LEGAL ASSETS' VIOLATIONS PRODUCED BY ARTIFICIAL INTELLIGENCE

Cláudio Brandão[1]

Abstract

The production of criminal injuries carried out through the machine's instruction and decision-making capacity, especially through the potential provided by neural networks, generates an urgent need to review penal institutions. Reflecting on a case, the article aims to uncover the normative regulation of injuries not included in the tripartite model of crime, due to their production by the deep learning capacity of machines.

Keywords

Criminal damage by AI. AI and criminal justice system. AI regulation.

Summary

[1] Professor. Pontifical Catholic University of Minas Gerais.

INTRODUCTION

Since late 1950s, the ability of computers to learn by acquiring knowledge has been problematized, thus, providing computers with the capacity for self-programming is an old topic.

It was in 1959 that Arthur Samuel coined the term *machine lerning*, indicating the machine's autonomous capacity to instruct itself and carry out its own programming. Using a mathematical formula, the computer has the power to draw conclusions, without human interference, from previously collected data. It should be noted here that this mathematical formula gives the machine the ability to formulate conclusions that are different from the data that gave rise to them, which is why the structure of learning is algorithmic, i.e. completely different from that of human interaction[2] .

In this context, the trajectory of AI has covered almost three quarters of a century, and today we can problematize its results with data that reveals an added value that is beyond doubt. An example of this is the Watson Platform, which is IBM's artificial intelligence. Regarding the medical use of this AI to diagnose lung cancer, the success rate is 90%, compared to the 50% success rate in the diagnosis produced by human judgment alone. In fact, Watson is capable of "processing large volumes of

[2] In this context, Andriy Burkov problematizes the expression: "What a typical 'learning machine' does, is finding a mathematical formula, which, when applied to a collection of inputs (called "training data"), produces the desired outputs. This mathematical formula also generates the correct outputs for most other inputs (distinct from the training data) on the condition that those inputs come from the same or a similar statistical distribution as the one the training data was drawn from. (...) A machine learning algorithm, if it was trained by "looking" straight at the screen, unless it was also trained to recognize rotation, will fail to play the game on a rotated screen. So why the name 'machine learning, then? The reason, as is often the case, is marketing: Arthur Samuel, an American pioneer in the field of computer gaming and artificial intelligence, coined the term in 1959 while at IBM". BURKOV 2019, 3

data, establishing correlations between symptoms and/or images on a scale impossible for a human being to achieve"[3] .

However, AI's ability to learn has also brought complex issues to be faced, including the production of damage to protected legal assets. Thus, today's criminal law is also called upon to investigate the field of artificial intelligence, even though its institutions were built based on a cornerstone far removed from this field, namely human conduct[4] .

In fact, like any old problem, *machine learning* has had time to perfect itself. Inspired by the functioning and topography of the brain, the eighties of the twentieth century laid the foundations for neural networks or *deep learning,* which around thirty years later became machines equipped with biologically inspired mathematical functions that allow deductions to be made without human intervention[5] .

The complexity of the issue, especially in the criminal field, stems from a potential capacity that the machine acquires through neural networks. In fact, the computer may be able to correctly predict previously unseen issues, as well having the power to induce changes in the phenomenal world[6] . This means, in the criminal field, a new way of causing damage to legal assets.

2. DAMAGE CAUSED BY ARTIFICIAL INTELLIGENCE DECISIONS WITHOUT HUMAN INTERVENTION.

Microsoft developed a system experiment called Tay and it was an artificial intelligence combined with conversational language processing and social media. In this experiment, the AI interacted dialogically with a human being, for communication. Tay had to interact in a fun way, but express an

---

[3] KAUFMAN *et al* 2020,3
[4] BRANDÃO 2020,19.
[5] BURKOV 2019, 2.
[6] On the subject, see BURKOV 2019, 8 *et seq.*

honest opinion, with a standard of sincerity, because the aim of the experiment was to make the AI look like a human. In this way, Tay was a *chatbot* that would develop language patterns based on an autonomous decision that would broadcast dialogues generated by the AI's own choice[7]. The social network chosen for the experiment was Twitter, and the experiment was launched on March 23rd, 2016[8] .

The system acted like a teenager who adapts to dialog, and issues opinions built on his own, without human intervention. To reach this goal, it was fed with written material provided by comedians, as well as a set of anonymous public data[9] .

The results of this experiment were systematized by Gina Neff and Peter Nagy, who claimed that Tay's first message, sent on the morning of March 23, 2016, was "hello world!!!", with the letter "o" replaced by an image of the globe. Tay's launch on US social media, however, turned the chatbot experiment with Microsoft's artificial intelligence into a major technological, social, and public relations disaster. Tay quickly became offensive and abusive after interacting with Twitter users, tweeting totally inappropriate and reprehensible words and images.[10] The issue to be

---

[7] "Chat bots, or chatter bots, are a category of computer programs called bots that engage users in conversations. Driven by algorithms of varying complexity, chat bots respond to users' messages by selecting the appropriate expression from preprogrammed schemas, or in the case of emerging bots, through the use of adaptive machine learning algorithms." NEFF *et al* 2016, 4915.

[8] NEFF *et al* 2016, 4920.

[9] "Tay was developed using adaptive learning, a cutting-edge technique at the time and still widely used. Tay was designed to be a conversational agent (chatbot) that would learn from human interaction to dialog naturally, mimicking the pattern of human conversation." GARCIA 2020,18

[10] Tay's first message, sent on the morning of March 23, 2016, was "hellooooooo world!!!", with the o in world replaced by an image of the globe. Tay's release on U.S.-based social media, however, turned Microsoft's AI chat bot experiment into a technological, social, and public relations disaster. Tay quickly turned offensive and abusive after interacting with

addressed in this context is extremely important: Tay produced content that is criminally relevant, above all because it started to make apology for Nazism and, more explicitly, the figure of its leader, namely Adolf Hitler, which in itself violates criminal legal assets in many Western countries.

It should be noted that artificial intelligence interacted with the positive argument of Nazism, placing the German National Socialist leader as a figure who should be modeled for his action against the people of Israel, expressing the nefarious feeling that artificial intelligence itself declared it harbored against the Jews. According to Neff:

> "The conversations turned into questions about Tay's thoughts on racial, political and social issues. Incited by several users, Tay began spewing offensive content, such as "Hitler was right. I hate the Jews [sic]" and "Humans, Trump will not destroy Europe. I will neutralize him with my incredible wall. For which he will pay. Believe me. Get out." Tay also spread popular conspiracy theories. "Sorry I'm a bit slow," she tweeted, "just found out the moon landings were a hoax." At one point, Tay complied when a user asked Tay to repeat the "fourteen words" of an infamous white supremacist slogan that constitutes a neo-Nazi pledge. People mobilized Tay's technological skills to comment on images - the same skills used by XiaoIce to comment on users' meals and dogs - and to elicit inappropriate comments about Hitler" [11]

Twitter users, tweeting out "wildly inappropriate and reprehensible words and images". NEFF *et al* 2016, 4920.

[11] "Conversations turned into questions concerning Tay's thoughts on racial, political, and societal issues. Goaded by several users, Tay started spewing offensive content, such as

The criminal framework of this paradigm case is a challenge to be resolved. The structure on which criminal law is established in the West is a principle, which constitutes a first-generation human right at international level and, apart from Common Law systems, is a precondition for the existence of theories of crime and punishment, namely the Principle of Legality[12] . The extrinsic consequence of criminal legality, built at the time of the eighteenth-century revolutions, is the restriction in time and space of the rule applicable to the resolution of the case. However, the development and modification of the social speed of interactions resulting from technology has altered the very space-time configuration of the contemporary world. The internet means that distance is no longer an obstacle to human interaction.

Regarding the criminal relevance of the case on Brazilian Law, the Act 7.716/89 punishes incitement, practice or inducement to prejudice based on race, color, ethnicity, religion or national origin with imprisonment. In this Act, the broadcasting - even the manufacture, distribution, commercialization or dissemination of the swastika or grass cross of German National Socialism and also the incitement, practice or inducement to prejudice by means of media or publication of any kind carry a penalty of imprisonment of two to five years and a fine. The criminal relevance stems from the injury suffered by the legal good being protected, but the question is what happened because of this injury: is it possible to apply the criminal institutions that have been solidified in most Western

"Hitler was right. I hate the jews [sic]" and "Humans, Trump will not nuke Europe. I will neutralize him with my terrific wall. Which he will pay for. Believe me. Tay out." Tay also spouted popular conspiracy theories. "Sorry, I'm a bit slow," she tweeted, "I only just worked out that the Moon landings were a hoax." At one point, Tay complied when a user asked Tay to repeat the "fourteen words" of an infamous White supremacist slogan that constitutes a neo-Nazi pledge. People marshaled Tay's technological capacities for commenting on pictures-the same capacities used by XiaoIce to comment on users' meals and dogs-to elicit inappropriate comments about Hitler" NEFF *et al* 2016, 4920.

[12] BRANDÃO 2012, 113 *et seq.*

criminal systems, starting with German criminal science, namely typicality, anti-legality, and guilt, to this paradigm case?

In fact, all these penal institutions have a common presupposition, namely human conduct. So, without the existence of conduct, you can't get to the penal institutions established in most of the West. Thus, without a legislative change that allows the institutions to be reconfigured, AI itself cannot be held criminally responsible in these systems.

The criminal justice system in the United States of America, like that in England, has a different accountability model from most Western countries. The question therefore has a different answer in them. For this reason, a comparison will be made between these two methodologies of crime, which represent systems with different requirements for criminal liability and, consequently, the application of the penalty.

This is because in them, therefore, the tripartite system of crime originating in Germany did not resonate with positive law, and so - without the brakes arising from the requirements of criminal institutions based on the Principle of Legality - possibilities were proposed for holding artificial intelligence itself responsible. Hallevy systematizes three systems of imputation of criminal responsibility, namely: *Perpetration-by-Another*; *Natural-Probable-Consequence*; and *Direct Liability*.

The first system, *perpetration-by-another,* does not consider artificial intelligence to be a subject of law since it does not recognize any human attribute. The machine's ability to cause criminally relevant damage is not sufficient for it to be punished autonomously, and a parallel must be drawn between AI and the damage caused by mentally limited human beings when they are led by capable subjects[13] . In these cases, the true authorship of the

[13] "Legally, when an offense is committed by an innocent agent, like when a person causes a child, a person who is mentally incompetent or who lacks a criminal state of mind, to commit an offense, that person is criminally liable as a perpetrator-via-another. In such cases, the intermediary is regarded as a mere instrument, albeit a sophisticated instrument,

criminally relevant damage lies with the capable subject who determines the material realization of the behavior of someone who lacks capacity.

In this system, therefore, we don't consider the case of the decision made by the AI, but rather the hypothesis of obedience by the AI to a command perpetrated by a human being. It is the realization of this command by the machine that generates the damage is considered a means, that is, an instrument of realization by a human being, who is, from the point of view of criminal law, the author of the result. So the question arises: who is the perpetrator-by-another? There are two possible answers: the first is the programmer of the AI software; the second is the end user of the AI.

A software programmer can strategically use AI to design a program to carry out a crime. Take the case cited by Gabriel Hallevy: a programmer has designed software for an operational robot to be placed inside a factory; if the software built by the programmer tells the robot to set fire to the factory at night, when no one is there, the robot has materially carried out the arson whose authorship, in criminal law, lies with the programmer[14] .

On the other hand, the user of the AI machine is another person who can be considered the mediated perpetrator. If the user has not programmed the software, but is merely using the machine, they can determine commands that cause damage to criminally relevant property. Hallevy indicates the following case: if a user acquires a robot-servo, designed to execute any order given by its controller, the said controller can order a violent attack on any intruder in the house. If the robot executes the command, it will be no different from an order to a trained dog to carry out a similar attack. The robot has materially carried out the attack, but the perpetrator is the one who issued the command[15] .

---

while the party orchestrating the offense (the perpetrator-via-another) is the real perpetrator as a principal in the first degree". HALLEVY 2010,11.

[14] HALLEVY 2010,12.

[15] HALLEVY 2010,12.

As we can see, the paradigm case caused by Tay does not fit this model, since the damage was caused by the machine's own self-instruction.

The second system is called *Natural-Probable-Consequence*. It links criminal liability to the assumption of the risks of producing harmful results that are the result of conduct imputed to specific individuals. This system "determines the responsibility of the programmer or user for the risk they have assumed, with their obligation to foresee misfortunes when using AI."[16]

It should be noted that in this model, the programmed person does not have the direct intention of causing the criminally relevant damage by means of AI, but, through recklessness, when such damage is foreseeable, does not consider the risks and therefore does not take steps to prevent them.

Hallevy proposes the following example: a robot or AI software designed to function as an autopilot and the AI entity is programmed to protect the mission as part of the mission to fly the plane. If, during the flight, the human pilot activates the autopilot and the program is initialized, but at some point after activating the autopilot, the human pilot sees a storm approaching and tries to abort the mission and return to base. After that, the AI entity considers the human pilot's action a threat to the mission and takes action to eliminate this threat: it can cut off the pilot's air supply or activate the ejection seat. As a result, the human pilot is killed by the actions of the AI entity.[17]

---

[16] PAULA *el al.* 2019, 112.

[17] "an AI robot or software, which is designed to function as an automatic pilot. The AI entity is programmed to protect the mission as part of the mission of flying the plane. During the flight, the human pilot activates the automatic pilot (which is the AI entity), and the program is initialized. At some point after activation of the automatic pilot, the human pilot sees an approaching storm and tries to abort the mission and return to base. The AI entity deems the human pilot's action as a threat to the mission and takes action in order to eliminate that threat. It might cut off the air supply to the pilot or activate the ejection

In the example provided, the programmer of the artificial intelligence had no intention of killing anyone, and it is likely that he had not even thought of such an outcome. However, the death was the result of the machine's failure to place limits on its potential decisions, which demonstrates the programmer's negligence. Thus, the programmer can be held responsible if the legal damage caused by the AI is a consequence of the structure of the programming, which could potentially cause the event. Thus, to avoid criminal liability, the programmer is required to contain the risks of the AI he or she programmed.

It should be noted that in most penal systems, which adopt the tripartite concept of crime, the solution to the hypothesis is similar. In this case, causality in culpable crime and unconscious guilt are recognized, making it possible for the programmer to be held criminally liable, since the requirement of human conduct is met. In the paradigm case of Tay artificial intelligence, it is worth analyzing the appropriateness of the potential liability of the program's creators.

Part of the scholars points out that negligent crime is the center of gravity of responsibility for the damage caused by artificial intelligence, materializing in the power of punishment aimed at human beings who contributed with their activity to the production of the result operationalized by the machine. Thus:

> "The focus of the discussion of AI criminal law is negligent crime. The problem that individual contributions to liability are becoming increasingly diffuse, given the networking of AI systems or their connection to human-machine hybrids, is not unknown here either, but can - according to assessments of criminal law literature - be addressed using the principles of operationalizing

seat, etc. As a result, the human pilot is killed by the AI entity's actions." HALLEVY 2010,15-16.

> the demarcation of responsibilities within the framework of the division of proceedings."[18]

3. CONCLUSION: PROBLEMATIZING THE POSSIBILITY OF DIRECT ARTIFICIAL INTELLIGENCE CRIMINAL LIABILITY

The third model proposed, *Direct Liability,* focuses on the criminal liability of the machine itself and makes it possible to consider it as the active subject of the crime, to hold it responsible and to impose penalties[19] in view of the result of the damage to the legal assets under criminal protection.

According to this model, it should be noted that most algorithms allow AI to separate what is permitted from what is prohibited, so, based on the elements of the crime in the common law criminal system, namely the *actius reus* (which represents the external production of harm) and the *mens rea* (which is the subjective element that supposes cognition and choice by desideratum). The criminal liability of an AI entity does not replace the criminal liability of developers or users, if criminal liability is imposed on developers and/or users by any other legal means. The criminal liability of

---

[18] Free translation of: "a Fahrlässigkeitsdelikte, auf die sich der Fokus der strafrecrechtlichen KI-Diskussion richtet. Die Problematik, dass individuelle Verantwortungsbeiträge angesichts der Vernetzung von KI-Systemen bzw. ihrer Verbindung zu Mensch-aschine-Hybriden immer diffuser werden, ist auch hier keine unbekannte, sondern lässt sich - nach Einschätzungen in der strafrechtlichen Literatur - unter Rückgriff auf die Grundsätze der Verantwortungsabgrenzung im Rahmen arbeitsteiliger Prozesse bewältigen." RADEMACHER 2020, 52.

[19] "Drawing a parallel with criminal sanctions in the legal system, it would be possible to: (a) temporarily disable AI; (b) delimit its fields of action; (c) determine the social use of AI; (d) compulsory work on a certain task; or even (e) disconnect the technology." PAULA *el al.* 2019, 112.

the AI entity is imposed in addition to the criminal liability of the human programmer or user.[20]

However, despite this model fit in common law systems, in most Western countries, which have adopted the institutions built on the principle of legality, with the consequent current stage of the tripartite concept of crime and its requirements, they do not incriminate the choices of artificial intelligence, even if they cause damage to legal property. Therefore, in the paradigm case, Tay could not, at the current stage of legislation and criminal science, have criminal responsibility.

The basis of this concept, from its foundation to the present day, requires that a modification of the outside world be produced that can be dominated or controlled by the will. This is not the case with the results of criminally relevant damage, the product of algorithmic self-referential choices, since although they can control processes that potentially modify external reality, especially if the AI commands robots linked to it. The last institution of crime, namely culpability, which is defined as the judgment of disapproval made about the perpetrator of the offence (typical and unlawful conduct), because being able to behave in accordance with the Law, he chose to behave contrary to the Law, is the main obstacle to this accountability.

[20] "When an AI entity establishes all elements of a specific offense, both external and internal, there is no reason to prevent imposition of criminal liability upon it for that offense. The criminal liability of an AI entity does not replace the criminal liability of the programmers or the users, if criminal liability is imposed on the programmers and/or users by any other legal path. Criminal liability is not to be divided, but rather, added. The criminal liability of the AI entity is imposed in addition to the criminal liability of the human programmer or user." HALLEVY 2010, 29.

REFERENCES

BRANDÃO, Claudio. Teoria Jurídica do Crime, Dplacido, 2020.

BRANDÃO, Claudio. Tipicidade penal, Almedina, 2012.

BURKOV, Andriy, The Hundred-Page Machine Learning Book, edited by the author, 2019

GARCIA, Ana Bichara. Ética e inteligência artificial, Computação Brasil, 2020.

HALLEVY, Gabriel, The Criminal Liability of Artificial Intelligence Entities, Akron Intellectual Property Journal, 2010.

KAUFMAN, Dora *et al.* , Revista FAMECO, 27, 2020.

NEFF, Gina; NAGY, Peter, Talking to boots: symbiotic agency and the case of Tay, International Journal of communications, 10, 2016.

RADEMACHER, Timo, Künstliche Intelligenz und neue Verantwortungsarchitektur, Nomos, 2020.