

CRIMINAL LIABILITY AND ARTIFICIAL INTELLIGENCE

RESPONSABILIDADE PENAL E INTELIGÊNCIA ARTIFICIAL

MATEUS DE OLIVEIRA FORNASIER

Pós-Doutorado em Direito pela University of Westminster (Reino Unido). Doutor em Direito pela Universidade do Vale do Rio dos Sinos (UNISINOS). Professor do Programa de Pós-Graduação em Direito (Mestrado e Doutorado) da Universidade Regional do Noroeste do Estado do Rio Grande do Sul – UNIJUI (Ijuí/RS).

ORCID: [<http://orcid.org/0000-0002-1617-4270>].

Lattes: [<http://lattes.cnpq.br/3316861562386174>].

DOI: [<https://doi.org/10.54415/rbccrim.v187i187.63>].

mateus.fornasier@unijui.edu.br

Recebido em: 19.04.2021

Aprovado em: 03.11.2021

Última versão do autor: 04.11.2021

ÁREA DO DIREITO: Digital

ABSTRACT: This article aims to relate legal personality and responsibility related to intelligent artificial entities. Its hypothesis is that not only general AI or superintelligence (still non-existent) can theoretically commit criminal acts: even common AI systems currently available can be trained by criminals to commit acts that violate fundamental rights. And culpable situations (due to negligence, recklessness or malpractice of programmers, designers and directors of AI development companies, for example) must be considered in the event of crimes involving AI. At least currently, therefore, alternatives more consistent with current Criminal Law should be considered for criminal situations involving AI, such as new criminal types in the current law against criminally responsible humans, or further emphasize civil punishment for damages caused by autonomous artificial entities. Specifically, its objectives are: i) to present the possibilities of criminal use of AI, as well as the legal impossibility of attributing criminal liability to

RESUMO: Este artigo objetiva relacionar a personalidade e a responsabilidade jurídicas de entes artificiais inteligentes. Sua hipótese é de que não apenas a IA geral ou a superinteligência (ainda inexistentes) podem teoricamente cometer atos criminosos: mesmo os sistemas comuns de IA atualmente disponíveis podem ser treinados por criminosos para cometer atos que violam os direitos fundamentais. E situações culpáveis (por negligência, imprudência ou imperícia de programadores, designers e diretores de empresas de desenvolvimento de IA, por exemplo) devem ser consideradas em caso de crimes envolvendo IA. Pelo menos atualmente, portanto, alternativas mais consistentes com o Direito Penal hodierno devem ser consideradas para situações criminais envolvendo IA, como novos tipos de crimes na lei vigente contra seres humanos criminalmente responsáveis, ou enfatizar ainda mais a punição civil por danos causados por entidades artificiais autônomas. Especificamente, seus objetivos são: i) apresentar as possibilidades de uso criminoso

artificial entities endowed with this technology at the current moment of technological evolution; ii) to explore new opposing theoretical and legal possibilities – in other words, what should be done for, one day, when there is a general AI, endowed with as much awareness as a human being and, therefore, perhaps guilty. Methodology: hypothetical-deductive procedure method, qualitative approach and bibliographic review technique.

KEYWORDS: Artificial intelligence – Criminal liability – Personality.

da IA, bem como a impossibilidade jurídica de se atribuir responsabilidade criminal a entes artificiais dotados dessa tecnologia no atual momento da evolução tecnológica; ii) explorar novas possibilidades teórico-jurídicas opostas – quais sejam, o que deve ser feito quando houver uma IA geral, dotada de tanta consciência quanto um ser humano e, portanto, quiçá culpável. Metodologia: método de procedimento hipotético-dedutivo, abordagem qualitativa e técnica de revisão bibliográfica.

PALAVRAS-CHAVE: Inteligência artificial – Responsabilidade penal – Personalidade.

SUMÁRIO: 1. Introduction. 2. Possibilities of Criminal Use of AI and the current impossibility of criminal liability of entities endowed with AI. 3. On theoretical possibilities of criminal liability of AI in the future. 4. Conclusion. 5. References.

1. INTRODUCTION

A humanoid robot called Sophia has generated controversy around the world because it has received citizenship and made presentations in the media since 2016. Although Hanson Robotics, the company that manufactured it, praised its creation as representing the future of artificial intelligence (AI), thinkers from the most varied areas have been less optimistic regarding the capabilities of such an artificial entity, describing it as a sophisticated puppet or chatbot (a type of conversational robot). In fact, they say, it is very likely that Sophia's performances were choreographed to promote specific political and economic interests, such as the promotion of the interests of the technology industries and their government promoters (PARVIAINEN; COECKELBERGH, 2020).

Socially, the limits of technology have been overcome each day more quickly, and popular culture has helped in this process by provoking, through their most popular representations (films, games, etc.), issues about values related to the human person and the attributes that differentiate them from other entities; but the specific legal aspects related to the personality fall behind in this regard, with sporadic and punctual advances. In spite of the fact that popular culture creates useful representations, its fictions do not allow the formulation of a coherent model for the entities to which legal personality could (or not) be attributed (ARNOLD; GOUGH, 2018, p. 31-33). AI is an expression used

to define a very wide range of technologies, comprehended from physical entities, such as autonomous cars, to other more abstract ones, such as parts of software – and such a spectrum must be increasingly nuanced and extended in its limits, given that the evolution of technology can bring several other ways still unthinkable. Thus, regulations relating to any form of AI cannot simply take a “one size fits all” approach, nor should it be expected that any intelligence capable of developing knowledge by itself should also be regulated effectively (CHEN; BURGESS, 2019).

It is very important to study criminal liability of AI. AI-based technology has become commonplace, but in most countries around the world there is still no specific regulation for its use, which results in a legal vacuum on the matter. Thus, when damage due to its use may occur, liability could, in theory, be assumed by several parties – from consumers, through producers and third parties (such as robot trainers or messengers), to the robot itself, and the definition of who (or what) is the responsible one will depend on how each country considers AI normatively: in the case of countries that follow the Saudi Arabian line, Sophia robot is considered a citizen – and, if this happens in other countries in the world, robots may be entitled to procedural legitimacy rights, to obtain equality before the law, in addition to others related to human citizenship. Thus, thinking about the legislation concerning AI is something that goes through many complexities, including those related to the involvement of many stakeholders. Several normative frameworks regarding accountability and legal personalization of such existing devices can be considered – from equating AI with living things, such as humans or pets or common products, to creating entirely new concepts for a legal framework that regulate AI-based systems (SUMANTRI, 2019).

Therefore, this article has the following question as a research problem: how should to criminal punishment be considered when entities endowed with AI are involved in crimes? As a hypothesis, it is assumed that not only general AI or superintelligence (still non-existent) can theoretically commit criminal acts: even common AI systems currently available can be trained by criminals to commit acts that violate fundamental rights. And culpable situations (due to negligence, recklessness or malpractice of programmers, designers and directors of AI development companies, for example) must be considered in the event of crimes involving AI. At least currently, therefore, alternatives more consistent with current Criminal Law should be considered for criminal situations involving AI, such as new criminal types in the current law against criminally responsible humans, or further emphasize civil punishment for damages caused by autonomous artificial entities.

The general objective of this article, which has an exploratory nature, a hypothetical-deductive procedure method, a qualitative approach and a bibliographic review research technique, is to relate the personality and the legal responsibility, with respect to these characteristics when related to intelligent artificial entities. In order to achieve that general objective, the development of the article was divided into two parts. The first of them aims to present the possibilities of criminal use of AI, as well as the legal impossibility of attributing criminal responsibility to artificial entities endowed with this technology at the current moment of technological evolution. The second one, on the other hand, explores new opposing theoretical and legal possibilities – which are what should be done for, one day – when there is a general AI, endowed with as much awareness as a human being and, therefore, perhaps to blame.

It is necessary to say that the article aims to provide a more comprehensive analysis, without delving deeply into the specific dogmatic difficulties of Brazil. That occurs because the discussion on this theme is still too incipient, and legal research must start with general political and philosophical issues firstly, in order to foment future research regarding specific cases and legal orders.

2. POSSIBILITIES OF CRIMINAL USE OF AI AND THE CURRENT IMPOSSIBILITY OF CRIMINAL LIABILITY OF ENTITIES ENDOWED WITH AI

An unintended consequence of the recent increase in AI research is the potential reorientation of AI technologies to facilitate criminal acts, which is theoretically viable thanks to published experiments on fraud automation targeting social media users and demonstrations of market manipulation simulated by IA. But since research on the criminal use of AI is still relatively incipient and inherently interdisciplinary, ranging from socio-legal studies to formal science, there is little certainty about the future of such a criminal use. A first good example systematic and interdisciplinary analysis of the literature on plausible threats of criminal uses of AI was made by King et al. (2020), who provided ethics experts, public policy makers and law enforcement organizations with a synthesis of current problems and a possible solution space:

(I) Areas: the areas of criminal use of AI can be better understood by listing the use of AI in interrogations, and theft and fraud in virtual spaces (online games with intangible assets that hold value in the real world, for example); and autonomous agents committing market manipulation. Although social engineering attacks can be considered plausible to occur, as well as homicidal and terrorist acts (using drones and autonomous vehicles, for example), there is still no evidence of these types of occurrences;

(II) Dual use: the digital nature of AI facilitates its dual use, that is, the use of applications designed for legitimate activities in malicious activities. If left untreated, this risk can lead to social rejection and overly strict regulations on AI-based technologies;

(III) Security: AI is taking an offensive role, in conjunction with defensive AI systems that are being developed and deployed to increase its resilience and robustness, and to combat threats related to its use as they arise. Although public policies can play a fundamental role in the mitigation and solving the risks of dual use after the implementation of applications liable to this type of use, it is at the design stage that these risks are dealt with in the most appropriate way. Overconfidence in AI can be counterproductive, however; thus, in addition to AI, the focus of research on security against the criminal use of AI should focus on people-users and social factors as well;

(IV) People: although psychological factors (trust in technology, for example) can play significant roles for the occurrence of crimes with the use of AI, there is still no research on the personal factors that can create perpetrators of crimes of this type (programmers and users, for example);

(V) Organizations: AI can play a role in criminal organizations (drug trafficking cartels, for example) and on the dark web, in criminal services sold directly to users, as a minor element in a general AI-fueled crime (for example, enabling the hacking of personal profiles for the most varied purposes).

The enthusiasm for the AI promises overcame the well-known challenge of implementing even weak AI – that is, an AI that does not “think for itself”, as if it were analog to a human conscience. However, weak AI can cause considerable costs and inconvenience when improperly or incorrectly implemented. Bots that deliver malware or spam, or automated forecasting systems incorrectly classifying the attributes of a decision-making domain, are good examples of damage that can be caused even by weak AI (BROADHURST et al., 2019).

Given the developments being made in the field of AI, there is a tendency to use AI as both an offensive and defensive tool. As AI is being developed, the methods of defense and attack against which we must develop defenses will also evolve reflexively, resulting in an evolving risk scenario. By adopting the use of AI, defensive systems can be significantly increased and better prepared for future attacks that may involve the misuse of AI.

AI-powered malware (even when such an AI is weak) is a threat that will become more prominent and poses a significant risk to a variety of potential targets – which comes with even more risk when aggregating Internet of

Things (IoT) and biometry. Predictive systems present risks if they are provided with inadequate data – stemming from prejudiced biases (gender, race, sexual orientation, origin, etc.), obtained in violation of the privacy of classified individuals, etc. – or attacked in a way that compromises AI decision-making, or that alters their own data sources and authenticity. For these reasons, offensive AI represents a significant opportunity for illegitimate uses and the ability to allow increasingly complex and effective attacks by malicious actors – especially in relation to attacks to financial institutions, transportation and logistics-related organizations, and other examples of industries already using weak AI applications.

As AI becomes more prominent and widely understood, it can also proportionally increase the frequency of damaging attacks on AI systems themselves. Malicious actors are able to exploit the “learning” processes of such systems and “teach” them again, inducing other forms of behavior. The introduction of data that was designed to “confuse” the system with incorrect classification can cause interruption and errors in the automation.

Despite significant advances, no AI application has yet developed an “awareness” (albeit remote) or independent motivation that approaches General Artificial Intelligence (GAI) and, therefore, the creators of systems can still be held responsible for criminal use in the majority of law systems around the world. AI technology has not advanced enough to warrant moral considerations – which could, for example, lead to considering an entity endowed with AI as having a unique identity and a moral personality. However, the unintended consequences of AI applications, as they evolve as exemplars of general-purpose technologies, will define new ethical challenges and provide new opportunities for criminal activities, as well as for crime prevention. However, the integration of AI into many manifestations of everyday life – from consumer monitoring to enhanced surveillance of individuals in public areas – will continue to be driven by investments from both commercial companies and governments. These developments will also reshape the opportunities for criminal activity to emerge.

In a recent past, several countries have begun to consider using Criminal Law to punish research, support and application of cloning technology – in this sense, Germany, Japan, China and France have formulated special norms containing criminal provisions to regulate and prohibit human reproductive cloning. Thus, maybe with the advent of a strong AI, that defies the pillars of traditional Criminal Law – which relies on an act originating from a guilty individual due to his conscience developed to a certain degree – this will become a strategy to be adopted by countries as well, in order to contain the risk represented by such

technology (LEI, 2019). Despite being an interesting path, however, this is still just speculation.

The use of AI systems in criminal activities is a new challenge for law enforcement, since such devices are capable, for example, of revealing information and imitating human personal characteristics – and thus the software in the hands of criminals makes potential victims much more vulnerable. Perhaps AI will be used not only to commit fraud, but also to facilitate violent crimes (against life, terrorism, etc.). For many regulatory fields, the use of AI can be regulated considering the adaptation of already existing norms, but not in the sphere of Criminal Law. Since errors in the legislation here have extremely serious consequences, some concepts and norms need to be revised (DREMLIUGA, 2020):

(I) The traditional concept of culpability should be revised, and new criteria should be established to assess a person's degree of culpability in case of using AI for crimes. Such a new paradigm could not be universal, as it has to obey national legal traditions, but it has some general points. The main challenge here is that a person responsible for serious consequences could not be considered as possessing a relevant degree of responsibility, as an innocent person could come to be recognized as responsible. New legislation must be drafted in relation to the sufficient characteristics of AI technologies – such as the fact that AI could have its own cognition and will, in any sense;

(II) The use, possession, and transfer of some types of AI must be controlled by governments – mainly AI systems that could significantly threaten society, the State and public order, such as those that combine AI with items and substances traditionally prohibited (weapons, explosives, narcotics, etc.), for which existing law cannot be considered enforceable because the combination of AI with something can be substantially more dangerous than the simple sum of its parts. Controlling the possession, use and transfer of dangerous AI systems is much more difficult than controlling traditional dangerous items. AI software is a computer code that could easily be copied, downloaded or inserted into equipment, which means that legal control and criminal liability must be introduced with great care to avoid abusing of personal rights and freedoms;

(III) The recognition of the legal personality of AIs in Criminal Law should also be avoided, as the artificial transfer of knowledge related to the experience of considering companies as criminally responsible is inappropriate for the case of AI, since the objectives of liability and punishment would not be achieved. There are no formal obstacles to creating such a personality, but that would not correlate either with the current legal order or with general morality. In addition, humans have no moral obligation to recognize AI as a legal entity.

As analyzed above, then, the punishment of artificial entities endowed with AI – which are hitherto endowed with weak AI, which is not self-conscious or even analogous to human intelligence – is not something plausible in the criminal field. But in intentional or culpable situations in which humans and AI are jointly involved in the occurrence of crimes, the punishment of the human, and not the entity itself, should be considered. This does not mean, however, that one day general AI or superintelligence will not be developed – and when that day comes, new theories and punitive strategies must be thought of. Therefore, the following section will deal with this matter.

3. ON THEORETICAL POSSIBILITIES OF CRIMINAL LIABILITY OF AI IN THE FUTURE

Osmani (2020) shows that the issues on liability regarding AI systems go beyond the limits of the foundations of traditional Criminal Law, and that several of the definitions of modern Criminal Law are not well-tuned with the era of advanced AI. Thinking of robots as being subject to Law is not a simple solution due to the uncertainty of many issues involving ethics and robots. While giving civic status to AI systems may cause some regulation for the industry related to that technology, many other problems, especially those related to the intentional action and culpability of AI systems, still need to be extensively researched and debated in order to achieve plausible resolutions someday. Therefore, the model of responsibility centered on the individual must be overcome in favor of a model of responsibility centered on the organization, and corporate criminal responsibility is a good solution whenever the liability link of individual agents cannot be conclusively established. However, specific norms and concepts in the field of AI will be necessary – not only with regard to separate statutes, but rather with a *sui generis* logic – in order to properly establish forms of criminal punishment related to AI.

In this sense, it is necessary to understand whether society considers it possible to consider some kind of legal personality to intelligent entities. Although it analyzed a very small sample of individuals (N = 3315), the research by Lima et al. (2020a) on the perception of online users about the legal personality of AI and robots noted that while people do not tend to attribute any state of mind to electronic agents and also that they are not willing to grant physical independence or asset ownership to AIs and robots – factors that could serve as a basis and prerequisites for civil or criminal liability – they consider that robots and AI entities should be responsible for their actions, and thus, worthy of punishment. In this

sense, it is illustrated that the establishment of the figure of the electronic legal entity is still not feasible without a wide legal reformulation or development of systems that satisfy the liability requirements – and, although the granting of legal status to electronic agents may not be viable currently, the punishment of AIs and robots can become imperative, since these systems are widely deployed and cause damage (LIMA et al., 2020b). The general public could hold AI and robots causally responsible for their actions and believe that they should be punished for their actions, then.

Moreover, many consider that humans are empirically considered retributivists when dealing with other humans. And the punishment of electronic agents could be considered psychologically beneficial for individuals and for society as a whole, in this sense. According to Lima et al. (2020b), scholars have already criticized the fact that retributivism (that is, revenge) is usually the main aspect to attribute punishment to transgressors, and the retributivist feeling is more strongly maintained when the transgressor is factually punished. On the other hand, the need for revenge is more easily overlooked if the wrongdoer does not suffer the consequences. Therefore, it may be possible to use the punishment of electronic agents as a way to motivate victims to abandon their retributivist feeling, leading to improvements in individual post-traumatic feelings and to a society that focuses its punishment on reform, instead of revenge.

Complexity of assigning criminal responsibility for AI systems requires investigation by the various stakeholders who could be held responsible for the system design, from the hardware vendor, through programmers and installers, to company policy-makers, to oversee the functioning of the system (MAJUMDAR; RONALD; RAUTDESAI, 2019). It should also be investigated whether the AI entity can be considered as having a “leading mind” in order to address the determination of liability, which, depending on the situation, may attract indirect liability, breach of contract or criminal liability. In addition, the issue of determining liability collides in parallel with the question whether the legal personality may be granted to AI or not.

Kingston (2016) identifies that three factors at least are essential to assess the responsibility of an AI system: (I) the definition of AI – whether it is a product or a service (something that is still a reason for controversy in the legal debate); (II) For the consideration of criminal offenses, which mens rea is necessary – it does not seem likely that AIs will violate norms that require the knowledge of the perpetrator to characterize the occurrence of a crime, but it is possible that they violate the rules that require the knowledge analog to that of “a reasonable man”, and it is almost certain that they could violate crimes of strict

responsibility; (III) If the limitations of the AI systems are communicated to the user/ consumer – as the AI systems have general and specific limitations, the legal cases on these issues can be based on the specific formulation of notices about such limitations.

With regard to the question of the insertion of machines in the general framework of Criminal Law, it is necessary to distinguish between the mental and material elements of crime, as well as overcoming the difficulty that jurists present when dealing with the constantly evolving autonomy of AI-equipped machines, so that the chain of responsibility through notions of legal cause and culpability may be interrupted. The increasing autonomy of AI machines in terms of legal causation and culpability is likely to suffer different forms of discipline in national legal systems, and later those normative experiences will be formalized in international frameworks (PAGALLO, 2016).

As for the *mens rea* of a new generation of robotic crimes, the hypotheses of robots endowed with free will, autonomy, or human moral sense, must be disregarded, at least in the short and medium terms. And when, in the more distant future, machines of this type appear, Law as a whole (not only its Criminal branch) must undergo a total review of fundamentals. And as for the acts of new robotic crimes, the loss of significant human control over strong AI agents must be taken seriously, as systems with deep learning change and improve, and then could trigger artificial processes that would cause an unstoppable machine, or a machine that adapts itself only in accordance with human needs. Thus, the verifiability of such systems and the reflection about their undesirable behaviors must be improved so that these machines are kept under control.

As AI is being used in an increasing range of activities – from driving automobiles, through operations in the financial markets, reaching applications in the Judiciary and Medicine – it represents an increasing risk not only related to the commitment of civil unlawful acts by artificial entities, but also to activities that for humans or corporations could be considered crimes. As they argue that the current Criminal Law is insufficient for these cases due to gaps in criminal liability, Abbott and Sarch (2019) suggest that the direct criminal punishment of the artificial entity itself, even when acting autonomously, would be inadmissible according to basic penal principles that demand legal capability for finding someone guilty, as well as to the requirement of a guilty mind – although analogies to corporate criminal responsibility and imputation principles in vogue allow theoretical constructions of AI criminal punishment.

The use of legal fictions to solve difficult conceptual issues or practical problems – such as creating specific types of mental elements for AI or bad behavior

by its developers – gives the theory of Criminal Law an impressive plasticity, which corroborates the transformation of that legal branch in a pragmatic tool for the solution of social problems. But legal fictions should be used with caution, as their excessive use risks public trust and the rule of law, since their uncontrolled proliferation can lead to widespread injustice, whether by punishing in a much more severe way than guilt requires, either for punishing innocents.

The criminally punishing AI can carry significant costs and require radical legal changes. Therefore, modest changes in existing Criminal Law norms that target people, together with potentially expanded Law of Torts, would constitute better solutions for crimes committed with the use of AI. A suitable alternative would involve modest expansions to Criminal Law, including the typification of new crimes of negligence centered on inappropriate design, bad operation and insufficient testing of AI applications, and possible criminal penalties for designated parties that do not comply with legal duties – added, of course, to a greater degree of relevance given the civil liability for those offenses.

By their turn, Lemley and Casey (2019) make an important counterpoint between explicability and possibilities of judicially ordering a conduct based on an AI-based robot (more precisely, in modern AI techniques, such as deep learning and artificial neural networks). It can be ordered that a robot – or, more realistically, its designer or owner – be punished for the damage it causes. But it turns out to be much more difficult for a judge to “order” a robot, instead of a human, to assume or refrain from certain conduct, as robots cannot directly obey Court orders that are not written in computer code. And filling the translation gap between natural language and programming code is generally more difficult than expected – mainly because of modern AI techniques that enable machines to learn and modify their decision-making processes autonomously. If such procedures are not explicable, it will be impossible to know how to inform the artificial autonomous entity about how it must behave in a desirable way. Thus, many of the current significant legal doctrines on punishment, obligations and procedural acts must be rethought, simply because the social role of robots was not seriously considered in the past, although they tend to increasingly become more present in society and more fundamental to human sociability. Then, although in the current criminal systems of the world AI cannot suffer any kind of criminal execution – requiring a natural or legal person to be responsible if there is a crime resulting from activities involving such technology –, with technological evolution, AI will be much more sophisticated in the future and, most likely, close to human capabilities (RAHMAN; HABIBULAH, 2019).

On the other hand, Kirpichnikov et al. (2020) consider seriously the possibility of making AI criminally liable. AI would, for the authors, be able to correct their

own behavior under the influence of coercive measures. However, Criminal Law and public security policies must be adapted to this new reality. In the same sense, in addition to demonstrating that AI systems can engage in activities that would constitute crimes if carried out by humans, Lagioia and Sartor (2019) considered that certain AI systems, in particular those that implement a “Belief, Desire and Intention” (BDI) approach, may also have the cognitive states relevant to Criminal Law (cognition/awareness and volition/intention, as required for *mens rea*).

BDI approach has been used to not only capture practical human cognition, but also to model and develop AI architectures. Systems with that AI paradigm have these three “mental” attitudes:

(I) Belief: which represents the agent’s informational status, including perceptual content, but also additional information obtained through reasoning, especially through the use of inference rules;

(II) Desire: which incorporates the motivational state, with desires representing objectives or situations that the agent would like to accomplish. In addition, an agent may have several desires, which may be in conflict;

(III) Intention: which represents the result of the agent’s deliberation, what the agent decided to do because of his beliefs and desires. Intentions are action plans to which the agent has committed, and must be mutually consistent.

An artificial agent with a BDI architecture will implement its current intentions as soon as it has the opportunity to do so, without further deliberation. However, it can revise and abandon its intentions before its implementation based on new deliberations. BDI architecture has the following components:

(I) A “stock” of beliefs, which contains the information that the AI being has (that is, that he “learned” about the world). With a perceptual processor, the agent observes the environment and interprets the data coming from the sensors, providing information for storing new beliefs;

(II) A set of goals/desires adopted by an agent – and, in order to achieve these objectives, the agent will build plans and store them in its “library” (memory);

(III) A library of plans, which an agent can use to achieve a specific objective – being that each plan is composed of three main elements:

i) Body: the concrete set of actions necessary to fulfill a plan;

ii) Conditions of invocation: the circumstances under which the agent must consider and activate a plan if it believes that such circumstances occurred;

iii) Termination conditions: circumstances that define the conditions under which the agent can reconsider his current intention, based on new circumstances that may render a plan unattainable.

Furthermore, some systems may also have the level of rationality and responsiveness needed for criminal liability like that – and, for that, an AI system would need, at least, to have a level of prudence that enables it to anticipate the negative effects of punishment in compliance with its goals.

Given this situation, in which AI agents could be considered capable of criminal punishment due to their level of autonomy, Law could require limiting the machine's autonomy in certain areas – especially in those involving intentional damage to humans, even if for legitimate reasons – in view of the risks that come with such tasks, particularly the loss of human life. However, as it is not always possible or desirable to limit the autonomy of AI systems, crimes in which such systems are involved need more legal responses. In many cases, no human user/controller would be criminally liable, as he/she could not be considered an accomplice or instigator of AI crimes. On the other hand, civil liability may not deter sufficiently unlawful acts committed by (or with) AI, since such form of liability is limited only to the compensation of the victims. Thus, the deployment of AI criminal systems could be addressed primarily by a broad interpretation of the notion of recklessness (*dolus eventualis*), so that implanters could be held criminally responsible for AI crimes whenever they are in a position to enable considerations about the occurrence of crimes by their systems.

The creation and deployment of criminal AI systems can be seen as a separate criminal offense, at least when potential victims are endangered. And in some cases, a necessary precaution – the omission of which can lead to criminal liability – may consist of endowing the AI system with a normative architecture. It may also become possible, in the future, to directly punish criminal AI systems by devising appropriate measures. If Law develops in that direction, AI systems should have a sufficient level of reason and responsiveness, at least in the sense that their actions can be influenced by the prospect of implementing criminal sanctions.

Simmler and Markwalder (2019) consider that criminal liability for robots is possible only if it is useful and necessary for stabilizing the rules of the penal system – which would happen only with the fulfillment, by the robots, of the personality requirements and of the capacities inherent to it: characteristics that are socially attributed, not constituting a biophysical fact. A possible attribution of criminal responsibility to robots is therefore a process subject to the function of the system. As a consequence, there are three main alternative paths, which the theory of Criminal Law could follow in the context of the “Law of robotics”:

(I) Insisting on the traditional approach to criminal liability, which is based on freedom of choice, on the ideal of the autonomous human being, and on the exclusive application of the concept of personhood to humans. As a result,

robots would never be held responsible, regardless of technological progress and the problems it may cause. Thus, it would be necessary to focus on the programmer or operator of the robot for criminal liability issues;

(II) applying criminal responsibility functionally, expressing a censure for the lack of socially visible compliance with the Law, seems to be more realistic according to the current state of the art of technology, and it also has the potential to destabilize the rules. This notion of criminal responsibility is not based on recognizable human free will or freedom of choice, but on a construction that is possible if technological developments take place, in relation to robots, so that they can be experienced as “equal” – constituted as recipients of normative expectations in social interaction (and not means), in the same way that human individuals and corporate entities are today;

(III) Taking advantage of the discussions raised by technological developments to rethink the traditional approach to criminal liability. Thus, traditional elements contained in this notion, based on “free will” and “humanity” should be overcome – not in order to exclude them completely from the system, but rather, to look for suitable functional equivalents, free from concepts and ideas that continue to destabilize the currently prevailing understanding of criminal liability, but that nonetheless ensure that the purpose of Criminal Law – stabilization of rules through punishment – is fulfilled. This path would not only pave the way to face new challenges regarding the growing relevance of robotics, but would also allow orienting criminal legal doctrine towards a more ideal way of fulfilling its function.

4. CONCLUSION

Perhaps one day humanity will be able to technologically reach the point of creating a general AI, and when it happens, ideas about what Law, guilt and responsibility are should be developed (and transformed) enough for allowing legal imagination to deal with that truly revolutionary novelty. That will be the moment from which all ideas resulting from the biopsychosocial evolution of humanity will be considered by entities more evolved than their creators, and who will probably carry out well-considered and precise rational calculations about their importance for the continuity in a situation of peace. Although it is more frequent to think about the possibilities of criminalizing autonomous artificial entities – which would require the development of the so-called general AI, or even of a superintelligence – even weak AI applications can represent terrible damage to people and organizations. Especially with the use of AI-equipped bots

that spread malware and through the improper use of AI systems already inserted in everyday life, which could be “retrained” by malicious agents so that they behave differently – erroneously classifying the data that store and/or making decisions that are offensive to fundamental rights, such as the privacy of users, their honor and image, etc.

The development of AI forces governments, legislators and researchers to reconsider traditional conceptions of Criminal Law. The traditional theory of culpability does not correspond to trends in the development of AI systems, as it does not allow us to consider that AI has some own analogue of free will and cognitive skills. Furthermore, it is difficult to take into account the fact that it is often not a criminal, but others (such as a developer) who are responsible for the final outcome of a committed crime.

Although criminal punishment of artificial autonomous entities may seem theoretically possible – by creating mental states and analogies to penalize companies, for example – it requires caution before the pillars of current Criminal Law. It would require radical changes that could undermine public trust and the rule of law, bringing disproportionate or innocent punishments. Thus, simpler changes, such as the introduction of some new types of penalties related to criminally responsible humans in relation to the acts of such apparatuses, as well as the increase in the significance of civil punishment for the damages caused by AI, would be a much more democratic and much more plausible theoretical path to follow.

AI systems could be structured on an architecture that endows them with beliefs, desires and intentions (BDI), which will provide them with sufficient levels of rationality – so that, theoretically they could be criminally punished directly. However, as long as Law is not fully reprogrammed in this sense, criminally liable humans must be considered and inserted into Law’s framework, mainly due to possible fraud and recklessness. One can also think about the autonomy limitations required by Law regarding machines that take on tasks that put human lives at risk.

Criminal liability must be rethought due to the possibilities brought about by the evolution of robotics. In this exercise, traditional and metaphysical elements of guilt – mainly “free will” and “humanity” – must be removed from legal system, in favor of a possibility of experimenting with robots, in their social interactions with humans, as “similar” ones. But that means not only undermining the system of such metaphysical elements, but rather, analyzing the possibility of fulfilling the function of Criminal Law – stabilizing its rules through the application of punishment. That is, it is not just a question of attributing or not legal

personality to robots (and, as a consequence, culpability for criminal purposes), but to analyze whether they behave in a similar way to humans in their social interactions.

5. REFERENCES

- ABBOTT, Ryan; SARCH, Alex. Punishing Artificial Intelligence: Legal Fiction Or Science Fiction. *University of California Davis Law Review*, v. 53, n. 1, p. 323-384, 2019. Available at: [https://lawreview.law.ucdavis.edu/issues/53/1/articles/53-1_abbott_sarch.html]. Access in: 15 apr. 2021.
- ARNOLD, Bruce Baer; GOUGH, Drew. Turing's People: Personhood, Artificial Intelligence and Popular Culture. *Canberra Law Review*, v. 15, n. 1, p. 4-33, 2018. Available at: [www7.austlii.edu.au/cgi-bin/viewdoc/au/journals/CanLawRw//2017/1.html]. Access in: 15 apr. 2021.
- BROADHURST, Roderic et al. *Artificial Intelligence and Crime*: Report of the ANU Cybercrime Observatory for the Korean Institute of Criminology. Korean Institute of Criminology; Australian National University Cybercrime Observatory; College of Asia and the Pacific: Canberra, 2019. Available at: [https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3407779]. Access in: 15 apr. 2021.
- DREMLIUGA, Roman. How development of Artificial intelligence technology will cause changes in crime and criminal law. In: KOKURYO, Jiro; MARACHE, Catharina; WALSH, Toby. *AI for Everyone: Benefitting from and building trust in the technology*. [S/l]: AI Access, 2020.
- CHEN, Jiahong; BURGESS, Paul. The boundaries of legal personhood: how spontaneous intelligence can problematise differences between humans, artificial intelligence, companies and animals. *Artificial Intelligence and Law*, n. 27, p. 73-92, 2019. DOI: [<https://doi.org/10.1007/s10506-018-9229-x>].
- KING, Thomas C. Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Science and Engineering Ethics*, v. 26, n. 1, p. 89-120, 2020. DOI: [<https://doi.org/10.1007/s11948-018-00081-0>].
- KINGSTON, John K. C. Artificial Intelligence and Legal Liability In: BRAMER, Max; PETRIDIS, Miltos (Ed.). *Research and Development in Intelligent Systems XXXIII: SGA12016*. Cham: Springer, 2016. DOI: [https://doi.org/10.1007/978-3-319-47175-4_20].
- KIRPICHNIKOV, Danila et al. Criminal Liability of Artificial Intelligence. *E3S Web of Conferences*, v. 159, p. 1-10, 2020. DOI: [<https://doi.org/10.1051/e3s-conf/202015904025>].
- LAGIOIA, Francesca; SARTOR, Giovanni. AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective. *Philosophy & Technology*, v. 33, n. 3, p. 433-465, 2019. DOI: [<https://doi.org/10.1007/s13347-019-00362-x>].

- LEI, Wei. Legal risk and criminal imputation of weak artificial intelligence. *IOP Conference Series: Materials Science and Engineering*, v. 490, n. 6, p. 1-6, 2019. DOI: [https://doi.org/10.1088/1757-899X/490/6/062085].
- LEMLEY, Mark A.; CASEY, Bryan. Remedies for Robots. *The University of Chicago Law Review*, v. 86, n. 15, p. 1311-1396, 2019. Available at: [https://lawreview.uchicago.edu/publication/remedies-robots]. Access in: 15 apr. 2021.
- LIMA, Gabriel et al. Explaining the Punishment Gap of AI and Robots. *Arxiv*, p. 1-10, 2020a. Available at: [https://arxiv.org/abs/2003.06507]. Access in: 15 apr. 2021.
- LIMA, Gabriel et al. Will Punishing Robots Become Imperative in the Future? In: BERNHAUPT Regina et al. (Ed.). *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, p. 1-8, 2020b. DOI: [https://doi.org/10.1145/3334480.3383006].
- MAJUMDAR, Priyanka; RONALD, Bindu; RAUTDESAI, Rupal. Artificial Intelligence, Legal Personhood and Determination of Criminal Liability. *Journal of Critical Reviews*, v. 6, n. 6, p. 323-325, 2019. DOI: [http://dx.doi.org/10.22159/jcr.06.06.48].
- OSMANI, Nora. The Complexity of Criminal Liability of AI Systems. *Masaryk University Journal of Law and Technology*, v. 14, n. 1, p. 53-82, 2020. DOI: [https://doi.org/10.5817/MUJLT2020-1-3].
- PAGALLO, Ugo. AI and bad robots: the criminology of automation. In: MCGUIRE, M. R.; HOLT, Thomas J. (Ed.). *The Routledge Handbook of Technology, Crime and Justice*. London: Routledge, 2016, p. 643-653.
- PARVIAINEN, Jana; COECKELBERGH, Mark. The political choreography of the Sophia robot: beyond robot rights and citizenship to political performances for the social robotics market. *AI & Society*, p. 1-10, 2020. DOI: [https://doi.org/10.1007/s00146-020-01104-w].
- RAHMAN, Rofi Aulia; HABIBULAH, Rizki. The Criminal Liability of Artificial Intelligence: Is It Plausible to Hitherto Indonesian Criminal System? *Legality*, v. 27, n. 2, p. 147-160, 2019. DOI: [https://doi.org/10.22219/jihl.v27i2.10153].
- SIMMLER, Monika; MARKWALDER, Nora. Guilty Robots? Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence. *Criminal Law Forum*, v. 30, p. 1-31, 2019. DOI: [https://doi.org/10.1007/s10609-018-9360-0].
- SUMANTRI, Viony Kresna. Legal Responsibility on Errors of the Artificial Intelligence-based Robots. *Lentera Hukum*, v. 6, n. 2, p. 333-348, 2019. DOI: [https://doi.org/10.19184/ejhl.v6.i2.10154].



PESQUISAS DO EDITORIAL



ÁREA DO DIREITO: Digital

Veja também Doutrinas relacionadas ao tema

- A imputação de responsabilidade civil por danos decorrentes do uso de sistemas totalmente autônomos, de Julia Pinto Lemos – *RD Tec* 9;
- Da possibilidade de responsabilização cível e criminal do ente dotado de inteligência artificial, de Fernando Rodrigues Peres e Larissa Guidorizi de Barros – *RD Tec* 11; e
- E-personality aos robôs dotados de inteligência artificial: uma terceira via de personalidade? utopia ou realidade? – a necessidade de discutir a interação entre seres humanos e robôs, de João Sérgio dos Santos Soares Pereira – *RD Tec* 7.