INTERNET POLICY REVIEW
Journal on internet regulation

Volume 8 | Issue 3

New perspectives on ethics and the laws of artificial intelligence

Eduardo Magrani
FGV Law School; Ibmec; PUC-Rio, Rio de Janeiro, Brazil, eduardomagrani@gmail.com

Published on 13 Sep 2019 | DOI: 10.14763/2019.3.1420

Abstract: The continuous interaction between intelligent devices, sensors and people points to the increasing number of data being produced, stored and processed, changing, in various aspects and increasingly, our daily life. This increasing connectivity and symbiotic interaction among humans and intelligent machines brings significant challenges for the rule of law and contemporary ethics. Do machines have morality? What legal liability regime should we adopt for damages arising from increasingly advanced artificial intelligence? Which ethical guideline should we adopt to orient its advancement? In this paper we will discuss the main normative and ethical challenges imposed by the advancement of artificial intelligence.

Keywords: Artificial intelligence, Morality, Ethics

Article information

Received: 26 Sep 2018 **Reviewed:** 10 Dec 2018 **Published:** 13 Sep 2019
Licence: Creative Commons Attribution 3.0 Germany
Competing interests: The author has declared that no competing interests exist that have influenced the text.

URL: http://policyreview.info/articles/analysis/new-perspectives-ethics-and-laws-artificial-intelligence

Citation: Magrani, E. (2019). New perspectives on ethics and the laws of artificial intelligence. *Internet Policy Review*, *8*(3). https://doi.org/10.14763/2019.3.1420

INTRODUCTION

With the growing dissemination of ‘Big Data’ and computing techniques, technological evolution spread rapidly and increasingly intelligent algorithms have become a great resource for innovation and business models.

This new context based on the concepts of Web 3.0, internet of things and artificial intelligence, depends on the continuous interaction between intelligent devices, sensors and people generating a huge amount of data being produced, stored and processed, changing, in various aspects, our daily life (Magrani, 2017).

The increasing connectivity and symbiotic interaction among these agents,[1] bring a significant challenge for the rule of law and contemporary ethics, demanding a deep reflection on morality, governance and regulation.

What role should intelligent things play in our society? Do machines have morality? What legal liability regime should we adopt for damages arising from increasingly advanced artificial intelligence (AI)? Which ethical guidelines should we adopt to orient its development? In this paper we will discuss the main normative and ethical challenges imposed by the advancement of artificial intelligence.

TECHNOLOGY IS NOT NEUTRAL: AGENCY AND MORALITY OF THINGS

Peter-Paul Verbeek in his work *Moralizing Technology: Understanding and Designing the Morality of Things* aims to broaden the scope of ethics to better accommodate the technological age, and in doing so, reveals the inseparable nature of humanity and technology. Following Verbeek's contributions, technologies can be considered "moral mediators" that shape the way we perceive and interact with the world and thus reveal and guide possible behaviours. Since every technology affects the way in which we perceive and interact with the world, and even the way we think, no technology is morally neutral – it mediates our lives (Verbeek, 2011).

Technical artifacts, as explained by the theorist Peter Kroes, can be understood as man-made *Things (objects)*, which have a *function* and a *plan of use*. They consist of products obtained through technological action, designating the attitudes we take daily to solve practical problems, including those related to our desires and our needs. Technical artifacts involve the need for rules of use to be observed, as well as for parameters to be created in relation to the roles of individuals and social institutions in relation to them and their use (Vermaas, Kroes, van de Poel, Franssen, & Houkes, 2011).

Technical artifacts, as specific objects (Things) with their own characteristics have a clear function and usage plan. Besides, they are subject to an evaluation analysis as to whether they are good or bad and whether they work or not. Thus, it is possible to observe the great importance that the *function* and the *plan of use* have in the characterisation of a technical artifact. These two characteristics are intimately connected with the goals that the individuals who created the object seek with it, so that they do not stray from the intended purposes (Vermaas et al., 2011).

Faced with this inseparability, the questioning of the morality of human objectives and actions extends to the morality of technical artifacts (Vermaas et al., 2011). Technology can be used to change the world around us and individuals have goals – be they private and / or social – that can be achieved with the help of these technical artifacts and technologies. Considering that the objectives sought by the humans when creating a technical artifact are not separated from the characteristics of the object itself, we can conclude that the technical artifacts have an intrinsically moral character.

Therefore, alongside the technical artifacts, which can represent the simplest objects, with little capacity for interaction/influence, to the more technologically complex ones, we have the sociotechnical systems, which consist of a network that connects humans and *things*, thus possessing greater capacity for interaction and unpredictability (Latour, 2001).

For a regulatory analysis, this concept is even more fundamental (Vermaas et al., 2011).

Precisely because of its complexity embodied in a conglomerate of 'actants' (in relation to Bruno Latour's conception of actor-network theory), causing sociotechnical systems to have even less predictable consequences than those generated by technical artifacts. In addition, they generate a greater difficulty to prevent unintended consequences, and to hold agents liable in case of harm, since the technological action, reflected in the sociotechnical system, is a sum of actants' actions, entangled in the network in an intra-relation (Barad, 2003).

TECHNICAL ARTIFACTS AND SOCIOTECHNICAL SYSTEMS: ENTANGLED IN INTRA-RELATION

To illustrate the difference between the concepts of technical artifact and sociotechnical system, we can think of the former being represented by an airplane, and the second by the complex aviation system. The sociotechnical system is formed by the set of interrelated agents (human and non-human actants - *things*, institutions, etc.) that work together to achieve a given goal. The materiality and effects of a sociotechnical system depend on the sum of the agency of each actant. However, there are parameters of how the system should be used, which means that these systems have pre-defined operational processes and can be affected by regulatory laws and policies.

Thus, when a tragic accident involving an airplane occurs, it is necessary to analyse what was in the sphere of control and influence of each actor and technical artifact components of the sociotechnical network. Quite possibly we will observe a very complex and symbiotic relationship between the components that led to this fateful result (Saraiva, 2011). Moreover, this result is often unpredictable, due to the autonomy of the system based on a diffused and distributed agency among all components (actants).

These complex systems bring us to debate the liability and ethics concerning technical artifacts and sociotechnical systems. Issues such as the liability of developers and the existence of morality in non-human agents - with a focus here on technological objects - need a response or, at least, reflections that contribute to the debate in the public sphere. [2]

Bruno Latour's theory offers progress in confronting and discarding the formal binary division between humans and non-humans, but it places objects with different complexities and values at the same level. Given this context, from a legal and regulatory point of view, assigning a different status to technical artifacts and sociotechnical systems, according to their capacity for agency and influence is justifiable and should be endowed with different moral status and levels of liability. It is necessary, then, to distinguish the influence and importance that each *thing* also has in the network and, above all, in the public sphere (Latour, 2001).

HELLO WORLD: CREATING UNPREDICTABLE MACHINES

For this analysis, we will focus on specific things and technologies, aiming at advanced algorithms with *machine learning* or robots equipped with artificial intelligence (AI), considering that they are technical artifacts (Things) attached to sociotechnical systems with a greater potential for autonomy (based largely on the processing of 'Big Data') and unpredictability.

While technical artifacts, such as a chair or a glass, are artifacts "domesticated" by humans, i.e., more predictable in terms of their influence and agency power, it is possible to affirm that intelligent algorithms and robots are still non-domesticated technologies, since the time of interaction with humans throughout history has not yet allowed us to foresee most of the risks in order to control them, or to cease them altogether.

Colin Allen and Wendell Wallach (Wallach and Allen, 2008) argue that as intelligent Things - like robots [3] - become more autonomous and assume more responsibility, they must be programmed with moral decision-making skills for our own safety.

Corroborating this thesis, Peter-Paul Verbeek, while dealing with the morality of Things understands that: as machines now operate more frequently in open social environments, such as connected public spheres, it becomes increasingly important to design a type of functional morality that is sensitive to ethically relevant characteristics and applicable to intended situations (Verbeek, 2011).

A good example is Microsoft's robot Tay, which helps to illustrate the effects that a non-human element can have on society. In 2016, Microsoft launched an artificial intelligence programme named Tay. Endowed with a *deep learning* [4] ability, the robot shaped its worldview based on online interactions with other people and producing authentic expressions based on them. The experience, however, proved to be disastrous and the company had to deactivate the tool in less than 24 hours due to the production of worrying results.

The goal was to get Tay to interact with human users on Twitter, learning human patterns of conversation. It turns out that in less than a day, the *chatbot* was generating utterly inappropriate comments, including racist, sexist and antisemitic publications.

In 2015, a similar case occurred with "Google Photos". This was a programme that also learned from users to tag photos automatically. However, their results were also outright discriminatory, and it was noticed, for example, that the bot was labeling coloured people as gorillas.

The implementation of programmes capable of learning and adapting to perform functions that relate to people creates new ethical and regulatory challenges, since it increases the possibility of obtaining results other than those intended, or even totally unexpected ones. In addition, these results can cause harm to other actors, such as the discriminatory offenses generated by Tay and Google Photos.

Particularly, the use of artificial intelligence tools that interact through social media requires reflection on the ethical requirements that must accompany the development of this type of technology. This is because, as previously argued, these mechanisms also act as agents in society, and end up influencing the environment around them, even though they are non-human elements. It is not, therefore, a matter of thinking only about the "use" and "repair" of new technologies, but mainly about the proper ethical orientation for their development (Miller, Wolf, & Grodzinsky, 2017).

Microsoft argued that Tay's malfunctioning was the result of an attack by users who exploited a vulnerability in their programme. However, for Miller et al. this does not exempt them from the responsibility of considering the occurrence of possible harmful consequences with the use of this type of software. For the authors, the fact that the creators did not expect this outcome is part of the very unpredictable nature of this type of system (Miller et al., 2017).

The attempt to make artificial intelligence systems increasingly adaptable and capable of acting in a human-like manner, makes them present less predictable behaviours. Thus, they begin to act not only as tools that perform pre-established functions in the various fields in which they are employed, but also to develop a proper way of acting. They impact the world in a way that is less determinable or controllable by human agents. It is worth emphasising that algorithms can adjust to give rise to new algorithms and new ways to accomplish their tasks (Domingos, 2015), so that the way the result was achieved would be difficult to explain even to the programmers who created the algorithm (Doneda & Almeida, 2016).

Also, the more adaptable the artificial intelligence programmes become, the more unpredictable are their actions, bringing new risks. This makes it necessary for developers of this type of programme to be more aware of the ethical and legal responsibilities involved in this activity.

The Code of Ethics of the Association for Computing Machinery(Miller et al., 2017) indicates that professionals in the field, regardless of prior legal regulation, should develop “comprehensive and thorough assessments of computer systems and their impacts, including the analysis of possible risks”.

In addition, there is a need for dedicated monitoring to verify the actions taken by such a programme, especially in the early stages of implementation. In the Tay case, for instance, developers should have monitored the behaviour of the bot intensely within the first 24 hours of its launch, which is not known to have occurred(Miller et al., 2017). The logic should be to prevent possible damages and to monitor in advance, rather than the remediation of losses, especially when they may be unforeseeable.

To limit the possibilities of negative consequences, software developers must recognise those potentially dangerous and unpredictable programmes and restrict their possibilities of interaction with the public until it is intensively tested in a controlled environment. After this stage, consumers should be informed about the vulnerabilities of a programme that is essentially unpredictable, and the possible consequences of unexpected behaviour (Miller et al., 2017).

The use of technology, with an emphasis on artificial intelligence, can cause unpredictable and uncontrollable consequences, so that often the only solution is to deactivate the system. Therefore, the increase in autonomy and complexity of the technical artifacts is evident. They are endowed with an increased agency, and are capable of influencing others but also of being influenced in the sociotechnical system in a significant way, often composing even more autonomous and unpredictable networks.

Although there is no artificial intelligence system yet that is completely autonomous, with the pace of technological development, it is possible to create machines that will have the ability to make decisions in an increasingly autonomous way, which raises questions about who would be responsible for the result of its actions and for eventual damages caused to others (Vladeck, 2014).

APPLICATION OF NORMS: MAPPING LEGAL POSSIBILITIES

The ability to amass experiences and learn from massive data processing, coupled with the ability to act independently and make choices autonomously can be considered preconditions for legal liability. However, since artificial intelligence is not recognised today as a subject of law, it cannot be held individually liable for the potential damage it may cause.

In this sense, according to Article 12 of the United Nations Convention on the Use of Electronic Communications in International Contracts, a person (natural or an entity) on behalf of whom a programme was created must, ultimately, be liable for any action generated by the machine. This reasoning is based on the notion that a tool has no will of its own (Čerka et al., 2015).

On the other hand, in the case of damage caused by acts of an artifact with artificial intelligence, another type of responsibility is the one that makes an analogy with the responsibility attributed to the parents by the actions of their children or even the responsibility of animal owners in case of damage. In this perspective, the responsibility for the acts of this artifact could fall not only on its producer or programmers, but also on the users that were responsible for their "training" (Čerka et al., 2015).

Another possibility is the model that focuses on the ability of programmers or users to predict the potential for these damages to occur. According to this model, the programmer or user can be held liable if they acted deceitfully or had been negligent considering a result that would be predictable (Hallevy, 2010).

George S. Cole refers to predetermined types regarding civil liability: (i) product liability, (ii) service liability, (iii) malpractice, and (iv) negligence. The basic elements for applicability of product liability would be: (i) the AI □□should be a "product"; (ii) the defendant must be an AI seller; (iii) the AI must reach the injured party without substantive change; (iv) the AI □□must be defective; and (v) the defect shall be the source of the damage. The author sustains that the standards, in this case, should be set by the professional community. Still, as the field develops, for Cole, the negligence model would be the most applicable. However, it can be difficult to implement, especially when some errors are unpredictable or even unavoidable (Cole, 1990).

To date, the courts worldwide have not formulated a clear definition of the responsibility involved in creating AIs which, if not undertaken, should lead to negligent liability. This model will depend on standards set by the professional community, but also clearer guidelines from the law side and jurisprudence.

The distinction between the use of negligence rule and strict liability rule may have different impacts on the treatment of the subject and especially on the level of precaution that is intended to be imposed in relation to the victim, or in relation to the one who develops the AI

In establishing strict liability, a significant incentive is created for the offender to act diligently in order to reduce the costs of anticipating harm. In fact, in the economic model of strict responsibility, the offender responds even if he adopts a high level of precaution. This does not mean that there is no interest in adopting cautious behaviour. There is a level of precaution in which the offender, in the scope of strict liability will remove the occurrence of damage. In this sense, if the adoption of the precautionary level is lower than the expected cost of damages, from an economic point of view, it is desirable to adopt the precautionary level (Shavell, 2004). But even if the offender adopts a diligent behaviour, if the victim suffers damage, she will be reimbursed, which favours, in this case, the position of the victim (Magrani, Viola, and Silva, 2019).

The negligence rule, however, forms a completely different picture. As the offender responds only when he acts guilty, if he takes diligent behaviour, the burden of injury will necessarily fall on the victim, even if the damage is produced by reason of a potentially dangerous activity. Therefore, the incentive for victims to adopt precautionary levels is greater, because if they suffer any kind of loss, they will bear it (Magrani, Viola, and Silva, 2019).

Should an act of an artificial intelligence cause damages by reason of deceit or negligence, manufacturing defect or design failure as a result of blameworthy programming, existing liability rules would most often indicate the “fault” of its creators. However, it is often not easy to know how these programmes come to their conclusion or even lead to unexpected and possibly unpleasant consequences. This harmful potential is especially dangerous in the use of artificial intelligence programmes that rely on *machine learning* and especially *deep learning* mechanisms, in which the very nature of the *software* involves the intention of developing an action that is not predictable, and which will only be determined from the data processing of all the information with which the programme had contact. Existing laws are not adequate to guarantee a fair regulation for the upcoming artificial intelligence context.

The structure contained in the table below, produced in a UNESCO study (UNESCO, 2017), contains important parameters that help us think about these issues, at the same time trying to identify the different agencies involved.

Table 1. From UNESCO (2017)

Decision by robot	**Human involvement**	**Technology**	**Responsibility**	**Regulation**
Made out of finite set of options, according to preset strict **criteria**	Criteria implemented in a legal framework	Machine only: deterministic algorithms/robots	Robot’s producer	Legal (standards, national or international legislation)
Out of a range of options, with room for flexibility, according to a preset **policy**	Decision delegated to robot	Machine only: AI -based algorithms, cognitive robots	Designer, manufacturer, seller, user	Codes of practice both for engineers and for users; precautionary principle
Decisions made through human-machine **interaction**	Human controls robot’s decisions	Ability for human to take control over robot in cases where robot’s actions can cause serious harm of death	Human beings	Moral

Although the proposed structure is quite simple and gives us important insights, its implementation in terms of assigning responsibility and regulating usage is complex and challenging for scientists and engineers, policymakers and ethicists, and eventually it will not be sufficient for applying a fair and adequate response.

HOW TO DEAL WITH AUTONOMOUS ROBOTS: INSUFFICIENT NORMS AND THE PROBLEM OF ‘DISTRIBUTED IRRESPONSIBILITY’

Scientists from different areas are concerned and deliberate that conferring this autonomous "thinking" ability to machines will necessarily give them the ability to act contrary to the rules they are given (Pagallo, 2013). Hence the importance of taking into consideration and investigating the spheres of control and influence of designers and other agents during the creation and functional development of technical artifacts (Vladeck, 2014). [5]

Often, during the design phase, the consequences are indeterminate because they depend partly on the actions of other agents and factors beside those of the designers. Also, since making a decision can be a complex process, it may be difficult for a human to even explain it. It may be difficult, further, to prove that the product containing the AI was defective, and especially that the defect already existed at the time of its production (Čerka et al., 2015).

As the behaviour of an advanced AI is not totally predictable, and its behaviour is the result of the interaction between several human and non-human agents that make up the sociotechnical system and even of *self-learning* processes, it can be difficult to determine the *causal nexus* [6] between the damage caused and the action of a human being or legal entity. [7]

According to the legal framework we have today, this can lead to a situation of "distributed irresponsibility" (the name attributed in the present work to refer to the possible effect resulting from the lack of identification of the *causal nexus* between the agent's conduct and the damage caused) among the different actors involved in the process. This will occur mainly when the damage transpires within a complex sociotechnical system, in which the liability of the intelligent *thing* itself, or of a natural or legal person, will not be obvious. [8]

'WITH A LITTLE HELP FROM MY FRIENDS': DESIGNING ETHICAL FRAMEWORKS TO GUIDE THE LAWS OF AI

When dealing with artificial intelligence, it is essential for the research community and academia to promote an extensive debate about the ethical guidelines that should guide the construction of these intelligent machines.

There is a strong growth of this segment of scientific research. The need to establish a *regulatory framework* for this type of technology has been highlighted by some initiatives as mentioned in this section.

The EU Commission published in April 2019 the document "Ethics guidelines for trustworthy AI" with guidelines on ethics in artificial intelligence. According to the guidelines, trustworthy AI should be: "(i) lawful - respecting all applicable laws and regulations; (ii) ethical - respecting ethical principles and values; and (iii) robust - from a technical perspective" (HLEG AI, 2019).

The guidelines put forward a set of seven key requirements that AI systems should meet in order to be deemed trustworthy. According to the document, a specific assessment list (hereunder) aims to help verify the application of each of the key requirements:

> - Human agency and oversight: AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches;
> - Technical robustness and safety: AI systems need to be resilient and secure. They

> need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented;
> - Privacy and data governance: besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimized access to data;
> - Transparency: the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations;
> - Diversity, non-discrimination and fairness [9]: unfair (algorithmic) bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle;
> - Societal and environmental well-being: AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered;
> - Accountability: mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured.
>
> (HLEG AI, 2019)

Similar to this well-grounded initiative, many countries, companies and professional communities are publishing guidelines for AI, with analogous values and principles, intending to ensure the positive aspects and diminish the risks involved in AI development. In that sense, it is worth mentioning the recent and important initiatives coming from:

i. Future of Life Institute – Asilomar AI;
ii. Berkman Klein Center;
iii. Institute Electrical and Electronic Engineers IEEE;
iv. Centre for the study on existential risks;
v. K&L gates endowment for ethics;
vi. Center for human-compatible AI;
vii. Machine Intelligence Research Institute;
viii. USC center for AI in society;
ix. Leverhulme center for future of intelligence;
x. Partnership on AI;
xi. Future of Humanity Institute;
xii. AI Austin;
xiii. Open AI;
xiv. Foundation for Responsible Robotics;
xv. Data & Society (New York, US);
xvi. World Economic Forum's Council on the Future of AI and Robotics;
xvii. AI Now Initiative;
xviii. AI100.

Besides the great advancements on ethical guidelines designed by the initiatives hereinabove, containing analogous values and principles, one of the most complex discussions that pervades the various guidelines that are being elaborated is related to the question of AI's autonomy.

The different degrees of autonomy allotted to the machines must be thought of, determining what degree of autonomy is reasonable and where substantial human control should be maintained. The different levels of intelligence and autonomy that certain technical artifacts may have must directly influence the ethical and legal considerations about them.

ROBOT RIGHTS: AUTONOMY AND E-PERSONHOOD

On 16 February 2017, the European Parliament issued a resolution with recommendations from the European Commission on *civil law* rules in robotics. The document the European Parliament issued ("Recommendations from the European Commission on civil law rules in robotics 2015/2103 – INL") advocates for the creation of an European agency for robotics and artificial intelligence, to provide the necessary technical, ethical and regulatory expertise. The European Parliament also proposed the introduction of a specific legal status for smart robots as well as the creation of an insurance system and compensatory fund [10] with the aim of creating a protection system for the use of intelligent machines.

Regarding the legal status that could be given to these agents, the resolution uses the expression "electronic person" or "*e-person*". In addition, in view of the discrepancy between ethics and technology, the European proposition rightly states that dignity, in a deontological bias, must be at the centre of a new digital ethics.

The attribution of a legal status to intelligent robots, as designed in the resolution, it is intended to be one possible solution to the legal challenges that will arise with the gain of autonomy of intelligent Things. The European Parliament's report defines "intelligent robots" as those whose autonomy is established by their interconnectivity with the environment and their ability to modify their actions according to changes.

With the purpose of building up on this discussion, the Israeli researcher Karni Chagal-Feferkorn performs the analysis on robot autonomy to help us differentiate the potential of responsibility in each case. To Chagal-Feferkorn, in order to resolve the liability issue, it is crucial to think on different levels of robot's autonomy (Chagal-Feferkorn, 2018). Nevertheless, she is aware that given the complexity of the artificial intelligence systems, the classification is difficult to implement, since the autonomy is not a binary classification.

Two possible metrics raised for assessing autonomy are freedom of action of the machine with respect to the human being and the capacity of the machine to replace human action. Such metrics are branched and complex with several possible sub-analyses and, according to Chagal-Feferkorn, these tests should also consider the specific stage of the machine decision-making process (Chagal-Feferkorn, 2019).

To illustrate, Chagal-Feferkorn designed the following table (hereunder), with a metric showing the possibility for machines to substitute humans in complex tasks and analysing also the decision making capacity of the machine (Chagal-Feferkorn, 2019). The more machines get closer to a "robot-doctor" stage, the more reasonable it would be to attribute new forms of accountability, liability, rights or even an electronic personhood.

Table 2. From Chagal-Feferkorn (2019).

	Roomba robot	Autopilot	Autonomous vehicle	Robo-doctor
Success rates not measurable?				
Responsible for more than two OODA loop stages?			+	+
Independently selects type of info to collect?			?	+
Independently selects sources of info to collect from?				+
Dynamic nature of sources of info?				+
Replaces professionals in complex fields?		?	?	+
Life and death nature of decisions?		+	+	+
Real time decisions required?		+	+	?

One criteria used by Chagal-Feferkorn is the OODA [observe-orient-decide-act] cycle. [11] Since the analysis of autonomy is complex, Chagal-Feferkorn states that we should observe the characteristics of different decision-making systems. These systems manifest themselves in four different stages, according to the OODA cycle, affecting different justifications for liability concerning machines. These four points are: (i) Observe: collect current information from all available sources; (ii) Orient: analyse the information collected and use it to update its reality; (iii) Decide: decide the course of action; (iv) Act: implement its decision.

Considering the stages of the OODA cycle used by Chagal-Feferkorn, the more the characteristics of the system are analogous to traditional products / things, the greater the possibility of being embedded in the logic of consumer law. However, advanced robots and algorithms, because of its specific characteristics, might be classified differently from traditional consumer products and, therefore, needing a differentiated treatment and responsibility perspective.

The parameters for assigning responsibility in accordance with consumer law are defined and precise. However, as the complexity of systems increases, in the case of 'doctor robots', for instance, as a specific example brought in the study, the number of scenarios and justification for assigning responsibility depend on a number of factors. The doctor robots' example correspond to the last stage of autonomy thought of by Chagal-Feferkorn, in which algorithms of reasoning are programmed to be capable of replacing human beings in highly complex activities, like medical activities of diagnosis and surgery.

In order for the degree of autonomy-based responsibility to be measured, one should consider the size of the parameter matrix that the algorithm judges before the final decision-making and how much of that decision was decisive for the damaging outcome. It is necessary to consider that the more stages of OODA a system is able to operate, the greater the unpredictability of the manufacturer on the decisions taken by artificial intelligence (Magrani, Viola, & Silva, 2019). [12]

In the case of the robot doctor, for instance, it is up to the machine to decide to what extent it should consider the medical history of the patient and the more independent of human action these decisions are, the further the human responsibility will be. On the contrary, it would be possible to programme the machine in such a way as to consult a human being whenever the percentage of certainty for a decision-making is below a certain level, but the establishment of such issues would also imply an increase in the responsibility of the manufacturer (that should also be based on a deontological matrix type). The limit of action of the machine will be determinant for the attribution of responsibility (Magrani, Viola, & Silva, 2019).

Although our technology has not yet developed robots with sufficient autonomy to completely replace human beings for complex tasks, such as the case of doctor robots, if this moment arrives, we should have theoretical mechanisms to implement this type of attribution of responsibility without provoking chilling effects on technological innovations.

For the time being, and according to the consumerist logic, the responsibility should be attributed to the manufacturer. Nevertheless, considering the possibility of robots reaching more independence with respect to humans, fulfilling the four stages of OODA, the aforementioned logic of accountability of the consumer chain may not be applicable. This would trigger the need to assign rights and eventually even a specific personality to smart robots with high autonomy level, besides the possibility of creating insurance and funds for accidents and damages involving robots.

Because we are not yet close to a context of substantial or full robotic autonomy, such as a ‘strong AI’ or ‘general artificial intelligence’, there is a sizeable movement against the attribution of a legal status to them. Recently, over 150 experts in AI, robotics, commerce, law, and ethics from 14 countries have signed an open letter denouncing the European Parliament’s proposal to grant personhood status to intelligent machines. [13] The open letter suggests that current robots do not have moral standing and should not be considered capable of having rights.

However, as computational intelligence can grow exponentially, we should deeply consider the possibility of robots gaining a substantial autonomy on the next years, stressing the need for the attribution of rights.

Considering the myriad of possibilities, the Italian professor and researcher Ugo Pagallo states:

> Policy makers shall seriously mull over the possibility of establishing novel forms of accountability and liability for the activities of AI robots in contracts and business law, e.g., new forms of legal agenthood in cases of complex distributed responsibility. Second, any hypothesis of granting AI robots full legal personhood has to be discarded in the foreseeable future. (...) However, the normative reasons why legal systems grant human and artificial entities, such as corporations, their status, help us taking sides in today’s quest for the legal personhood of AI robots. (Pagallo, 2018)

One of the important features to consider is the learning speed and individual evolution of the robot (based on data processing and deep learning), which may represent in some cases the infeasibility of an educational process, thus limiting its moral and legal liability. But how could one punish a robot? It cannot be as simple as “pulling the plug”. In this case, there are two viable options: rehabilitation and indemnification. The first would involve reprogramming the guilty robot. The second, would be to compel the same to compensate the victim for the damage caused. In such a context, the European resolution is relevant. The proposition in assigning a new type of personhood, an electronic one, considering the characteristics of intelligent Things, coupled with the idea of □□compulsory insurance or a compensatory fund can be an important step.

The new European proposal reflects, therefore, a practical and prompt response to the previously mentioned problem of “distributed irresponsibility”, which occurs when there is no clear connection between an agent and the harm generated (unclear causal nexus between agents and damages).

In view of a causal nexus that cannot be identified directly, for some scholars, we can infer its presumption from the economic group, making it possible to repair the damages caused by facilitating the burden of proof for the victim. However, when we think of the damages that can occur within complex sociotechnical systems, we can have an unfair or unassured application of the causal nexus and legal liability. This is because we are often talking about the action caused by a sum of agencies of human beings, institutions and intelligent things with autonomy and agency power of their own. In this case, the focus on the economic group, despite being able to respond to several cases of damages, may not be sufficient for the fair allocation of liability in the artificial intelligence and internet of things era.

Therefore, as a pragmatic response to this scenario of uncertainty and lack of legal appropriateness, the European proposal suggests that in case of damages the injured party may either take out the insurance or be reimbursed through the compensatory fund linked to the intelligent robot itself.

Beside the concern that this legal arrangement could lead to a convenient tool for companies and producers to disproportionately set aside their responsibility before users and consumers, this step should be closely followed by a continuous debate on the ethical principles that should guide such technical artifacts. Furthermore, this discussion must be coupled with an adequate governance of all the data used by these agents. In observance of these factors, the recommendation is that the development of these intelligent artifacts be fully oriented by the previously described values, such as: (i) fairness; (ii) reliability; (iii) security (iv) privacy and data protection; (v) inclusiveness; (vi) transparency; and (vii) accountability.

GOVERNING INTRA-ACTION WITH HUMAN RIGHTS AND BY DESIGN

One point worth considering in this context is that flaws are natural and that they can be considered even desirable for the faster improvement of a technical artifact. Therefore, a regulatory scenario that would extinguish all and any flaws or damages would be uncalled for. AI-inspired robots are products with inherently unforeseeable risks. “The idea of avant-garde machine learning research is for robots to acquire, learn, and even discover new ways of interactions without the designer’s explicit instruction. The idea of artificial general intelligence (which is admittedly looking far into the future) is to do so even without any implicit instruction” (Yi Tan, 2018). Therefore, we could say that those technologies are “unforeseeable by design”.

From a legal standpoint, it is fundamental to keep in mind the new nature of a diffused liability, potentially dispersed in space, time and agency of the various actants in the public sphere. In that sense, we need to think about the context in which assumptions on liability are made. The question that is presented to us is not only how to make computational agents liable, but how to reasonably and fairly apply this liability.

The idea of a shared liability between the different agents involved in the sociotechnical network seems a reasonable perspective, requiring, in order to attribute a fair liability to each one, the analysis of their spheres of control and influence over the presented situations and over other agents (humans and non-humans), considering their intra-relation (intra-action) (Barad, 2003).

However, we are still far from obtaining a reasonable consensus [14] on the establishment of appropriate legal parameters for the development and regulation of intelligent Things, although we already see many advancements concerning ethical guidelines.

These agents can influence relationships between people, shaping behaviours and world views, especially and more effectively when part of their operation have technological complexity and different levels of autonomy, as it happens in the case of artificial intelligence systems with the capacity of reasoning and learning according to *deep learning* techniques in artificial neural networks (Amaral, 2015).

In view of the increasing risks posed by the advance of techno-regulation, amplified by the dissemination of the 'Internet of Things' and artificial intelligence, the rule of law should be seen as the premise for technological development, or as a meta-technology, which should guide the way technology shapes behaviour rather than the other way around - which often results in violation of human and fundamental rights.

For law to act properly as a meta-technology, it must be backed by ethical guidelines consistent with the age of hyperconnectivity. In this sense, it is necessary to understand the capacity of influence of non-human agents, aiming to achieve a better regulation, especially for more autonomous technologies, thinking about preserving the fundamental rights of individuals and preserving the human species.

The law, backed by an adequate ethical foundation, will serve as a channel for data processing and other technological materialities avoiding a techno-regulation harmful to humanity. In this new role, it is important that the law guides the production and development of Things (technical artifacts) in order to be sensitive to values, for example, regulating privacy, security and ethics by design. In a metaphor, law as meta-technology would function as a pipeline suited to the digital age, through which all content and actions would pass.

With technology moving from a simple tool to an influencing agent and decision maker, law must rebuild itself in the techno-regulated world, incorporating these new elements from a meta-perspective (as a meta-technology), building the normative basis to regulate the ethics of new technologies through design. To do so, we must enhance and foster human-centred design models that are sensitive to constitutional values □□(*value-sensitive design*).

Governing AI with the mentioned ethical principles (fairness; reliability; security; privacy; data protection; inclusiveness; transparency; and accountability) and the "by design" technique, are an important step to try to follow the pace of technological innovation, at the same time as trying to guarantee effectiveness of the law.

CONCLUSION

It is evident that these intelligent artifacts are consistently exerting more influence in the way we think and organise ourselves in society and, therefore, the scientific and legal advance cannot distance itself from the ethical and legal issues involved in this new scenario.

In that sense, new ontological and epistemological lenses are needed. We need to think about intelligent Things not as mere tools but as moral machines that interact with citizens in the public sphere, endowed with intra-acting agencies, entangled in sociotechnical systems.

Legal regulation, democratically construed in the public sphere, should provide the architecture for the construction of proper legal channels so that non-human agents can act and be developed within the prescribed ethical limits. To design adequate limits for the AI era, we must recognise Things as agents, based on a post-humanist perspective, but with a human rights based approach to guide its development.

Certainly, the reasons to justify an *electronic personhood* are not there yet. Nevertheless, since computational intelligence can grow exponentially, as well as their level of interaction on our daily lives and on the connected public sphere, with the gain of new stages of autonomy, we must inevitably think about the possibilities of establishing new forms of accountability and liability for the activities of AI, including the possibility of attributing rights, subjectivity and even an *e-personhood* in the future.

The granting of an electronic personality is the path suggested by the European Parliament for smart robots and we cannot reject this recommendation, as a future regulation, depending on the degree of autonomy conferred on AIs. Such construction, however, is not immune to criticism, notably as regards the comparison between an AI and a natural person. [15]

As evidenced, the discussion about ethics and responsibility of artificial intelligence still navigates murky waters. However, the difficulties arising from technological transformations of high complexity cannot prevent the establishment of new regulation that has the capacity to reduce the risks inherent in new activities and, consequently, the production and repair of damages (Magrani, Viola, & Silva, 2019). The exact path to be taken still remains uncertain. Nevertheless, it is already possible to envision possibilities that can serve as important parameters. In the wise words of the Italian philosopher Luciano Floridi: “The new challenge is not technological innovation, but the governance of the digital”.

ACKNOWLEDGEMENTS:

This article counted on the collaboration of Beatriz Laus in the translation.

REFERENCES

Amaral, G. (2014). *Uma dose de pragmatismo para as epistemologias contemporâneas: Latour e o parlamento das coisas* [A dose of pragmatism for contemporary epistemologies: Latour and the parliament of things]. São Paulo: Digital de Tecnologias Cognitivas.

Barad, K. (2003). Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. *Signs: Journal of Women in Culture and Society*, *28*(3). doi:10.1086/345321

Castro, M. (2009). *Direito e Pós-humanidade: quando os robôs serão sujeitos de direitos* [Law and post-humanity: when robots will be subject to rights]. Curitiba: Juruá.

Čerka, P., Grigienė, J., & Sirbikytė, G. (2015). Liability for damages caused by artificial intelligence. *Computer Law & Security Review*, *31*(3), 376–389. https://doi.org/10.1016/j.clsr.2015.03.008

Chagal-Feferkorn, K. A. (2019). Am I an Algorithm or a Product? When Products Liability Should Apply to Algorithmic Decision-Makers. *Stanford Law & Policy Review*, *30*, 61–114. Retrieved from https://www-cdn.law.stanford.edu/wp-content/uploads/2019/05/30.1_2-Chagal-Feferkorn_Final-61-114.pdf

Cole, G. S. (1990). Tort Liability for Artificial Intelligence And Expert Systems. *Computer/Law Journal*, *10*(2), 127–231. Retrieved from https://repository.jmls.edu/jitpl/vol10/iss2/1/

Domingos, P. (2015). *The Master Algorithm: how the quest for the ultimate learning machine will remake our world*. New York: Basic Books.

Doneda, D., & Almeida, V. A. F. (2016). What Is Algorithm Governance? *IEEE Internet Computing*, *20*(4), 60–63. doi:10.1109/MIC.2016.79

Hallevy, G. (2010). The Criminal Liability of Artificial Intelligence Entities: From Science Fiction to Legal Social Control. *Akron Intellectual Property Journal*, *4*(2), 171–201. Retrieved from https://ideaexchange.uakron.edu/akronintellectualproperty/vol4/iss2/1/

High-Level Expert Group on AI (HLEG AI). (2019). *Ethics guidelines for trustworthy AI* [Report / Study]. Brussels: European Commission. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Vermaas, P., Kroes, P., van de Poel, I., Franssen, M., & Houkes, W. (2011). A Philosophy of Technology: From Technical Artefacts to Sociotechnical Systems. *Synthesis Lectures on Engineers, Technology, and Society*, *6*(1). doi: 10.2200/S00321ED1V01Y201012ETS014

Latour, B. (2001). *A Esperança de Pandora: Ensaios sobre a realidade dos estudos científicos* [Pandora's Hope: Essays on the reality of scientific studies]. São Paulo: EDUSC.

Magrani, E. (2017). Threats of the Internet of Things in a techo regulated society A New Legal Challenge of the Information Revolution. *ORBIT Journal*, *1*(1). doi:10.29297/orbit.v1i1.17

Magrani, E., Silva, P., Viola, R. (2019). *Novas perspectivas sobre ética e responsabilidade de inteligência artificial* [New perspectives on ethics and responsibility of artificial intelligence]. In C. Mulholland, & A. Frazao (Eds.), *Inteligência Artificial e Direito: Ética, Regulação e Responsabilidade* [Artificial Intelligence and Law: Ethics, Regulation and Responsibility]. São Paulo: RT.

Matias, J. (2010). Da cláusula pacta sunt servanda à função social do contrato: o contrato no Brasil [From the pacta sunt servanda clause to the social function of the contract: the contract in Brazil]. In E. Vera-Cruz (Ed.), *O sistema contratual romano: de Roma ao direito actual* [The Roman contractual system: from Rome to current law]. Coimbra; Lisboa: Coimbra Editora; Faculdade de direito da universidade de Lisboa.

Miller, K. W., Wolf, M. J., & Grodzinsky, F. S. (2017). Why We Should Have Seen That Coming: Comments on Microsoft's tay "Experiment," and Wider Implications. *ORBIT Journal*, *1*(2). doi:10.29297/orbit.v1i2.49

Pagallo, U. (2013). *The Law of Robots: Crimes, Contracts and Torts*. Dordrecht: Springer. doi:10.1007/978-94-007-6564-1

Pagallo, U. (2018). Vital, Sophia, and Co.—The Quest for the Legal Personhood of Robots. *Information*, *9*(9). doi:10.3390/info9090230

Saurwein, F., Just, N., Latzer, M. (2015). Governance of algorithms: options and limitations. *info*, *17*(6), 35–49. doi:10.1108/info-05-2015-0025

Shavell, S. (2004). *Foundations of economic analysis of law*. Cambridge, MA: Belknap Press of Harvard University Press.

UNESCO. (2017). *Report of COMEST on robotics ethics*. Paris: World Commission on the Ethics of Scientific Knowledge and Technology. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000253952

Verbeek, P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago; London: The University of Chicago Press.

Vladeck, D. C. (2014). Machines without principals: liability rules and artificial intelligence. *Washington Law Review*, *89*(1), 117–150. Retrieved from https://digitalcommons.law.uw.edu/wlr/vol89/iss1/6/

Wallach, W. & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Yi Tan, C. (2018, December 11). Artificial Intelligence, Artificial Persons, and the Law. *Becoming Human: Artificial Intelligence Magazine*. Retrieved from https://becominghuman.ai/artificial-intelligence-artificial-persons-and-the-law-2cce322743b6

FOOTNOTES

[1] Better understood by the expression "actant" in Latour's theory.

[2] In its Habermasean definition.

[3] The 2005 UN Robotics Report defines a robot as a semi or fully autonomous reprogrammable machine used for the well-being of human beings in manufacturing operations or services.

[4] "Deep learning is a subset of machine learning in which the tasks are broken down and distributed onto machine learning algorithms that are organized in consecutive layers. Each layer builds up on the output from the previous layer. Together the layers constitute an artificial

neural network that mimics the distributed approach to problem-solving carried out by neurons in a human brain.” Available at: http://webfoundation.org/docs/2017/07/AI_Report_WF.pdf.

5. The engineers are responsible for thinking about the values that will go into the design of the artifacts, their function and their use manual. What escapes from the design and use manual does not depend on the control and influence of the engineer and can be unpredictable. That’s why engineers must design value-sensitive technical artifacts. An artifact sensitive to constitutionally guaranteed values (deliberate in the public sphere) is a liable artifact. It also necessary to think about the concepts of “inclusive engineering and “explainable AI”, to guarantee non-discrimination and transparency as basic principles for the development of these new technologies.

6. With this regard, to enhance the transparency and the possibility of accountability in this techno-regulated context, there is nowadays a growing movement in civil society demanding the development of “explainable artificial intelligences”. Also, the debate around a “right to explanation” for algorithmic and autonomous decisions that took place on discussions around the General Data Protection Regulation (GDPR) is also a way to achieve the goals of transparency and accountability since algorithms are taking more critical decisions on our behalf and is increasingly hard to explain and understand its processes.

7. ‘Causal nexus’ is the link between the agent’s conduct and the result produced by it. Examining the causal nexus determining what were the conducts, be they positive or negative, gave rise to the result provided by law. Thus, to suggest that someone has caused a certain fact, it is necessary to establish a connection between the conduct and the result generated.

8. This legal phenomenon is also called by other authors as “problem of the many hands” or “accountability gap”.

9. For the purposes of this article, although “fairness” can be understood as a broader term, it is addressed here on the topic of AI with a smaller scope, focused on algorithmic fairness. It is not in the scope of this article to expand the discussion of algorithmic fairness in special. A deeper exploration of this concept deserves a specific article focused on each guiding principle.

10. The type of insurance that should be applied to the case of intelligent robots and which agents and institutions should bear this burden is still an open question. The European Union’s recent report (2015/2103 (INL)) issued recommendations on the subject, proposing not only mandatory registration, but also the creation of insurance and funds. According to the European Parliament, insurance could be taken by both the consumer and the company in a similar model to those used by the car insurance. The fund could be either general (for all autonomous robots) or individual (for each category of robot), composed of fees paid at the time of placing the machine on the market, and / or contributions paid periodically throughout the life of the robots. It is worth mentioning that, in this case, companies would be responsible for bearing this burden. Despite this proposal, however, the topic continues open to debate, with new alternatives and more interesting models - such as private funds, specific records, among other possibilities - that will not be the subject of a deep analysis in this thesis.

11. *OODA* means the "*observe–orient–decide–act*" orientation cycle, a strategy developed by military strategist John Boyd to explain how individuals and organisations can win in uncertain and chaotic situations.

12. Parts of this subsection were built upon a recent and unpublished work of the author, in co-

authorship (Magrani, Viola, & Silva, 2019), and cited here to bring an updated vision of the author in dialogue with other recent publications.

13. The characteristics most used for the foundation of the human personality are: consciousness; rationality; autonomy (self-motivated activity); the capacity to communicate; and self-awareness. Another possible social criterion is to be considered a person whenever society recognises one (we can even apply the Habermasian theory here, through a deliberative process in the public sphere). Other theorists believe that the fundamental characteristic for the attribution of personality is sensibility, which means the capacity to feel pleasure and pain. The legal concept of a person is changeable and is constantly evolving. For example, afro-descendants have once been excluded from this category, at the time of slavery. Therefore, one cannot relate the legal concept of a person to *Homo sapiens*. A reservation is necessary at this point because even if robots can feel and demonstrate emotions as if they were sensuous, the authenticity of these reactions is questioned since they would not be genuine, but at most a representation (or emulation), analogous to human actors when they simulate these emotions in a play, for example, feelings in certain roles, not being considered by many as something genuine. Because of this, the Italian jus-philosopher Ugo Pagallo calls this ‘artificial autonomy’.

14. In the present article, it is argued that the consensus must be constructed according to Jurgen Habermas’s proposal, that is, through dialectical conflicts in the public sphere.

15. Such criticism, however, can be overcome by instruments already available on legal regulation. The recognition that the AI expresses a centre of interests would already be more than sufficient to admit that it has subjectivity and therefore deserving at least some rights. Nothing would prevent the granting of subjectivity to AIs as a mid-term regulation and leaving the path open for a future grant of an effective e-personality depending on the degree of autonomy (based on a matrix type). As an initial measure, it would play an important role in guaranteeing the reparation of victims, avoiding a scenario of ‘distributed irresponsibility’.