such as guilt, shame, and apology.

Discussion

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. One of the problems confronting contemporary cultural constructs in the context of machine ethics is that Jurisprudence is not progressing quickly enough given the speed of technological advance. Though there are various types of machine autonomy, human laws are made for human beings who we assume have certain core cognitive capacities underwriting morality, 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 of moral behaviour specific to the machines in their roles in their organizations, as members of swarms or otherwise.

It is also important to recognize that legislators are tardy in pursuing these issues in these terms, from the point of view of various emergent machine moralities, given the pace of technique. This is worrying because there is simultaneously operative a confused notion that technical progress is tantamount 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, including the creation of extra wealth being distributed fairly.

Duly analysed history can recommend 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, essentially constituted by armies conquered and foreign citizens. Similarly, we have the possibility of enjoying more and more machines as slaves, as it is already happening to some extent, for they liberate us from effort that can be allocated to them without moral claim to equal treatment. But we would like everyone to be freer and equally profit from 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. And, situations in which the human has no possibility of competing with machines proliferate. Hence, as a consequence of the impact of such technologies, a new social contract is indispensable in which the labour/capital relationship is reformulated and updated taking into account the increase in the sophistication of machines with cognition and autonomy.

If a machine is going to replace me, a human being, it should do it completely including social obligations, even if this requires that they express certain core cognitive capacities. Consider that, through the wealth created by my work, I contribute to the Social Security that supports the current retirees. In doing so, I contribute to the National Health Service, and 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 the human has to mean replacing in all these aspects!

Another noteworthy point linked with the safety of technologies is signalled here. 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, repeated, ongoing collaborative study reports of unsuspecting entities, namely from McKinsey&Company (McKinsey 2017, 2018), the Pew Research Center (Pew 2018), the OECD (2018), PricewaterhouseCoopers (PwC 2020), etc., which point to an increase of 15-20% in additional unemployment in 2030, by virtue of AI alone. In China the displacement promises to be even more dramatic, reaching even 20% because, while in the Western world people can still access some social mobility, becoming cognitively specialized given a high educational starting point, in China the level of education is generally lower and therefore the ability of people to rise in their cognitive abilities and to stay ahead of increasingly intelligent machines is lesser and slower. Imagine, then, the employable 1.4 billion Chinese and the impact on them of ever more sophisticated AI. The topic of increased unemployment caused by AI in each of the very AI superpowers creating it, and which will be heavier in less developed countries, is well analysed in the recent book by Kai-Fu Lee (2018).

The real dangers of AI do not fit into the possibility of appearance of a Hollywood-like “Terminator”. Actual risks are that, at present, simplistic machines are making decisions that affect us without adequate human account of the implications. Though, by calling them "intelligent machines," people believe they are doing a good job. Such currently ongoing excessive selling of AI is quite pernicious in that respect. In addition, the AI now being sold is less than one tenth of what AI science is, and what applied AI may actually be. Serious AI is yet to come and will be much more sophisticated than current programs dubbed deep learning ones. The latter are quite limited, and we should not be giving so much power to such simplistic machinery, notably outside their circumscribed area of usage. But since they can to some extent replace humans, like radiologist technicians, car and truck drivers, call centres attendants, people in shopping centres security, they are oversold as profit making panacea. Hopefully optimistic that further progress will deliver them from their problems, humanity finds itself in the role of Epimetheus, and what we do next may set the stage for the future of the race, if not all animal life on Earth, as we uncritically move forward.

The author is a part of the project titled "Incentives for safety compliance in AI Race"[19] sponsored by the Future of Life Institute (FLI), a non-profit organization[20]. 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[21]. 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, would merit a separate article of its own. Here I wish only to alert the reader to the need for parity with other initiatives in the context of human flourishing, such as the "National Bioethics Commission". It will have to be independent, to be above all interests, and respond directly to the President of the Parliament, without depending on the whims of government for its continued legitimacy.

We cannot accept, as we hear in Europe and the United States, that companies that make driverless cars are exclusively responsible for them, and that if there is found a problem only then shall we see to it. We cannot accept that government officers are not responsible for the tests to which such cars should be subjected, or that they may simply delegate this responsibility to the companies themselves. In retrospect over recent events, we

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

[20] https://futureoflife.org

[21] For a summary of the project see T. A. Han, L. M. Pereira, T. Lenaerts (2019).

are afforded foresight over how such a process may unfold. Simply look at the recent crashes of the Boeing 737-MAX, in which the American Federal Aviation Authority (FAA) delegated quality checks to Boeing itself. Those who fail to learn from history… 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 indications, the result is some recommendations in the form of guidelines that firms should follow, in addition the nomination of private audit firms that will inspect those firms, while in parallel commissioning studies. So given, perhaps we will fall into the same scheme of investigators with interests in the affirmation of the very entities that they are tasked with examining such as that evident in the Boeing 737-MAX fiasco. Hence, the need for some independent, public 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 mention whenever suitable - is to give excessive power to simplistic machines, because of the risk that exists that they increasingly, systematically, put us into statistical drawers. What these deep learning machines do with big data is recognize specific patterns in the context of possible patterns. For certain things, it is great. For recognizing cancers in soft tissues using imaging that this not sensitive enough for the human eye, it's an excellent technique. But such machines cannot solve problems for which this technique is inadequate. They are unable to prospect, foresee and choose on the basis of reasoned hypotheses that justify and explain choices about the future. Nor can they reason counterfactually, knowing what we know now, about the consequences of alternative past choices[22]. For these pattern recognising machines, patterns are fixed, the future mimics the past. People and options are put into statistical drawers.

Note the following very emblematic example: In the US (ProPublica 2016; WIRED 2017) 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) at any given time. So, they will see in a Big Data record on people who have been paroled whether or not things 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. This

[22] Cf. L. M. Pereira and A. Saptawijaya (2017) plus L. M. Pereira and F. C. Santos (2019).

assumption is effectively contrary to what our EGT research has shown, that a population with a certain profile evolved along with social customs, with the lesson that this evolution must continue if morality is to remain anything like human morality. Instead, with machine learning, human beings with evolving futures are judged solely by the historical standard, a that standard – ironically - will be confirmed by yet another instance of its application. Morality is thus reduced to a self-fulfilling prophecy made from the bones of the past.

This is a clear example of real and effective misuse of simplistic algorithms with ethical implications. 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 application case, it is argued that the judges, busy as they are, and especially after lunch, decide worse! Yet, these processes are not confined to criminal law. They are also moving to an area that tells us even more of the role of technology in replacing distinctly human moral cognitive capacities. In the context for instance of care, consider 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 updated knowledge[23]. In the final analysis, there are no diseases in themselves, they always occur in a patient with a personal 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 instruments of measurement, with distinct granularity and obsolescence. There will be mistakes, and much as with the car that kills the pedestrian without human input, here we see, more systematically, machine algorithms making prognoses that inform actions resulting in life or death for those affected.

Conclusion:

We close this article with a short synthesis of the "terms of reference" for the AI scientific community, with finishing remarks on the evolutionary dimension of cognition. First, the topics that make up this problem area are summarised:

- 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.

[23] As an example, 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

- 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. Moreover 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, 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, begin to break down. All life is an evolutionary stage, where replication, reproduction, and genetic recombination have been testing individual group-combined solutions to intergenerational coordination problems, resulting in improved moral cognition and ethical action. Thus, the current state of knowledge in AI leads us to anticipate the redefinition of the human being's place in the world, posing challenges to several domains of inquiry all at once. From the onset these are problems for Philosophy, tackling questions such as what knowledge is, what humanity is, what values are and how morality may open individual agents to otherwise unthinkable perspectives in the interests of their society.
As far as ethical and concrete moral models and their evolution are concerned, the possibility arises of their objective simulation in computers, helping to overcome limits previously imposed by often subjective arm-chair speculation, even if shared. With regard to anthropological questioning, the traditional discussion "What is humanity?” is now replaced by a powerful and challenging problem around what is desirable, possible or likely for humanity to become given the anticipated crossbreeding of AI, genetic engineering and nanotechnology. 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 is 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 pressing problem associated with the impact on work and its inherent dignity, as well as with the production and distribution of wealth. Again, we anticipate that a reconfiguration of economic relations will result, not only from the automation of routine activities, but fundamentally from the arrival of robots and software that can replace doctors, teachers, and assistants in nursing homes (to mention caring professions which are not commonly believed to be replaceable by robots). Knowledge specifically about essential moral cognitive capacities 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 ethics becomes urgent as the knowledge ecosystem is greatly enriched by machines with increasing ethical impacts, and machines become active players in dimensions that, until now, have been attributed exclusively to humans.

If we do not exercise appropriate foresight, 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 unfortunates bound to suffer every manner of misfortune in the transition. 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 may eventually prove unstable. People will no longer be able to endure so much intended hypocrisy, so much lack of distribution of the wealth generated, so much automated fake news lies, and so on. Without proper values guidance there may be chaos. How we may commission the creation of machines that afford this guidance is a challenge of the ages, no more pressing than it is now. Should the future reflect lessons learned, the myth of Pandora may be replaced with an evolved version in which the irreversible cost of catastrophic error due to reckless optimism may be recognized from the start.

Acknowledgements

The author is very grateful for the many relevant comments, alternative wordings and suggestions from reviewer Jeffrey White (a native language speaker), almost all of which were adopted. The author is also very grateful for the extensive comments and suggestions proffered by António B. Lopes on an earlier draft of this paper. Support is acknowledged from grant RFP2-154 of the "Future of Life Institute", USA, and from project FCT/MEC NOVA LINCS PEst UID/CEC/04516/2019 of the "Fundação para a Ciência e a Tecnologia", Portugal. The author is also thankful for the 1-hour invited lecture, recorded in Portuguese, at the "Horizontes do Futuro" series promoted by town council "Câmara de Loulé", Portugal, on 21 February 2019, from whence an antecedent of this text initially arose.

References

Cave S et al. (eds.) (2020) AI Narratives – A History of Imaginative Thinking about Intelligent Machines. Oxford University Press, Oxford.

Hammerstein P, Stevens JR (eds) (2012) Evolution and the Mechanisms of Decision Making. The MIT Press, Cambridge.

Han TA, Pereira LM (2018) Evolutionary Machine Ethics. In: Bendel O (ed), Handbuch Maschinenethik, ISBN 978-3-658-17484-2, pp. 229-253, SpringerVS, Wiesbaden.

Han TA, Pereira LM, Lenaerts T (2019) Modelling and Influencing the AI Bidding War: A Research Agenda. In: Markham A et al (eds), Procs. AAAAI/ACM Conference on AI, Ethics, and Society, (AIES 2019), 27-28 January 2019, Honolulu. AAAI, Palo Alto.

Kershaw SP (2007) A Brief Guide to the Greek Myths. Constable & Robinson Ltd., London.

Lee K-F (2018) AI super-powers – China, Silicon Valley, and the New World Order. Houghton Mifflin Harcourt, New York.

Mayor A (2018) Gods and Robots – Myths, Machines, and Ancient Dreams of Technology. Princeton University Press, Princeton.

McKinsey&Company (2017) JOBS LOST, JOBS GAINED: WORKFORCE TRANSITIONS IN A TIME OF AUTOMATION. McKinsey Global Institute. https://technologyreview.us11.list-manage.com/track/click?u=47c1a9cec9749a8f8cbc83e78&id=66f78fce4f&e=d1762c0ec8. Accessed 1 May 2020.

McKinsey&Company (2018) Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey Global Institute. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy. Accessed 1 May 2020.

Moor, JH (2006) The Nature, Importance, and Difficulty of Machine Ethics. IEEE Intelligent Systems, pp.18-21, July/August 2006. http://www.psy.vanderbilt.edu/courses/hon182/The_Nature_Importance_and_Difficulty_of_Machine_Ethics.pdf

Nuremberg (1945-46) Superior orders. Wikipedia. https://en.wikipedia.org/wiki/Superior_orders#%22Nuremberg_defense%22 . Accessed 1 May 2020.

OECD (2018) Putting faces to the jobs at risk of automation. OECD POLICY BRIEF ON THE FUTURE OF WORK. https://www.oecd.org/employment/Automation-policy-brief-2018.pdf . Accessed 1 May 2020.

Parret H (1993) The Aesthetics of Communication: Pragmatism and Beyond. Springer-Science+Business Media, B.V., ISBN 978-94-010-4779-1, Dordrecht.

Pereira LM, Lenaerts T, Martinez-Vaquero LA, Han TA (2017) Social Manifestation of Guilt Leads to Stable Cooperation in Multi-Agent Systems. In: Das, S. et al. (eds), Procs. 16th Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2017), pp. 1422-1430, 8–12 May 2017, São Paulo, Brazil. AAAI, Palo Alto.

Pereira LM, Lopes AB (2020) Machine Ethics: From Machine Morals to the Machinery of Morality. In: Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE, volume 53), XXV+164 pp., ISBN 978-3-030-39629-9, Springer Nature, Cham.

Pereira LM, Santos FC (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, ISBN 978-3-030-32721-7, pp. 69-82, SAPERE series, ISSN 2192-6255, vol. 49, Springer, Berlin.

Pereira LM, Saptawijaya A (2016) Programming Machine Ethics. In: Studies in Applied Philosophy, Epistemology and Rational Ethics, (SAPERE, volume 26, 194 pages, ISBN: 978-3-319-29353-0, Springer, Berlin.

Pereira LM, Saptawijaya A (2017) **Counterfactuals, Logic Programming and Agent Morality.** In: Urbaniak R, Payette G (eds), Applications of Formal Philosophy: The Road Less Travelled, Springer Logic, Argumentation & Reasoning series, ISBN: 978-3319585055, pp. 25-54, Springer Nature, Cham.

Pereira LM, Saptawijaya A (2018) From Logic Programming to Machine Ethics. In: O. Bendel O (ed), Handbuch Maschinenethik, ISBN 978-3-658-17484-2, pp. 209-227, SpringerVS, Wiesbaden.

Pew RC (2018) ARTIFICIAL INTELLIGENCE AND THE FUTURE OF HUMANS. Pew Research Center *Internet & Technology*. https://www.pewresearch.org/internet/2018/12/10/concerns-about-human-agency-evolution-and-survival/ . Accessed 1 May 2020.

ProPublica (2016) Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing . Accessed 1 May 2020.

PwC (2020) How will automation impact jobs? PriceWaterhouseCoopers *Insights*. https://www.pwc.co.uk/services/economics-policy/insights/the-impact-of-automation-on-jobs.html#cta-1 . Accessed 1 May 2020.

White J. (2010) Understanding and Augmenting Human Morality: An Introduction to the ACTWith Model of Conscience. In: Magnani L., Carnielli W., Pizzi C. (eds) Model-Based Reasoning in Science and Technology. Studies in Computational Intelligence, vol 314. Springer, Berlin, Heidelberg.

White, J. (2012) Manufacturing morality, a general theory of moral agency grounding computational implementations: the ACTWith model. In: Floares, A. (ed.) Computational Intelligence. Nova Science Publishers, Hauppauge.

WIRED (2017) Courts Are Using AI to Sentence Criminals. That Must Stop Now. WIRED. https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/ . Accessed 1 May 2020.