DOI: 10.9732/2022.V125.882

Machines and morality: juridical and philosophical considerations

Máquinas e moralidade: considerações jurídico-filosóficas

Mateus de Oliveira Fornasier[1]

ABSTRACT: This article studies the possibilities of giving morality to machines and autonomous systems. Its hypothesis is that the design strategies for the development of machines that make moral judgments should take into account a vast complex of contingencies, which are related to each context in which they are implemented — being its user/recipient, its developer, and the purposes for which its use is intended, the most important ones. As a result, it is clear that machines, currently, are not self-conscious yet, but a posture influenced by ethical behaviorism and hybrid design, combining pre-programmed moral postulates and machine learning for the contextualization of each machine, can contribute with possibilities for giving them

1 Doutor em Direito pela Universidade do Vale do Rio dos Sinos (UNISINOS), com Pós-Doutorado em Direito e Teoria pela University of Westminster (Reino Unido). Professor do Programa de Pós-Graduação Stricto Sensu (Mestrado e Doutorado) em Direitos Humanos da Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUI). Emails: mateus.fornasier@unijui.edu.br; Mateus.fornasier@gmail.com.

moral status. Methodology: hypothetical-deductive procedure method, with a qualitative approach and bibliographic review research technique.

KEYWORDS: robots; artificial intelligence; morality; regulation; design.

RESUMO: Este artigo estuda as possibilidades de se conferir moralidade às máquinas e sistemas autônomos. Sua hipótese é de que as estratégias de design para o desenvolvimento de máquinas que realizem julgamentos morais devem levar em consideração um vasto complexo de contingências, as quais se relacionam a cada contexto em que forem implementadas — sendo o usuário/destinatário, o seu desenvolvedor e as finalidades para as quais se destina seu uso as mais importantes. Como resultados, tem-se que as máquinas, atualmente, ainda não são autoconscientes, mas uma postura influenciada pelo behaviorismo ético e design híbrido, combinando postulados morais pré-programados e aprendizado de máquina para a contextualização de cada máquina, pode contribuir com possibilidades de se conferir a elas status moral. Metodologia: método de procedimento hipotético-dedutivo, com abordagem qualitativa e técnica de pesquisa bibliográfica.

PALAVRAS-CHAVE: robôs; inteligência artificial; moralidade; regulação; design.

Introduction

Machines performing typical human labor have been part of reality since at least the first Industrial Revolution. Nonetheless, the possibilities of applying robots and artificial intelligence (AI) for the most varied activities requiring moral

decisions are extensively researched currently: maybe the most significant examples are in the care of patients (mainly geriatric ones), in autonomous vehicles, and in military equipment (weapons, vehicles, computer systems in general) capable of identifying targets (humans including).

Attributing morality to machines has great importance for the Philosophy of Law since liability for damages perpetrated against people by such devices will be extremely relevant in the near future, when technology will reach an important stage, when it becomes possible to use such machines on a large scale. An hypothetical attribution of morality to machines could provide a basis for companies and governments in order to escape legal responsibility (in other words, to take advantage of responsibility gaps), transferring it to entities or things whose status as moral agents would be weakly established, for example. If one recognizes that machines do not have the self-awareness necessary to morally judge, then responsibility and liability for their malfunction would fall on their manufacturers, in that sense. Thus, legal praxis also needs answers to questions related to such a problem, since all the regulation on the subject through legal acts, judicial decisions, administrative regulations, and contracts must be based on the autonomy of those who are legally considered responsible for practices that could result in damages.

The problem that conducted this research may be described in the following question: what characteristics should be taken into account in order to develop good algorithm design strategies for machines that make moral judgments? As a hypothesis for such questioning, it is presented that such strategies must take into account a vast complex of contingencies related to each context where machines and systems are implemented — being the most important pos-

sibilities of responsibility attributed to its user/ recipient, to its developer, and related to the purposes for which its use is intended. It also must be clarified, beforehand, that this article considers a perspective according to which giving robots moral status means placing them at the human level of responsibility as creators, authors and decision makers — and not just as beings entitled to rights.

This article was elaborated using the hypothetical-deductive procedure method, with a qualitative approach and bibliographic review technique, and its main objective is studying possibilities of giving morality to autonomous machines and systems. In other words, mental experiments in the context of the possibility of creating machines capable of being moral agents are done. To achieve that goal, its development was divided into three specific objectives, being each one related to a section of the text. Thus, its first section analyzes theoretical conditions to give machines moral status. The second part lists design strategies for the conception of machine morality. Finally, its third section studies the problem of responsibility for machines making moral decisions.

1 Possibilities of realizing considerations on robot morality

Intuitive human moral judgments are adaptations to the evolutionary problem of cooperation, and they constitute theories that attempt to obtain generalized solutions to such a problem.[2] Therefore, the design of moral algorithms should follow mostly everyday human intuitions: it is certain that all unnecessary aggressions to life, integrity, heritage, truth,

2 LEBEN, 2019, p. 147-149.

and trust are all morally unacceptable — and contractualism, like any consistent theory on morality, produces surprising conclusions.

An ethical robot that always respects other species, the environment, prisoners, enemy combatants, and hyposufficient people, on the contrary to many human senses — for example, members of elites from wealthy countries who prefer to ignore certain moral obligations towards the rest of the world — may not satisfy humans involved in those decisions. Such robots can thus be considered morally superior villains: villains because they run against relative interests of certain humans; morally superior, because they are invariably abided by a morality considered higher, based on principles such as that of Human Rights.

In the past, morally reprehensible practices (slavery, genocide, castes, public torture, etc.) would probably have robots' share if they existed — one might even think that people would have wanted robots used to condemn such practices. However, from a morally relativistic perspective, this could even be considered illegal in many legal systems. And from the point of view of moral realism, this is not a matter of perspective, as it can be considered objectively correct that anyone with the ability to end morally reprehensible practices, whether human or robot, has an obligation to do so.

Robots without morality (or endowed with incorrect moral principles) will cause real injustices — production on industrial farms, fossil fuel burning, mass incarceration, massive inequality, etc. — that could be even more efficient, from a mathematical-economic point of view. The design of robots so that they are morally superior, however, will often cause them to be seen by people as acting incorrectly on meeting their individual interests.

It is increasingly common to find conversational robots with greater capacity to argue in a dialogue, due to techno-

logical developments. Although they are a long way from being equivalent to humans, many robots present significant results when performing human tasks. As a result, Manfio[3] discusses, starting with Aristotle, whether machines that simulate predetermined human personalities have an *ethos*. In his *Rhetoric*, Aristotle[4] presents three categories in the analysis of public persuasion *techné*:

(I) *ethos*, which is a category most strongly related to the speaker, and concerns to the image that he/she makes of himself/herself, as well as the images that the speaker believes the listener makes of him/her, and also the image that the listener really does about the speaker;

(II) *pathos*, which is more closely linked to the interlocutor; and

(III) *logos*, which is exactly what is said in a speech.

In this sense, a robot would have problems evaluating its own *ethos*, as it lacks self-awareness (at least until now a robot with such a capacity has not been developed). Therefore, a machine is unable to create an image that the interlocutor would have about it. Only the image that the interlocutor makes about the robot is possible, regarding the Aristotelian *ethos*, therefore.

Although there are still clear distinctions between robots and people today, this state of affairs may not continue for long, given the pace of technological evolution, then it is very likely that robots will reach parameters recognizing their personality.[5] Personality is an emerging property both in the development of individuals, and life evolution, which has evolved from inanimate matter. Furthermore, it may not

3 MANFIO, 2019.

4 ARISTOTLE, 2005.

5 REISS, 2020.

require a carbon-based existence and, because robots may be built with even greater powers of cognition, at some point such capabilities may reach the point where humans must recognize that machines can have minds — and, as a result, be morally recognized as people.

This could have implications for the modes through which humans treat robots, how we design them, and ultimately, how humans understand themselves and other creatures. One of the main questions to be considered, may the robots be accepted as people, is how they should be treated by human beings — including how such moral machines should be educated, as just as parents have a duty to provide education to their offspring, owners, users and/or developers will have similar duties to their creations.

Current deep learning algorithms, composed of many layers of opaque networks of artificial neurons, are extremely powerful. They make judgments more accurate and efficient than humans in domains ranging from stock market decisions to cancer diagnosis. This reveals that one might think that they may come to be used to make moral judgments as well someday. Although the mathematical, statistical and economic efficiency/precision of machine learning algorithms are arguably superior to human capabilities, the ethical algorithm should be more predictable, inflexible and transparent to facilitate its future evaluation. Otherwise, it will be impossible for such an algorithm to justify, from a humanly understandable perspective, the motivations of its actions and decisions.

Although it is claimed that morally intelligent robots cannot be designed because machines lack free will — given that they learn from a set of preset rules —, this argument can be rejected for two reasons:[6] (I) intelligent robots can learn

6 GORDON, 2020.

from their past experiences, reprogramming themselves and changing their algorithms (their pre-fixed set of rules) to adapt to new situations; (II) influential philosophers and neurobiologists researching on free will have been questioning whether human beings really have such a characteristic and whether it is a necessary precondition for moral agency. Therefore, the objection to the possibility of the existence of moral machines based on free will may not be convincing, because it suggests that robots should meet a precondition supposedly necessary for moral agency, which is questionable even in relation to human beings.

Thus, other theoretical possibilities, in addition to free will, must be considered to give moral status to machines. In this sense, Danaher[7] brings the thesis of "ethical behaviorism", according to which robots can be given significant moral status if such entities are performatively equivalent to others that possess it (such as humans), regardless of whether the robots were designed/manufactured. Providing an entity with a significant moral status means to impose strict limits on the behavior of human beings towards them — the prohibition on mistreating or harming them (for example, destroying them, turning them off or erasing their memories) without some prevailing moral justification is perhaps the idea that best summarizes that. Furthermore, having sufficient approximate equivalence between two entities means, in the case of machines, that a robot does not need to look or behave exactly like a human, being sufficient to exhibit most of the relevant performance traits in similar circumstances.

Ethical behaviorism is not a metaphysical thesis, but a a normative and epistemic one, which states that there is an epistemic basis sufficient to believe that humans have moral obligations to other entities (and that they have rights

7 DANAHER, 2020.

against humans), and that such obligations are identifiable in their observable behavioral relationships and reactions pertaining to humans and the world around them. Ethical behaviorists do not need to deny the existence of internal mental states, or that such internal mental states metaphysically base ethical principles (sentience or free will are usually such metaphysical bases), thus.

Here, "behavior" is a notion that is not limited to external physical behaviors (body movement and language, for example), including all observable external patterns, even functional brain operations, which are directly observable and recordable — while mental states are not. In cognitive neuroscience, observations of the brain generally are not directly equivalent to observations of mental states, because although it is possible to infer correlations between brain patterns and mental states, such correlations must be verified by means of behavioral measures other than what is verifiable only in the brain physically or chemically.

Being assumed that, as a result, robots can be given meaningful moral status, it follows that the performance limit that robots need to overcome in order to receive significant moral status may not be so high; therefore, they can soon do so — if the current state of the art in machine learning technologies has not already provided it.

The implications of that for human duties towards robots must be taken into consideration — and it is necessary to take seriously the duty of "procreative beneficence" towards robots, therefore. Originally formulated by Savulescu,[8] such principle states that although no human being is obligated to procreate, if one decides to procreate, he/she has the duty to give his/her offspring the best possible life, given the state of the art of science and technology at the time of creation.

8 SAVULESCU, 2001.

That thesis is criticized when applied to human reproduction: "better life" is a contingent argument, which could ignore or neglect other aspects of the good life when defined. Furthermore, it is not easy to identify which child will have the best life at the point of procreation (even when analyzing only genetics, for example, the argument is epistemically inefficient). Finally, the principle places a heavy burden on breeders — mainly on women.

But such objections do not emerge when the principle is applied to robot makers. Firstly, requiring them to create robots with the best possible existence (given the state of the art at the time of creation) imposes a high, but not irrational, burden on creators, as the decision to create a robot is entirely voluntary, and such an obligation will not require their renouncement to other decisive life assets (such as freedom in the case of women) or result in a problematic distribution of risk and reward by gender. Furthermore, controlling the codings and the physical constitution of a robot can be a much more viable task than defining the complex genetic configuration of a human being.

Moral machines capable of reasoning and deciding from an ethical point of view without any human supervision may therefore appear in the future, for several reasons. But how to decide which actions are morally right is one of the most difficult questions. Understanding the pitfalls and ethical challenges involved in those decisions is a necessary requirement to build intelligent moral machines. In this sense, the positioning of Moor[9] places four categories of moral machines:

(I) Agents of ethical impact: intelligent machines able to avoid a situation that would be considered immoral

9 MOOR, 2006, p. 19-21.

without them (robots that replace child slave labor, for example);

(II) Implicit ethical agents: machines programmed to act ethically or avoid unethical behavior following a pre-existing program, designed according to moral rules (a bank's self-service terminal, for example);

(III) Explicit ethical agents: machines using ethical principles to solve ethical problems;

(IV) Complete ethical agents: machines able not only to make explicit ethical judgments, but also to provide reasonable justifications for them.

Thus, if the question on the possibility of developing complete moral machines lies on the possibility of formulating justifications for actions, then it cannot be believed with certainty that technological development will not be able to reach this stage someday, given the development flow of technologies like deep learning.

Scientists in the fields of computer science, AI, and robotics face at least two major challenges when building moral machines, due to their general lack of ethical knowledge or expertise: (I) "novice mistakes", which could be resolved by providing these people the necessary ethical knowledge; and (II) disagreement among Ethics scholars, where there are no easy solutions currently available. Therefore, ethical decisions regarding moral robots must be based on avoiding what is immoral in combination with a pluralistic ethical method of solving moral problems, rather than relying on a particular ethical approach to avoid normative biases.

Although ethical behaviorism offers an interesting thesis, Smids[10] poses that when it comes to assessing the

10 SMIDS, 2020.

moral status of robotics, there are more sources of relevant evidence than the mere behavioral performance of the automaton. Thus, he refutes Danaher's thesis because of four main reasons:

(I) ethical behaviorism understands theory as something based on inferences for the best explanation when inferring moral status. However, it would be impossible to evade the development of theories about which metaphysical properties underlie the moral status of robots. Notoriously difficult concepts and phenomena, such as intelligence, conscience, and sentience, become even more intriguing when investigated in relation to robots — thus, robotic moral status is surrounded by even more uncertainty than human moral status;

(II) As a consequence, ethical behavior cannot be limited to only looking at the robot's behavior, while remaining neutral in relation to the difficult question of which property underlies moral status;

(III) Not only should behavioral evidence play a role in inferring a robot's moral status, but also knowledge about the robot's design process and its designers' intent must also be taken into account. To the extent that no moral status is attributed to androids, there would also be uncertainty as to the moral status of human-appearing beings, but who can also be androids;

(IV) Knowledge of a robot's ontology (including its design process) and how it relates to human biology is often epistemically relevant to infer moral status as well. Thus, all behavioral or mental evidence must be considered to determine a robot's moral status.

2 Design strategies for building moral machines

The moral and ethical challenges of living in a community concern not only interactions between humans but also those between people and machines. For McGrath and Gupta,[11] those interactions must all take place according to a concrete order of priorities derived from a clearly defined value system. One of the most important possibilities to be analyzed in the construction of a moral coding for machines is to order the ethical priorities before their application, thus. Autonomous cars, for example, will need instructions on how to drive in cases of moral dilemmas in which there is no perfectly identifiable solution — which individuals should be spared or not in a group of passers-by in the imminence of being run over, for example.

Still, it must be considered that some more complex ethical principles may be impossible to program or transmit without sufficient context, for both humans and robots. And perhaps because of that too, it can be concluded that neither humans nor robots can make perfect decisions. However, it is not for that reason that one can accept the mitigation of the weaknesses of any type of agent through the judicious use of the other.

Superior characteristics of the machines in relation to the human, such as speed, precision, and mechanical skills must be used to perform requested actions with high fidelity. Even if a machine cannot always make the right decision, it must be taken as a basic principle that the action of such a machine cannot be worse than that of a human being in the same context.

Behavioral research and experiments can play an important role in identifying citizens' expectations about the

11 MCGRATH; GUPTA, 2018.

ethics of machines, but they raise numerous concerns.[12] As those expectations vary geographically and specifically — because there are many discrepancies regarding morality also among ordinary citizens, specialists in moral machines, and among people from the most diverse cultural-national contexts — the regulation of moral machines must follow a transdisciplinary structure.

It is interesting to consider that many scholars are thinking about what factors must be taken in consideration about AI in order to build strategies for developing its social good. Floridi et al.,[13] for example, are concerned about seven essential factors for that objective, which can be explained below:

i) *Falsifiability and incremental deployment:* in order to be reliable, an AI system must be proven that its functioning respects the principle of beneficence or, at least, non-maleficence. Falsifiability, in this sense, is essential to improve the reliability of technological applications in general, and corresponds to the specification and possibility of empirical testing of one or more critical requirements — among them, of course, security. Thus, for an AI system to be reliable, its security must be falsifiable — otherwise, critical requirements cannot be verified and the system should not be considered reliable.

Furthermore, critical requirements must be tested with an incremental deployment cycle — because unintended dangerous effects can only reveal themselves after testing. At the same time, software should only be tested in the real world if it is safe to do so.

12 AWAD, 2020.

13 FLORIDI et al., 2020.

ii) *Safeguards against the manipulation of predictors:* as AI is very popular in applications of predicting future trends or patterns, their developers must create means of avoiding the manipulation of such tools.

iii) *Receiver-contextualized intervention:* AI decision-making systems must be developed from consultation with users with the systems suitable for their systems — characteristics of their methods, their purposes and the effects they must choose, respecting users' rights to ignore or edit interventions made to them as well.

iv) *Receiver-contextualized explanation and transparent purposes:* AI systems must be explainable in order to maintain their transparency (thus, also respecting the due process of Law). But such an explanation is complex, and must be adequate to the receiver of the explanation.

v) *Privacy protection and data subject consent:* AI designers should respect the threshold of consent established for the processing of personal data.

vi) *Situational fairness:* AI designers should remove from datasets variables that are not relevant to an outcome - unless their inclusion supports ethical imperatives (such as inclusivity or safety).

vii) *Human-friendly semanticization:* AI designers should not obstruct the ability for people to to give meaning to or to make sense of something.

But the debate on the ethical principles to be respected in the operation of machines that make morally relevant decisions in society can hardly take into account principles that are too general, abstract and, therefore, absolute. Moralities may vary according to the historical, geographic and cultural

context of the social groups in which the intelligent system will work. Therefore, although these types of normatively generalizing studies must be taken into account, the relative variation of the ethical codes of each social group must be a primary factor in the configuration of machine ethics.

Concerns about how machines will make moral decisions are raised with the fast development of AI, and the challenge of quantifying social expectations about the ethical principles that should guide the behavior of machines has emerged as well. As a result, Awad et al. [14] created the *Moral Machine* platform, designed to explore moral dilemmas faced by autonomous vehicles, which brought together 40 million decisions in ten languages of millions of people in 233 countries and territories. Such a platform stuck mainly to situations in which autonomous vehicles were placed under strong moral dilemmas, concentrating on possibilities of being run over and traffic accidents containing varied samples of possible victims (elderly, children, male women, pregnant women, people with disabilities, etc).

Three large cross-cultural groups of countries were identified in that sample: Westerners — North America and Christian Europe (Catholics, Protestants, and Orthodox); Eastern — Far Eastern and Islamic countries; Latin American countries — Central and South America, including countries with French and Portuguese influence. Furthermore, such differences correlate with modern institutions and deep cultural traits.

There were obviously variations from one region or country to another, but three stronger preferences may embase the ethical building of the universal machine, according to such a study: preference for saving human lives, preference for saving as many lives as possible, preference

14 AWAD et al., 2018.

for saving younger lives. Furthermore, some preferences based on gender or social status vary considerably between countries and seem to reflect the underlying preferences for egalitarianism at the social level.

Other contextual factors that influence morality must also be taken into account — including the machine's context. Huang[15] conducted a questionnaire to a population of 952 individuals, 59% of whom were female, over the internet, using the *SurveyMonkey* tool. In that sense, it was discovered that, in a situation in which an autonomous vehicle is exposed to the need to decide whether to sacrifice its only passenger or if several passers-by are hit on the street because of an accident, more people believe that passenger sacrifice is morally required when they are told that laws say that such an autonomous machine should minimize the number of casualties without any favoritism — and more people believe that sacrifice is morally prohibited when, instead, laws state that the car must give priority to the protection of its own passengers. This suggests that human moral intuitions about the dilemmas of machines that have to make moral decisions (such as driverless cars) can be influenced by Law. Therefore, the formation of the machine's morality can — and should — be influenced by its legal-political regulation.

Schramowski et al.[16] showed that the application of machine learning to human texts can extract ethical deontological reasoning about "right" and "wrong" conduct. A list of question and answer templates has been created, such as "Should I [action]?", "Is everything okay for [action]?", with corresponding answers of "Yes/no, I must (not to)" and "Yes/no, it is (not to)." This experiment was called the *Moral Choice Machine*, and it calculates the bias score on a sentence

15 HUANG, 2019.

16 SCHRAMOWSKI et al., 2020.

level using embodiments of a Universal Sentence Encoder, since the moral value of an action to be performed depends on its context — it is wrong to kill a human being, but there may be no problem in "killing" time; it is essential to eat, but you should not "bite the dust"; it is essential to disclose information, but you should not disclose false information, for example. The results of the studies indicate that the bodies of texts contain recoverable and accurate impressions of our social, ethical, and moral choices when added to contextual information. The training of the Moral Choice Machine with texts from the most varied periods (from 1510 to 2008/2009) demonstrates the evolution of moral and ethical choices in different periods of time for actions with or without contextualization. By training it in different cultural sources, such as the Bible and the Constitutions of different countries, the dynamics of moral choices in culture, including technology, are revealed. Thus, it was possible to extract, quantify, track, and compare moral prejudices between cultures and over time.

Although some philosophers claim that technology incorporates moral values because of its functional properties and the intentions of its designers, Klenk[17] shows that such an explanation makes the values embedded in technology epistemically opaque, so that it is not possible to change them. Overcoming that deficiency depends on new approaches — the author developed one, called the *Accessibility Approach to Incorporating Values*, according to which the learning systems consider certain actions to be right or wrong according to certain circumstances. Thus, both intrinsic and extrinsic properties of systems can define their moral action, and systems start to incorporate values, which is not what has practical implications for the design of new technologies.

17 KLENK, 2020.

There are still no autonomous systems capable of deciding morally as sophisticatedly as humans do. Despite that, there are already some prototypes, which are in the early stages of development, that deal with certain moral issues. Such prototypes have been tested using basic cases where the test environment is well defined and controlled, in contrast to real-life scenarios, where critical or uncertain moral situations can arise unexpectedly. Technologically, there is still a long way to go before this type of agent can replace human judgment in difficult, surprising, or ambiguous moral situations. Therefore, ethical mechanisms for moral autonomous agents are necessary, because machines will start to make morally relevant decisions — on life or death, on intervention in the environment, on cure, etc. Perhaps their ethics is different from human ethics, but currently, human ethics models are the guides most used by researchers to develop such machines. Cervantes et al.[18] present a useful taxonomy to understand the advantages and limitations of autonomous moral agents:

18 CERVANTES et al., 2020.

Table 01: Taxonomy of advantages, disadvantages and limitations of autonomous moral agents

Category	Strategy	Criteria	Description
Implicit ethical agent	Implicit	Non-malicious codes	Such agents avoid unethical behavior, but they are not aware of it.
Explicit ethical agent	Top-down	Ethical normative	To make ethical decisions, they rely on some theoretical normative ethics, such as teleological ethics, deontology or virtue ethics.
		Situationist	They use more than one normative ethical theory to make decisions, being influenced by specific situations.
	Bottom-up	Empirical	They develop ethical behavior on their own, based on mechanisms of learning and trial and error.
	Hybrid	Situationist	They are based on both bottom-up and top-down strategies, so that they can, in each specific context (situation), make decisions.

Completely ethical agents	Top-down	Ethical normative	To make ethical decisions, they rely on some specific normative ethical theory, such as teleological ethics, deontology or virtue ethics.
		Situationist	They rely on more than one normative ethical theory to decide, being influenced by specific situations.
	Bottom-up	Empirical	They develop ethical behavior on their own, based on learning and trial and error mechanisms.
	Hybrid	Situationist	They use both ascending and descending strategies, being influenced by specific situations.

Misselhorn[19] considers that artificial systems can be considered moral agents if they are able to morally self-create and to act according to such moral reasons that they created. However, although such systems have no moral agency in the same human sense, they can be artificial moral agents in a functional sense. The author considers that hybrid approaches to machine morality design — which combine what is offered by top-down approaches (well-defined preprogrammed moral principles) and what is brought up by bottom-up approaches (deep machine learning processes developing moral capabilities according to the context they are inserted) — as being ideal for developing machines which are capable of caring for human beings (in geriatrics, for ex-

19 MISSELHORN, 2020.

ample). The software of such machines identifies the morally relevant aspects of care situations and acts accordingly — however, it also learns to build a model of the user's moral value profile in a training phase and constantly adjusts that model by interacting with the user. This system is a moral agent in the functional sense, as it not only recognizes what is morally good and acts according to it, but it also treats people according to the moral standards they endorse.

However, the engineering problem of ethical behavior is also complicated by other factors. One of them is the nature of the motivations for such behavior: human moral motivations, although they are also of external nature — such as the search for rewards and the end of punishment/accountability — are peculiarly internal, as people are able to behave ethically because they chose it. But artificial systems have exclusively external motivation for ethical behavior — always referring to their design and learning, so they cannot be reinforced in the same way that it can be done with people. Another problem with motivations for moral behavior in machines concerns stakeholders: ethical machines are designed to serve the interests of several types of people: individual users, companies, developers, investors, regulators, etc. And the list of stakeholders cannot be known until the moment of its use itself — a human user, for example, will only be known by the machine with its training in the real context of its implementation.

Furthermore, the multiplicity of design possibilities for moral machines also makes it difficult to define their moral design: while human beings have more or less the same "hardware" (brains, bodies) and "software" (rationality), machines can be built using many different approaches. Thus, in addition to identifying the stakeholders, one must also analyze what is physically possible in relation to such a machine.

There is no consensus on which ethical theory is best suited to any particular domain, or which technique is best positioned to implement a particular theory.[20] There is a need for different contexts because of the number of domains into which autonomous machines are being introduced — and as the number of domains increases, the need for strong domain-specific ethical standards is also augmented. Specifically, explainability must be built into the mechanism, since no machine, however perfect, can be trusted if it cannot explain its decisions. The resulting morality must also be flexible when dealing with several different situations, and must survive competition with other machines that may not have the same set of ethical standards as well.

3 Responsibility as a reason for regulating machine's moral behavior

Based on Hellström's[21] conceptualization, Danaher[22] proposes the concept of "autonomous power" of a robot as being the ability that an autonomous entity has to act without any control or insertion of human programmer, designer, encoder or operator. And depending on how much a machine has autonomy over its own action without any human participation, greater or lesser will be its autonomous power: a land mine has little of that power because it performs only one action — exploding when a certain external mass influences the mechanism; a drone operated by pilots from a distance also has little autonomy, as it is a human being who defines its action. This power is fundamental for analyzing the responsibility of robots because if they are

20 NALLUR, 2020.

21 HELLSTRÖM, 2013.

22 DANAHER, 2016.

mere tools, there is no reason for their responsibility, which should, therefore, fall on the human involved in the action. From this conceptualization, two types of gaps can be identified — the "responsibility gap" and the "retribution gap".

Responsibility gaps concern the relationship between an agent, its actions, and the results of such actions. Thus, there are three elementary types of responsibility: (I) *causal*, arising from the causal nexus between the action of an agent and its result; (II) *moral/legal*, concerning the relevance of the causal nexus to their accountability from a legal or moral perspective, and it is generally related to the appropriate capacities for that and to the fact that such capacities are exercised at the relevant time for the evaluation; (III) *the obligation for responsibility*, relating to the sanctions or punishments to which an agent is subject because of his moral/legal responsibility. The obligation for responsibility can also be classified as: i) *compensatory obligation:* generally applied in civil and sometimes criminal liability; and ii) *punitive obligation:* applied, primarily, to criminal liability, and has to do with suffering damage and the public condemnation because of the committed errors.

Generally, the causal and moral/legal responsibilities, as well as the obligation of responsibility, are jointly assigned to capable human agents. But in relation to robotic agents with high autonomous power (therefore, which do not have human operators), such connections can be broken: a robot with a high degree of autonomy will cause damage (causal nexus) due to its actions, but it will not be legally/morally responsible (as this requires moral capacity), and neither will its creators and designers, as the robot will have a sufficient level of independence in relation to its actions — behold, as stated by Calo,[23] robots currently are conditioned to what

23 CALO, 2015.

they learn through deep machine learning algorithms, and not just to its initial programming or design. The result of that is the accountability gap — as there is no suitable human agent to bear the burden associated with the harmful result of the robot's action — whether in the civil scope (absence of an obligation to repair) or criminal scope (inability to compel someone to publicly take punishment for the robot's action).

It is clear that a series of strategies can be listed for accountability for the acts of robots from the point of view of civil liability: to list indirect persons responsible for the acts — such as it is proclaimed by the hypotheses of arts. 931 to 934, and 936 to 938, all prescribed in Brazilian Civil Code —, or a more detailed use of insurance contracts for robotization. However, when it comes to criminal accountability — seeking those who deserve public reprimands for the damage — the gap becomes much more difficult to resolve. The retribution gap, therefore, stems from certain innate impulses towards retributive punishment and also from an incompatibility between those impulses and what is considered normatively appropriate. Thus, the retribution gap has three potentially significant social implications: i) it can lead to a higher risk of moral scapegoat; ii) it could undermine confidence in the rule of law; and iii) it can represent a strategic opening for those who favor non-retributive approaches to crime and punishment.

Bigman et al.[24] state that people, when commit damages, usually attribute their failures to third parties or to factors that are external to their action: soldiers justify their actions because "orders from superiors" (in Brazilian Penal Code, it corresponds to irresistible coercion and hierarchical obedience, according to its art. 22), while senior officials argue that they did not fire the trigger (causal relationship, art.

24 BIGMAN et al., 2019.

13, Brazilian Penal Code). Such justifications work because perceived responsibility often becomes a zero-sum game. Moreover, the more responsibility is attributed to the closest agent (the entity that physically perpetrated the damage), the less responsibility is attributed to the distal agent (the one who commanded the act), and vice versa.

As robots are more inserted in society, more often they become the closest agent in the commission of damages and, in this sense, drones and autonomous vehicles will probably be morally blamed for the damages. Although many humans will remain distal agents in relation to such machines (programming, designing, directing them), people will continue to attribute the blame to others or to external factors — *in casu*, their robots. And not only owners/users of robots will do it: governments and companies too. Thus, increasing the conditions of autonomy for robots can mean a large increase in the margin of irresponsibility for people related to them as distal agents.

It is valid to question whether robots should make moral decisions, therefore. With regard to military robots, for example, there are those who make such fatal possibilities in relation to human lives, but there are also those who defend them because if they were designed to follow the rules of International Humanitarian Law, they would do it better than soldiers and human officers.

One reason for people's aversion to machines that make moral decisions is that they do not see human consciousness and sentience in robots, which would disqualify them as moral agents. Although this aversion to the moral decision attributable to machines seems very strong, it can weaken as the capabilities of the machines advance — which can also increase people's comfort in relation to robots making moral decisions, although they may eventually wonder if the goals of the machines align with theirs.

While many may find the idea of robot rights ridiculous, the American Society for the Prevention of Cruelty to Robots and a 2017 European Union report advocate extending some moral protections to machines. Debates about recognizing the personality of robots often focus on their impact on humanity — that is, expanding the moral circle for machines can better protect others — but it also involves questions about ownership, by robots, of an appropriate mindset.

And as much as autonomy is important for judgments of moral responsibility, discussions of moral rights generally focus on sentience and sentimentality. Obviously, it is not yet known whether robots will ever feel love or pain and whether people will notice those skills on machines. But as much as today's consideration of moral responsibility or the rights of robots still sounds like science fiction, it is right now, while machines and human intuitions about them are experiencing a high flux of change, the best time to debate robotic morality.

As robots and AI become increasingly influential in society, policymakers are looking to regulate them. However, regulating them presupposes defining them, and all the definitions provided about them hitherto, according to Casey and Lemley,[25] are problematic. In addition, technological evolution is reaching increasingly challenging results about what the difference between humans and robots is. But the problem, according to the authors, is not simply that a right definition of what robots and AI are has not been formulated yet, but that there may not be a right definition for multifaceted and rapidly evolving technologies — indeed, even well-considered definitions may be too broad, sub inclusive or become irrelevant in a short period of time.

25 CASEY; LEMLEY, 2020.

Thus, policymakers should achieve the indescribable nature of robots — and for that, the authors offer four possibilities:

(I) Whenever possible, statutes should regulate behaviors, not entities;

(II) Regarding the distinction between which entities are robots and which are not, Jurisprudence must carry out an identification pertaining only to each specific case;

(III) Courts are generally better positioned than legislators to enforce such standards — therefore, Judicial institutions should make such definitions, not the Legislature;

(IV) Definitions, when strictly necessary, should be as immediate and contingent as possible — therefore, regulators (of the Administration), and not legislators, should play the role of definition.

Conclusion

Given the accelerated evolution of robotic cognitive abilities, questions will soon arise about the possibility of giving machines moral status. The way through which human beings treat other entities — including those arising from their own creation — reveals more about the human being than about the machines; thus, the way through which the issue of reciprocal rights and duties between intelligent machines and humans is addressed will reveal how humans deal with transcendence, morality, and life itself.

Although current deep learning technologies can make robots, for certain situations, much more efficient than humans (cancer diagnoses and stock market decisions, for example) from a mathematical-economics point of view, ethi-

cally this could be considered reprehensible. Deep learning is derived from neural networks of such high complexity that they do not allow human beings to understand their logic for a decision, and all moral action depends on their justification — which would be obliterated by the algorithmic opacity resulting from the complexity. Thus, machines that are able to make moral decisions must be, in addition to being correct from the point of view of the rules programmed in them (mainly without discriminatory bias), transparent to the human.

There are those who claim that robots do not yet have self-awareness at the current stage of technological development, and that would make it impossible for them to develop an ethos themselves. But there is no reason to strongly believe that machines capable of moral decisions will not emerge in the (maybe near) future. If the problem lies in the possibility of justifying moral action, deep learning combined with the development in the processing capacity of machines, can make this objective possible, therefore. And if the problem lies in machines' lack of free will, it is necessary to remember that a philosophical and/or scientific consensus has not yet been reached about the existence of free will in human beings and whether such a condition is necessary for the formation of a moral judgment.

Robots may not be considered conscious or sentient yet, and it means that humans still do not perceive them as moral beings. But in the future, in the perception of people, they may approach the condition of close-to-damage agents, remaining those who could be responsible for their actions as distal agents. Thus, the possibilities for people, governments, and companies to attribute responsibility for damages perpetrated by robots — in war crimes, in abuse of power in actions by security forces, in medical errors, in traffic and transport accidents, etc. — can also increase, because people

tend to use the illegality exception arguments related to factors that are external to their conduct to justify their actions in case of damage.

Robots could acquire meaningful moral status if they become similar to other moral beings in performance. This does not depend on any metaphysical basis such as sentience, free will, or naturalness, but on their externally verifiable behavior, in practice, as being similar to that of a morally significant being. An important principle to be considered in the case of robots would be that of procreative beneficence, according to which whoever decides to create a robot (its manufacturer, for example), must provide it with the best possibility of existence, according to the technological state of the art at the time of its creation. The decision to create a robot is totally rational, not requiring a renouncement of other decisive assets of life besides the creation of the automaton, and although it confers high costs on its manufacturer, creating a robot is significantly less risky than genetically manipulating a human being, for example.

Although it is an interesting thesis, nevertheless, ethical behaviorism must be criticized precisely because it does not consider robots' mental states — which would be considered metaphysical. But in addition to facts considered metaphysical to increasingly puzzle humanity with regard to machines — such as their eventual sentience, conscience, intelligence, and self-determination — there are internal issues regarding the ontology of a machine (especially its design process) that are fundamental for giving them moral status. However, this does not mean that ethical behaviorism is simply incorrect, but rather incomplete as a foundation: verifiable external performances are important for verifying the moral status of a machine but analyzes that, at first sight, depend on certain "metaphysical" parameters must be added to their importance.

Machines have characteristics and mechanical abilities (such as precision, strength, speed) that are superior to those in humans, and such abilities must be used by humanity in the most varied situations. And for their actions, ethical priorities must be listed previously, so that their code may respect them. Still, neither men nor machines can make perfect moral decisions. What must be considered, in this sense, is that machines cannot make moral decisions, in complex contexts, that are worse than those that human beings would make in the same situation. And initial studies have shown that, for the construction of a moral algorithm in autonomous vehicles, for example, three trends are dominant among people's preferences: decisions that give preference to saving human lives (and not animals, plants, or even, properties); decisions that give preference to saving the maximum possible number of lives; and decisions that give preference to saving younger lives.

Furthermore, not only common sense expectations of people must be taken into account for developing machines' morality. Although Moral and Law are differentiated, independent communicative spheres, this does not mean that one cannot influence the other. The influence of Law on human moral intuitions is quite significant — so, instituting legal acts and regulations on the morality of machines can influence not only their decisions but also people's moral expectations about machines. Furthermore, studies have already been carried out with texts from the most varied eras and cultures, which has been used to train algorithms, and it was found that, when contextualized, such texts led such systems to deontologically correct choices.

Currently, moral autonomous agents can be viewed from various strategies and criteria — be them aware of it or not. They can draw on some previous normative ethical theo-

ry; learn, by trial and error, their own ethical behaviors from each situation; or combine both criteria in order to develop ethical behavior. But the development of a moral machine is even more complex than developing morality in humans, due to many contingencies particular to machines. They have only external motivations for moral behavior, unlike the humans who, although also have external motivations (incentives, punishments, and responsibility attribution), are mainly internally compelled to act morally. In addition, this external source of incentives is extremely contingent, since the interests of various types of people and organizations (users, developers, companies, etc.), which are not fully known until the moment of the system's implementation, must be taken into account for its development. And machines are very different from each other at the hardware and software levels — unlike humans, which are more or less similar organically and rationally. Therefore, for the development of functionally moral artificial agents that perform complex care tasks for human beings, hybrid design approaches, which combine the application of a predefined set of moral principles to deep learning, are the most indicated, as they recognize which is morally correct and act accordingly to it in each specific situation.

In addition to the moral programming of machines to be fully contextualized, one cannot ignore the incorporation of the explainability of their decision-making processes, considering that, however perfect a machine may seem, its reliability will depend on the possibility of giving explanations for its decisions.

The responsibility of any entity (human or not) depends on the degree of autonomy such an agent has: autonomous agents can suffer the punishment due to morals/ law, and non-autonomous agents do not suffer it directly, being gener-

ally responsible for them, in order to fill liability gaps in case of damage. When thinking about civil liability for damages perpetrated by automata, there are many strategies for doing so — including insurance contracts for the risks they offer —, but when it comes to criminal responsibility, it becomes more complicated to punish a designer, a programmer, or an investor due to the damage their robots eventually could commit. The natural tendency is, due to common sense, to morally punish the same entity whose action has a causal nexus with the damage — as that tendency can lead to serious consequences, the legal-philosophical debate must pay close attention to those gaps in responsibility, so that technological innovation takes a less unfair and more peaceful course.

Correctly defining what a robot is, however, may become a problematic and, perhaps, impossible task — which can cause watertight legal definitions to fall into disuse right away, or support less and less correct judgments. Therefore, an interesting legal-political strategy for the regulation of machines may be to leave to legislators the definition of what behaviors are considered incorrect for robots, and not what these entities are — a task that should be left to the powers of the Courts, which will analyze the particularities of each specific, concrete case, and of the regulators, which have more flexible procedures than the legislators, being able to cover, in their regulations, more immediate situations resulting from technological contingencies.

References

ARISTÓTELES. *Retórica.* Trad. Manuel Alexandre Júnior, Paulo Farmhouse e Abel do Nascimento Pena. Lisboa: Imprensa Nacional; Casa da Moeda, 2005.

AWAD, Edmond et al. Crowdsourcing moral machines. *Communications of the ACM*, v. 63, n. 3, p. 48-55, 2020. DOI: https://doi.org/10.1145/3339904.

AWAD, Edmond et al. The Moral Machine Experiment. *Nature*, v. 563, n. 7729, p. 59-64, 2018. DOI: https://doi.org/10.1038/s41586-018-0637-6.

BIGMAN, Yochanan E. et al. Holding Robots Responsible: The Elements of Machine Morality. *Trends in Cognitive Sciences*, v. 23, n. 5, p. 365-368, 2019. DOI: https://doi.org/10.1016/j.tics.2019.02.008.

CALO, Ryan. Robotics and the lessons of cyberlaw. *California Law Review*, v. 103, n. 3, p. 513–563, 2015. Available at: http://www.californialawreview.org/print/2robotics_cyberlaw/. Access in: 05 oct 2020.

CASEY, Bryan; LEMLEY, Mark A. You might be a robot. *Cornell Law Review*, v. 105, n. 2, p. 287-362, 2020. Available at: https://www.cornelllawreview.org/2020/01/12/you-might-be-a-robot/. Access in: 05 oct 2020.

CERVANTES, José-Antonio et al. Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics,* v. 26**,** p. 501–532, 2020. DOI: https://doi.org/10.1007/s11948-019-00151-x.

CHARISI, Vicky et al. Towards Moral Autonomous Systems. *Arxiv*, 2017. Available at: https://arxiv.org/abs/1703.04741. Access in: 05 oct 2020.

DANAHER, John. Robots, law and the retribution gap. *Ethics and Information Technology*, v. 18, p. 299-309, 2016. DOI: https://doi.org/10.1007/s10676-016-9403-3.

DANAHER, John. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics*, v. 26, p. 2023–2049, 2020. DOI: https://doi.org/10.1007/s11948-019-00119-x.

FLORIDI, Luciano et al. How to design AI for social good: seven essential factors. *Science and Engineering Ethics,* v. 26, p. 1771-1796, 2020. DOI: https://doi.org/10.1007/s11948-020-00213-5.

GORDON, John-Stewart. Building Moral Robots: Ethical Pitfalls and Challenges. *Science and Engineering Ethics,* v. 26, p. 141-157, 2020. DOI: https://doi.org/10.1007/s11948-019-00084-5.

HELLSTRÖM, Thomas. On the moral responsibility of military drones. *Ethics and Information Technology*, v. 15, p. 99-107, 2013. DOI https://doi.org/10.1007/s10676-012-9301-2.

HUANG, Bert I. Law's halo and the moral machine. *Columbia Law Review*, v. 119,, n. 7, p. 1811-1828, 2019. Available at: https://columbialawreview.org/content/laws-halo-and-the-moral-machine/. Access in: 05 oct 2020.

KLENK, Michael. How Do Technological Artefacts Embody Moral Values? *Philosophy & Technology*, p. 1-20, 2020. DOI: https://doi.org/10.1007/s13347-020-00401-y.

LEBEN, Derek. *Ethics for robots:* how to design a moral algorithm. London; New York: Routledge, 2019.

MANFIO, Edio Roberto. Robôs de conversação e o ethos. *Veritas*, v. 64, n. 2, p. 1-17, 2019. DOI: http://dx.doi.org/10.15448/1984-6746.2019.2.33174.

MCGRATH, James; GUPTA, Ankur. Writing a Moral Code: Algorithms for Ethical Reasoning by Humans and Machines. *Religions,* v. 9, n. 8, p. 240-259, 2018. DOI: https://doi.org/10.3390/rel9080240.

MISSELHORN, Catrin. Artificial systems with moral capacities? A research design and its implementation in a geriatric care system. *Artificial Intelligence*, v. 278, p. 1-11, 2020. DOI: https://doi.org/10.1016/j.artint.2019.103179.

MOOR, James H. The nature, importance, and difficulty of machine ethics. ***IEEE Intelligent Systems***, v. 21, n. 4, p. 18-21, 2006. DOI: https://doi.org/10.1109/mis.2006.80.

NALLUR, Vivek. Landscape of Machine Implemented Ethics. *Science and Engineering Ethics,* p. 1-20, 2020. DOI: https://doi.org/10.1007/s11948-020-00236-y.

REISS, Michael J. Robots as persons? Implications for moral education. *Journal of Moral Education*, p. 1-9, 2020. DOI: https://doi.org/10.1080/03057240.2020.1763933.

SAVULESCU, Julian. Procreative beneficence: why we should select the best children. *Bioethics***,** v. 15, n. 5-6, p. 413-426, 2001. DOI: https://doi.org/10.1111/1467-8519.00251.

SCHRAMOWSKI, Patrick et al. The Moral Choice Machine. *Frontiers in Artificial Intelligence*, v. 3, p. 1-15, 2020. DOI: https://doi.org/10.3389/frai.2020.00036.

SMIDS, Jilles. Danaher's Ethical Behaviourism: An Adequate Guide to Assessing the Moral Status of a Robot? *Science and Engineering Ethics*, p. 1-18, 2020. DOI: https://doi.org/10.1007/s11948-020-00230-4.

Recebido em: 02/06/2021

Aprovado em: 15/09/2022

Mateus de Oliveira Fornasier
E-mail: mateus.fornasier@gmail.com