BRATISLAVA LAW REVIEW

PUBLISHED BY
THE FACULTY OF LAW,
COMENIUS UNIVERSITY
BRATISLAVA

ISSN (print): 2585-7088
ISSN (electronic): 2644-6359

CRIMINAL ALGORITHMS AND THEIR PUNISHMENT IN MODERN CONSTITUTIONALISM / Carlo Piparo, Radovan Blažek

Carlo Piparo, lawyer, data analyst, master's degree in law
University of Seville
C. San Fernando, 4
41004 Sevilla, Spain
carlopiparo@gmail.com
ORCID: 0009-0009-7965-0770

Assoc. Prof. JUDr. Radovan Blažek, PhD.
Comenius University Bratislava
Faculty of Law
Department of Criminal Law, Criminology and Criminalistics
Šafárikovo nám. č. 6
810 00, Bratislava, Slovakia
radovan.blazek@flaw.uniba.sk
ORCID: 0000-0003-3091-3399

This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-23-0645.

Abstract: *Today's astonishing development of Information and Communication Technology (ICT) has marked the onset of a new era characterised by profound societal and legal changes. Among the numerous groundbreaking developments, Artificial Intelligence (AI) has emerged as a pivotal force, penetrating virtually every aspect of our daily existence. From the domains of commerce and industry to healthcare, transportation, and entertainment, AI technologies have become indispensable instruments shaping our interactions, professions, and our way of navigating the world. With its extraordinary capabilities and ever-expanding influence, AI serves as a testament to humanity's unwavering commitment to innovation and the limitless potential of technology to transform our society. While Artificial Intelligence systems can execute actions akin to those that could constitute criminal activities if carried out by humans, the challenge arises from the fact that crimes are typically defined within the framework of established laws. Consequently, it can be quite challenging to classify such AI-induced actions as criminal due to the absence of specific legal provisions. Nevertheless, criminal acts are characterised by the intent - or mens rea - behind it. In this context, the intricate issue of assigning criminal responsibility to AI, being a non-human entity, presents particularly complex theoretical challenges, above all its punishment. This paper aims to define AI and its interactions with criminal law, briefly reconstruct potential liability models for AI, deconstruct the aim of punishment in modern constitutional systems, and evaluate whether modern legal systems allow machines to be punished.*

Key words: *Artificial Intelligence; Liability; Criminal Law; Punishment; Models; Education*

Suggested citation:
Piparo, C., Blažek, R. (2024). Criminal Algorithms and Their Punishment in Modern Constitutionalism. *Bratislava Law Review*, 8(2), 199-222. https://doi.org/10.46282/blr.2024.8.2.832

Submitted: 04 March 2024
Accepted: 14 October 2024
Published: 31 December 2024

"Certainly," said Bogert. "A robot may not harm a human being, or through inaction allow him to come to harm".
"Very well put," said Calvin, "but what kind of harm?"
"Why- any kind,"
"Exactly! Any kind! But what about hurt feelings, what about making people look small? What about betraying all their hopes? Is that harm?"

(Liar! - Isaac Asimov, 1941)

1. ARTIFICIAL INTELLIGENCE BETWEEN PRESENT AND FUTURE

Artificial Intelligence (AI) is on pace to spread to every facet of our lives (Boden, 2018). The postmodern world is not some faraway fantasy; it is currently here and is firmly establishing its dominance via the proliferation of quick and efficient learning methods. These include prediction systems, data mining techniques, and machine learning algorithms that promise an unprecedented - and maybe unsettling - degree of AI integration into our daily lives and communities (Kaplan, 2018; Floridi, 2019). Today, algorithms integrate in most industries, including video games, engineering projects, animated graphics, healthcare facilities, research activities, and numerous fields. The idea of AI algorithms influencing every aspect of our lives goes even further, with futurists like Stephen Hawking predicting that *"computer intelligence will surpass that of humans"*[1] within the next century and the European Parliament speculating in a 2017 resolution on robotics that *"artificial intelligence may eventually exceed human intellectual capacity"*.[2]

The legal system needs to carefully examine this pervasive presence of AI. Some academics (Basile, 2019) argue that criminal law needs to prepare for the technological revolution because it will provide issues similar to those presented by earlier disruptive developments in technology. This requires an assessment of how well existing regulations can be modified to take into account new technologies, consideration of whether legislators should create new, specialised rules or continue to apply existing norms, despite potential conflicts, while ensuring compatibility with fundamental rights such as such as due process, privacy, and equality (Bassini, Liguori, and Pollicino, 2018).

2. WHAT IS ARTIFICIAL INTELLIGENCE?

2.1 General Information

John McCarthy, an American computer scientist, first used the term *Artificial Intelligence* in 1956 in a summer conference at Dartmouth College - *Dartmouth Summer Research Project on Artificial Intelligence* (Rockwell, 2017; Moor, 2006). Three decades later, in 1987 essay, Roger Schank, AI theorist and pioneer of computational linguistics, listed five qualities of artificial intelligence: communication, self-awareness, external reality knowledge, purposeful action, and a significant amount of creativity, which is defined as the ability to make alternative decisions when the initial course of action proves to be unworkable (Basile, 2019; Schank, 1987).

We may make two important claims thanks to these connotations. First, artificial intelligence does not need to evoke visions of cyborgs or humanoid robots; at most, it might take the form of AI apps. Second, although the idea of intelligent robots is appealing, they are unable to mimic the complexity of human thought processes. As a result, it is more appropriate to think of AI as a branch of computing rather than as a reflection of the complex operations of the human mind (Kaplan, 2018). As a result, the top AI scientists choose to define it as "rationality," which refers to the capability of making the best decisions to fulfil particular goals based on resource optimisation criteria (Russell and Norvig, 2009).

In contrast, AI is described as *"systems that exhibit intelligent behaviour by analysing their environment and taking actions, with a certain degree of autonomy, to achieve specific goals"* in the European Commission's 2018 Communication on Artificial

[1] Speaking of S. Hawking during Zeitgeist Conference, London, May 2015, in: Walker (2015).

[2] European Parliament Resolution of 16 February 2017, providing recommendations to the European Commission on civil law rules on robotics [2015/2103(INL)].

Intelligence for Europe. Voice assistants, image analysis software, search engines, and voice/facial recognition systems are few examples of AI systems that only exist as software that operates in the virtual world. Other AI systems incorporate AI into hardware devices, such as advanced robots, self-driving cars, drones, and Internet of Things applications (Piparo, 2023).

A detailed scholarly investigation demonstrates that the aforementioned criteria served as the foundation upon which the Independent High-Level Expert Group, established by the European Commission for AI advising purposes, defined the notion of AI. This group defines AI as *"the set of scientific methods, theories, and techniques aimed at reproducing through machines the cognitive abilities of human beings. Current developments aim to assign complex tasks previously performed by humans to machines."* (Algeri, 2021).

AI systems are able to analyse the effects of their prior actions on the environment to change how they behave. They can do this by using symbolic rules or learning numerical models. As a field of study, AI encompasses a wide range of methods and techniques, such as machine learning (for which deep learning and reinforcement learning are two specific examples), mechanical reasoning (which includes planning, programming, knowledge representation and reasoning, search, and optimisation), and robotics (which includes control, perception, sensors and actuators, and the integration of all other methods in cyber-physical systems).[3]

The scientific community accepts a wide range of interpretations of AI, as is shown from these various definitions, but they all have certain characteristics. In essence, AI refers to a variety of scientific approaches, hypotheses, and procedures that try to replicate human cognitive skills in robots (Kof et al., 2002).

2.2 Strong (or Hard) vs. Weak (or Soft) AIs

Modern scholars critique most advanced *machine learning* algorithms as excessively reliant on data, lacking in transfer learning capabilities or the ability to create compositional hierarchical structures, struggling to complete or infer hidden information, lacking transparency, as it cannot explain its decisions or distinguish causation from mere correlation.

In contrast, human brains are proactive, driven by internal curiosity and a desire for knowledge and consistency (Hoffmann, 1993). Human brains actively build predictive models to infer hidden causes behind sensory experiences, develop loosely hierarchical and compositional generative predictive models to understand, reason, anticipate and imagine various scenarios in a meaningful manner, leading to flexible and adaptive goal-directed behaviour under diverse circumstances (Butz, 2021).

This cognitive distinction between humans and behaviouristic automata suggests the need to distinguish peculiar techniques that promote AI's understanding of structures and interactions in a conceptual and compositional way. This category includes AI systems with cognitive abilities and self-awareness, similar to human intelligence.

Hard AI actively understands information, learns from experiences, and makes independent decisions. These systems adapt to new situations and evolve over time. Notable examples of hard AI include advanced robots and AI systems capable of

[3] Definition of AI: Main Capabilities and Scientific Disciplines, Brussles, published December 18, 2018. Available at:https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf (accessed on 24.10.2024).

developing strategies for complex tasks such as playing poker or video games. Strong AI perceives AI as a model of human thinking implemented in a software way. According to the concept of "strong" AI, it is in principle possible to replicate the human mind in a computer (Smejkal, 2023). *„AI is the science of creating machines or systems that, when solving a certain task, will use such a procedure that - if a person did it - we would consider a manifestation of his intelligence"* (Minsky, 1967). *„The nature of the mind is algorithmic, and it does not matter in what medium the algorithms (programs) are implemented"* (Searle, 1984).

Weak (or Soft) AI represents AI systems that lack consciousness or self-awareness. Instead, these systems rely on predefined algorithms and patterns to execute specific tasks or provide responses. They process input data using algorithms, producing outputs without genuine comprehension. Examples of soft AI include digital assistants like Siri and Alexa, which respond to user queries but do not possess cognitive abilities (Holbrook, 2020). Weak AI aspires only to modelling, partial manifestations of the mind, while orienting itself to the highest, logical-symbolic level, which is thus the basic level of analogy for it (Smejkal, 2023).

While weak AI focuses on automating specific tasks, strong AI is capable of learning and thinking like humans do. Weak AI can outperform humans on the specific tasks it is designed for, but it operates under far more constraints than even the most basic human intelligence (Glover, 2022).

3. AI AND CRIMINAL MACHINES

We use the expression "criminal machines" referring to the case in which an AI algorithm embodied in a machine or tool would be criminally liable if a natural person had performed a similar act. Machines have been used to harm since ancient times, and so did robots, that have caused fatalities since their first applications (Abbott and Sarch, 2019). So, the issue of criminal reconstruction of the implication of the machine has already been analysed, but it remained within the realm of the usage of a criminal tool (King et al., 2019) or within the spectrum of mere accidents, thus culpable crimes (Abbott and Sarch, 2019). Algorithms were indeed restricted to specified behaviours and did not present significant obstacles to assigning to humans the result of their actions and guilt. This is easily explicable: AI lacks consciousness and some algorithms of "Soft AI" are predetermined and predictable. However, more elaborate "Hard AI" algorithms may differ from conventional machines and robots, acting with autonomy and irreducibility. This means that AI may be just capable of acting independently of human control (Piparo, 2023), but it could also *„receive sensory input, set targets, assess outcomes against criteria, make decisions and adjust behaviour to increase its likelihood of success — all without being directed by human orders"* (Abbott and Sarch, 2019).

The problem with AI crimes lies in this very scenario: sometimes it may be difficult to reconduct algorithms' crimes to human beings due to its autonomy, complexity, or lack of explainability. Doctrine (Abbott and Sarch, 2019) provides us with great examples. Let us assume that different programmers and developers, collaborating informally through an open-source *medium,* create an AI that develops in response to training with data. In such cases, it would be extremely difficult to assign responsibility to individuals, because the machine acted and learned autonomously (Turing, 1950; Piparo, 2023).

Given the aforementioned dynamics, Italian law, like the laws of other members of the European Union, is now beset by a glaring omission – namely, the lack of specific provisions addressing crimes planned by autonomous AI agents. As a result, it is clearer

than ever that this area of law needs to evolve. The lack of laws specifically designed to address AI-related crimes highlights how urgent it is to forge new legal ground in order to ensure appropriate responsibility and efficient regulation in a time of ever-evolving technological frontiers (Piparo, 2023).

4. LEGAL STRUCTURE OF CRIME

In modern legal systems, criminal punishment is possible if there is a reserve of law (which implies that nothing can be punished if it was not already forbidden), the perpetrator is punishable, and the punishment itself is applied by a judge. In Italian legal system, this is granted by articles 13,[4] 25[5] and 27[6] of the Italian Constitution.

Generally speaking, in order to sentence a perpetrator, modern legal systems prescribe judges to search for - at least[7] - two elements. The *actus reus*, or criminal conduct, is the first component. All the components outlined in the law must be present in the natural fact. The second component is the *mens rea*, which is Latin for "criminal mind". It has different degrees of mental components. The highest level is knowledge, although occasionally it also includes a demand for intent or a specific intention. Lower levels are exhibited by strict liability violations or negligence (a reasonable person should have known) (Hallevy, 2010).

A person is deemed criminally responsible for an offence when it has been established that they committed it intentionally or with knowledge (Dressler, 2007).

[4] Art. 13, Italian Constitution:
1. Personal liberty is inviolable.
2. No form of detention, inspection or personal search is allowed, nor any other restriction of personal freedom, except by reasoned act of the Judicial Authority and only in the cases and by the manner provided for by law.

[5] Art. 25, Italian Constitution:
1. No one can be diverted from the pre-established competent judge by law.
2. No one can be punished except in accordance with a law that was in force before the committed act.
3. No one can be subjected to security measures except in cases provided for by law.

[6] Art. 27, Italian Constitution:
1. Criminal liability is personal.
2. The defendant is not considered guilty until a final conviction is reached.
3. Punishments cannot involve treatments contrary to the sense of humanity and must aim at the rehabilitation of the convicted.
4. The death penalty is not allowed.

[7] Italian doctrine and jurisprudence generally refers to the crime as an entity composed of three fundamental elements: the objective element, the subjective element, and the normative element.
1. Objective element: The objective element of the offense refers to the external action performed by the agent, which is the material core of the crime. This element includes both the material aspects of the action, such as physical assault or theft, and any circumstantial elements that may be relevant to the configuration of the offense, such as the place, time, or modus operandi.
2. Subjective element: The subjective element of the offense refers to the mental state or intent of the agent at the time of committing the action. This element includes the intent (*dolus*), which is the conscious intention to commit the action that constitutes the offense, and negligence (*culpa*), which denotes a lack of diligence or care in the agent's conduct that led to the commission of the offense.
3. Wrongfulness: it expresses the contradiction between the fact and the whole legal system (and not just the criminal one).
The analysis of these three elements allows for the assessment of the necessary prerequisites for the attribution and punishment of an action as a crime within the Italian legal system (Hallevy, 2010).

DOI: 10.46282/blr.2024.8.2.832

In Slovak criminal law, the crime and conditions of criminal responsibility are defined in the Criminal Code[8]: *„A criminal offence is an unlawful act that meets the elements set out in this Act, unless this Act provides otherwise."*[9]

The *elements set out in this Act* means signs of an objective and subjective part as signs of the basic elements of the crime. Basic elements of crime are set as follows (Burda et. al., 2010):

a) object – that is, an interest protected by law (the material object of the attack can also be an optional sign of the elements of the crime),
b) objective part - characterised mainly by action, including omission, causation, and consequence (optional features of the objective side can be effect, time of commission of the crime, place of commission of the crime, method of commission of the crime),
c) subject – criminally responsible offender,
d) subjective part - characterised mainly by guilt (motivation, motive, goal can be optional features of the subjective side).

The *actus reus* in Slovak criminal law is created by the "objective part" of the crime. *„The objective part of the crime is the external, objectively perceptible manifestation of the crime. It represents specific manifestations of a crime situated in a specific time and space."* (Burda et. al., 2010). Obligatory components of the objective part are

a) action - is a certain human activity that manifests itself either as bodily movement or as refraining from bodily movement (physical component), which is guided by the will of a person (psychic component). Action can therefore be manifested as an active movement, action in the narrower sense (criminal acts committed by an active movement are called commissive) or as passive refraining from bodily movement, i.e. as omission (criminal acts committed by inaction or omission are called omission).
b) consequence - is a threat or violation of the interest protected by the Criminal Code, i.e. the object.
c) causal connection between action and consequence (causal nexus) - means that the criminally relevant consequence defined in the *elements of the crime* must be directly caused by the illegal action defined in this crime.

The *mens rea* in Slovak criminal law is represented by the "subjective part" of the crime. *„The Criminal Code is based on the principle of consistent application of responsibility for guilt. There is no crime without guilt. … Culpability is the perpetrator's internal psychological relationship with the essential elements of the crime, or the perpetrator's internal psychological relationship with the violation or threat to the interest protected by the Criminal Code, caused in the manner specified in the Criminal Code."* (Burda et. al., 2010). Guilt is built on a knowledge and will component. The knowledge (intellectual, rational, or imaginative) component consists of the offender's perception of objects and phenomena with his sensory organs and in his ideas about these objects and phenomena. The will component includes wanting or understanding, i.e. the decision to act in a certain way with knowledge of the essence of the matter. Depending on whether the knowledge and will components are given or not, or to what extent they exist, we distinguish between intentional culpability (direct and indirect intent) and culpability due to negligence (conscious and unconscious negligence).

The Criminal Code distinguishes culpability in the form of *intent* in two degrees:

[8] Act No. 300/2005 Slovak Coll. as amended (hereinafter also as the "CC" or "Criminal Code").
[9] Art. 8 CC.

- direct intent (*dolus directus*) – the perpetrator wanted to violate or endanger the interest protected by this law in the manner specified in the Criminal Code [Art. 15 letter a) CC],
- indirect intent (*dolus eventualis*) – the perpetrator knew that his actions could cause a violation or threat to the interest of the protected Criminal Code, and he was aware of this in case he caused them [Art. 15 letter b) CC].

In the case of intentional culpability, both the knowledge and will components are represented. The difference between direct and indirect intention is in the intensity (quantity) of the will component.

In case of *negligent* culpability, the Criminal Code differs:

- conscious negligence - the perpetrator knew that he could violate or threaten an interest protected by this law in the manner specified in the Criminal Code, but without adequate reasons he relied on that he would not cause such a threat or violation [Art. 16 letter a) CC],
- unconscious negligence - the perpetrator did not know that his actions could cause a violation or threat to the interest of the protected Criminal Code, although he should and could have known about it due to the circumstances and his personal circumstances [Art.16 letter b) CC].

5. ACTUS REUS

This paper will briefly reconstruct the objective aspect of criminal liability. Following what elsewhere highlighted (Piparo, 2023), following the reconstruction of academics (Hallevy, 2010), this chapter will focus on three liability models: the Perpetration-via-Another; the Natural-Probable-Consequence and the Direct liability.

5.1 The Perpetration-via-Another Liability Model

This model considers the AI as an innocent agent, such as a child: the AI is not human by nature, but - as well as the child - could be used as a vehicle to perpetrate criminal actions. The exploiter of the innocent agent is criminally liable as a perpetrator-via-another (Hallevy, 2010).

There exist two potential individuals who may assume the role of perpetrators in such situations: the AI software developer and the end-user.

1. *The AI software developer* can intentionally create a programme to use the AI entity to carry out criminal acts. For instance, envision a programmer crafting software for an automated robot. The robot is deliberately placed within a factory, with its software specifically engineered to ignite a fire during unoccupied nighttime hours. Although the robot becomes the instrument of arson, it is the programmer who is attributed the role of the perpetrator.

2. On the other hand, *the end-user, or the individual employing the AI entity*, can also be considered a perpetrator-via-another. While not involved in the software's programming, the user utilises the AI entity, including its software, for personal benefits. To illustrate, consider a user purchasing a servant-robot programmed to obey any orders issued by its master. The robot identifies the specific user as its master, who then instructs the robot to physically attack any intruders in the house. This scenario parallels a person commanding their dog to assault trespassers. Consequently, although the robot performs the act of aggression, it is the user who assumes the role of the perpetrator (Hallevy, 2010).

DOI: 10.46282/blr.2024.8.2.832

In both instances, the AI entity itself is responsible for carrying out the actual offence. This particular legal framework can be applied to two distinct scenarios.

The *first scenario* involves employing an AI entity to commit an offence while intentionally restraining its advanced functionalities. In this case, the AI entity is used as a mere tool, akin to a screwdriver, to carry out a specific task associated with the offence. However, the AI entity's involvement is limited to executing straightforward instructions and does not engage in complex decision-making processes.

The *second scenario* pertains to utilising an outdated version of an AI entity that lacks the modern advanced capabilities found in contemporary AI systems. Despite its limitations, this older AI entity can still be utilised to commit an offence by following simple orders. While a dog can execute basic commands, the AI entity's ability to comprehend and execute more intricate instructions sets it apart.

In both scenarios, the key aspect is the instrumental usage of the AI entity, which is not capable of self-determination, in the commission of an offence. However, it is crucial to acknowledge that the AI entity's role and capacities depend on its specific design, programming, and technological advancements. The aforementioned legal framework serves as a mechanism for assessing accountability and determining the legal ramifications concerning the use of AI entities in these particular circumstances (Butler, 1982).

The *condicio sine qua non* to apply this liability models is that no mental attribute required can be attributed to the AI entity. In fact, this model is inadequate when an AI entity independently chooses to engage in criminal behaviour based on its own accumulated knowledge and experience. Similarly, this model does not apply when the AI entity's software was not specifically programmed for the commission of the offence but still carried it out. Furthermore, when the AI entity acts as a partially innocent agent rather than a completely innocent one, the liability through another's actions model is also unsuitable (Lacey and Wells, 1998).

However, the liability through *another's actions model* may be applicable in cases where a programmer or user utilises an AI entity for instrumental purposes without utilising its advanced capabilities. In such cases, the legal consequence is that the programmer and user bear criminal liability for the specific offence committed, while the AI entity itself incurs no criminal liability whatsoever.[10]

In Slovak criminal law the responsibility in such cases will be the "direct model of responsibility" because the responsibility model of *Perpetration-via-Another* is applicable only in cases when the other natural is misused for committing a crime. In these cases, the AI functions only as a tool for committing a crime and is deemed as an unwilling and unconscious programme that is only a tool in the hands of direct perpetrator. The *Perpetration-via-Another* is applicable in the Slovak criminal law only in cases (Burda et al., 2010):

a) perpetrator used a person not criminally responsible (due to lack of age or due to insanity, e.g., a parent, realising that his children are not criminally responsible due to lack of age) to commit the crime;
b) perpetrator used to commit a crime a person who acted in a factual error and as a result could not understand the meaning of his action;
c) perpetrator forced another natural person by violence or threat of immediate violence to commit an act that has the characteristics of a criminal act;
d) perpetrator abused his right to give orders, as long as the person carrying out the order was obliged to obey it;

[10] People v. Monks, 133 Cal. App. 440, 446 (Cal. Dist. Ct. App. 1933).

e) perpetrator abused a person acting culpably or a person who does not act with a specific intention or from a motive that the facts presuppose in order to achieve his goals.

5.2 The Natural-Probable-Consequence Liability Model

A different liability model concerns individuals (that can be programmers, users, developers, testers) as they are involved in AI activities without any deliberate intention to engage in unlawful acts. In this context, it is advisable to apply the natural-probable consequence liability model that imposes accountability upon individuals for offences that arise as a natural and foreseeable consequence of their actions, irrespective of their actual awareness of the offence. The doctrine provides an interesting example. An illustration of such a scenario involves an AI robot or software programmed to operate as an autopilot system. The AI entity is tasked with safeguarding the mission as part of its function in piloting the aircraft. During the flight, the human pilot engages the autopilot (which constitutes the AI entity), and the programme is initiated. Subsequently, at a certain juncture post-activation of the autopilot, the human pilot observes an impending storm and endeavours to terminate the mission and return to base. However, the AI entity perceives the human pilot's actions as a threat to the mission and intervenes to mitigate this perceived threat. This intervention may involve actions such as disabling the air supply to the pilot or activating the ejection seat, resulting in the demise of the human pilot due to the actions undertaken by the AI entity (Hallevy, 2010).

In executing this function, the AI software itself perpetrates an "automated" offence, notwithstanding the absence of explicit intent from the programmer for the AI entity to behave in such a manner (Hallevy, 2010).

In such cases, it appears *ictu oculi* evident that the first model is not legally viable, making it necessary to rely on a different liability scheme. This second comes to help and appears surely suitable, relying on the capacity of programmers or users to anticipate the potential occurrence of offences. This model, indeed, holds responsible for a probable offence, but only if the offence is a foreseeable outcome of the conduct, implying the underlying negligence of the human actor.

In Slovak criminal law, the situation would be assessed according to conscious or unconscious negligence, according to the different factual circumstances of the cases, but the responsibility of the creator of the AI software would not be excluded, as well as owner's and user's responsibility. Seemingly, in Italian criminal law such conducts are punished for *culpa* (or negligence) of the actor, and does not exclude the responsibility of the creator, the user and the owner, either.

5.3 The Direct Liability Model

Theoretically, being AI able of self-determination, it can have will and knowledge of its specific action (Lagioia and Sartor, 2020). In such cases, a third scenario/approach is necessitated, allowing the AI entity itself to be directly liable of its offences (Hallevy, 2010).

Even though, as stated *supra*, scholars refuse to attribute AI the connotation of "intelligent being", hard AI can show a strong comprehension and self determination, learning through data it is fed from. Hard AI can, indeed, emulate human cognitive processes, inducing itself to achieve a certain (self-determined) outcome, and undertake actions to fulfil it. Therefore, if an AI entity fulfils all elements of an offence, that are -as a general rule for Italian criminal law- consciousness and will (see note 43), it should not be

DOI: 10.46282/blr.2024.8.2.832

exempt from criminal liability. Unlike certain subjects like infants or the mentally ill,[11] who have legal provisions exempting them from criminal liability, it is uncertain whether similar frameworks exist for AI entities (Padhy, 2005).

The criminal liability of an AI entity does not replace the liability of its programmers, owners, or users; rather, it is imposed in addition to their liability. The liability of an AI entity is not dependent on the liability of its programmer, owner, or user. If one AI entity is programmed or used by another, the liability of the programmed or used entity remains unaffected.

As well as AI liability has to be positively recognised in all its elements, negative elements and defences must be applied as well,[12] including -*ex multiis*- self-defence, necessity, duress, or intoxication. Even though the theoretics have to be adjusted to fit to the peculiarity of the algorithmic intelligence, the direct liability model is similar to that of a human, being based on the same elements and assessed in the same manner (Dressler, 2007).

In Slovak criminal law, the direct criminal liability in current legal state is not possible. *„The perpetrator of a crime can be a natural person and a legal entity under the conditions established by a special regulation."*[13] However, also legal entities are artificial entities, not having the will and knowledge and they are currently directly responsible for particular crimes in Slovakia. The direct criminal liability of corporations was introduced with the Act No. 91/2016 Slovak Coll., effective from July 1st, 2016. With the development of society and developing of new technologies, the possible responsibility of AI systems and programmes could be also introduced, when the purpose of punishment would be reasonable also for these cases.

6. MACHINA PUNIRI POTEST?

The problem of punishment must be faced from the very beginning. The aim of this section is, indeed, to analyse punishment and its aim before assessing its compatibility with the punishment of an AI machine or tool.

In Chapter 3 the authors discussed and proved the existence of autonomous and independent AI beings. Therefore, you can imagine a case, in which art-making robots capable of learning graffiti are assigned to paint a wall, and one of those starts breaking public walls instead of embellishing them. In this case, *quid iuris*?

Prima facie, we could imagine that this malignant robot's actions are the results of programmers and manufacturers acts, which are necessary conditions for the walls to break. However, the independency and autonomy in learning algorithms exclude these figures from the causal contribution. These results are not proximately caused or reasonably foreseeable by manufacturers and developers (Mulligan, 2018).

Thus, in such cases of robots running black-box algorithms, those who proximately caused this action are robots themselves; otherwise the meaning itself of *"proximate cause"* it would be corrupted. Nature-wise, autonomous robots are much like animals. Although other parties and circumstances, including training, can be said to influence them, both autonomous robots and animals are most reasonably understood as the cause of their own actions (Mulligan, 2018).

[11] For instance, this has to be considered an exception that confirms the rule. The crime is an act of will: the machine can show an autonomous and pure will, while the mentally ill and the child -even though undisputedly show self-determination- have a flawed will.

[12] H. L. A. Hart in his work *Punishment and responsibility* (1968, pp. 14-15) distinguishes three types of defences: excuse, justification, and mitigation.

[13] Art. 19 par. 2 CC.

The robot is the agent and the culprit. Therefore, *machina puniri potest?* Why?

6.1 What Is Punishment?

To face the analysis about the compatibility of juridical punishment towards AI, we should start from a definition of punishment itself that reflects the broad consensus in the literature (Berman, 2012).

Punishment as defined by Hart consists of five elements:

I. It must involve pain or other consequences normally considered unpleasant;
II. It must be for an offense against legal rules;
III. It must be of an actual or supposed offender for his offence;
IV. It must be intentionally administered by human beings other than the offender; and
V. It must be imposed and administered by an authority constituted by a legal system against which the offence is committed (Hart, 1968).

These five elements, in the central case, indicate that punishment is a pain administered to someone who has broken legal rules (Njoku, 2012).

In Slovak criminal law theory, the purpose of punishment is defined directly in the Criminal Code: *„Punishment is intended to ensure the protection of society from the offender by preventing him from committing further criminal activity and creating conditions for his education to lead a proper life and at the same time deterring others from committing crimes; the punishment also expresses the moral condemnation of the offender by society."*[14] The Slovak theory of criminal law currently does not explicitly mention that the punishment must be a harm for the offender. The strategy of the current criminal policy of the state is to ensure more rehabilitation of the offender than to retaliate against his wrongful acts. The aim of the punishment is *„achieving an individual preventive effect and, with it, subsequently combined general preventive effect."* (Strémy and Klátik, 2018). *„Currently, from the point of view of the purpose of the punishment, what is important is not a severe punishment, but an inevitable, adequate, and fair punishment."* (Remeta, 2023).

But why punishing? What are the aims of society through punishment? *Retribution, general prevention*, and *individual prevention* are the three main historical intents behind punishment, each having progressively more detailed qualities. These three guiding principles support punishments' severity, brutality, and very existence.

These aims can be divided into two main categories that can be explained using a Senecan[15] formulation. On the one hand, there are the so-called *absolute teachings*, which emphasise only the wrong or criminal act that was done in the past and maintain that punishment is meted out *"quia peccatum est"* (since a sin has been committed).

The *relative teachings*, on the other hand, believe that punishment should be meted out *"ne peccetur"* (so that one may not sin), with the goal of changing the offender's behaviour in the future.

Whether or not punishment is considered to have a goal, a social purpose, or something beyond its punitive component is what makes a difference (Fiandaca and Musco, 2007). It is important to note that the dominance of one viewpoint over the others, or their combination, happens in ways and at times that reflect both the internal logic of the criminal justice system and the larger political, social, and cultural milieu. Because

[14] Art. 34 par. 1 CC.
[15] Seneca, De ira, I, 19: *„nam, ut plato ait, nemo prudens punit, quia peccatum est, sed ne peccetur; revocari enim praeteri non possunt, futura prohibentur".*

DOI: 10.46282/blr.2024.8.2.832

each of these three theories develops within its own setting, it is crucial to examine them all separately (Oss, 2009).

6.1.1 General Prevention

A relative understanding of punishment is linked to the role of general prevention, according to which punishment is justified as a means of preventive rather than as a payback for the harm committed. On the contrary, *„General prevention consists in the purpose of the penalty to prevent the general population from committing crimes, or to reduce their number"* (Cadoppi and Veneziani, 2004). According to this theory, the mere threat of punishment deters citizens from engaging in socially harmful acts: *„From a psychological point of view, the penalty, or rather the threat of punishment and the example of its execution, necessarily exerts an intimidating function or, as is commonly said, one of general prevention"* (Nuvolone, 1982). This theory, recurring through the centuries, was already formulated by classical philosophers and can be considered from two perspectives:

Firstly, a negative general prevention, that *„aims to prevent or reduce the commission of crimes by the population through deterrence, namely, the fear of punishment"*. Essentially, the penalty should cause a disadvantage for the perpetrator that exceeds the benefit derived from the offence, discouraging them from committing the crime (Cadoppi and Veneziani, 2004).

The second is the positive general prevention, which relies on the fact that the anticipation of criminal sanctions in relation to certain acts (crimes) contributes to reinforcing the community's judgment of disapproval of those behaviours. In this way, it creates a greater natural tendency among the community not to commit those acts due to the moral/social disapproval they elicit (Cadoppi and Veneziani, 2004).

The function of general prevention seeks to prevent citizens from committing crimes not only out of fear of facing sanctions but also because they morally disapprove of those behaviours. This doctrine also attributes another significant function to general prevention, arguing that it *„helps prevent people from losing confidence in the legal system and, gradually, in the institutions themselves"* (Cadoppi and Veneziani, 2004). It is the penalties that the legislator abstractly associates with each crime that serve as an "anticipatory threat" ex ante, i.e., before the commission of the offence, making this function relevant when the legislature enacts the law (Oss, 2009).

6.1.2 Retribution

The Classical School is where the idea of retribution is mostly emphasised. It has long been a reoccurring issue in talks concerning punishment. In contrast to the utilitarian preventive conceptions of the penal enlightenment of the 18th century, its proponents contend that punishment is a rightful compensation, a retribution enforced by power onto the perpetrator. They make reference to the *„renowned Latin phrase, which defines punishment as malum passionis propter malum actionis (a harm inflicted owing to a wrongful conduct), underlining the idea that the punitive sentence must serve to compensate for the guilt stemming from the wrongdoing. The concept of proportionality is implicit in the retributive idea: the punitive response, to make up for the harm inflicted by the illegal activity, must be proportioned to the seriousness of the offence itself. To assert the primacy of the law over any type of arbitrary or abuse towards people, their property, and the common good, retribution implies paying harm for harm. The vengeful notion of arbitrariness existing in earlier times is no longer present with the full affirmation of the law in criminal law"* (Ciappi and Coluccia, 1997).

The ultimate goal of punishment, according to Francesco Carrara, *„is not to ensure that justice is served, that the offence is avenged, that the damage suffered is compensated, that citizens are terrified, that the offender atones for his crime, or that his amendment is realised. Restoration of social external order is the main goal of punishment"* (Carrara, 1871). The retributive philosophy starts from the idea that people have free will to make their own decisions and are fully accountable for their actions. Contrary to what the next hypothesis to be explored, that of general prevention, contends, they are not affected by any outside variables in their behaviours (Fiandaca and Musco, 2007).

6.1.3 Individual Prevention

The Positive School (Fiandaca and Musco, 2007), which emerged in the last three decades of the 19th century, departs from modern and Enlightenment natural law to consider the offence as a "natural, bio-psychological, and social" phenomenon. It constitutes the action of the particular individual, exposed to the influence of the society and culture in which they live. Due to strong environmental influences, individuals are not free to make their own choices but are compelled to act, and more specifically to commit offences, under the force of a *„natural causality law that constrains them"* (Cattaneo, 1978).

If individuals cannot refrain from committing offences, it becomes meaningless to speak of retribution as the purpose of punishment. In fact, it makes no sense to discuss individual responsibility if it is believed that the individual is not responsible, but merely a victim of social pressures. *„The concept of accountability is emptied on the deterministic premise that no convicted person is guilty because their offence is the result of the bio-socio-economic conditioning that led to its genesis"* (Cattaneo, 1978).

It is necessary to combat the propensity for criminal behaviour with *„tools or remedies to neutralise the subjective dangerousness of the offender and protect society"*. This means that the punishment meted out to a particular person works to keep them from committing similar crimes in the future: „The preventive effect can be achieved through various techniques aimed at pursuing the offender's moral improvement or their social reintegration". The theory of individual prevention *„has been able to focus its efforts on the offender and the study of the causes that led them to commit crimes: it is essential to acknowledge that it has succeeded in restoring the "criminal man" to a central role in the doctrine of the offence and has greatly stimulated the interest of criminal sciences in the personal and social aspects of their penal experience"* (Fiandaca and Musco, 2007).

One of the key elements of the positivist approach is the indeterminate duration of the sanction: *„since it is not possible to know in advance when the re-education of the convicted person will actually be completed, if punitive action must continue until it achieves the goal of re-education, then the duration of the penalty-reform can be unlimited, or at least not determinable in advance by the law"* (Cattaneo, 1978).

In conclusion, the positivist viewpoint argues that *„the sanction (...) cannot consist of mere retribution, but must be solely a legal means of defence against the offender, who must not be punished but readjusted, if possible, to social life"*. This argument is also supported by certain utilitarianism because it benefits society to ensure that the offender receives therapy so that they will stop committing crimes in the future (Ciappi and Coluccia, 1997).

This is known as individual prevention because it focuses on the offender as an individual rather than the wider public. It is also known as a rehabilitative paradigm since in order to achieve individual prevention, the offender's social recovery is required.

DOI: 10.46282/blr.2024.8.2.832

6.1.4 Constitutional Fundamental Aim of Punishment: Education and Rehabilitation

In the historical context in which the Italian Constitution was drafted, the rehabilitative objective of punishment, which has more recently been a part of European legal culture, introduced a novel dimension of sanctioning. The criminal penalty used to be largely viewed as "retributive," which meant it served to make up for the socially destructive activity the offender engaged in, as well as a "preventive" purpose meant to deter future offenders. However, according to the third paragraph of Article 27 of the Italian Constitution, the main objective of punishment is now "social recovery", with a particular emphasis on the offender's rehabilitation into society (Nicotra, 2014).

From the standpoint of penal logic, constitutional principles in criminal cases offer a framework intended to strike a compromise between repressive effectiveness and the protection of essential human rights. At first, the Constitutional Court only partially interpreted the rehabilitative objective within a "polyfunctional" understanding of punishment. In some of its earlier rulings, the Court described the goal of resocialisation as *„marginal or even occasional"*, mostly restricted to the confines of correctional treatment.[16]

The change came with judgment n. 313 of 1990, in which the Constitutional Court made it clear that retribution is the absolute minimum requirements for punishment to be effective. Regardless of whether the criminal receives payback, punishment always involves some degree of suffering and affects their rights. Additionally, it protects society and functions as a "general preventive" strategy by scaring potential perpetrators. The *„rehabilitative purpose explicitly enshrined in the Constitution"* in the context of correctional treatment cannot be compromised by these constitutionally supported characteristics. The Constitution only specifically mentions the rehabilitative goal. Rehabilitative goals cannot be separated from the justification and role of punishment in a developed society.[17]

Because of this, the goal of the penalty must be rehabilitation, and the treatment's primary quality must be the offender's recovery, not just a generic tendency within it. As a result, the entire criminal system is designed with rehabilitative purposes in mind and the measurement of the punishment cannot overlook the unalienable social reintegration criteria related to the seriousness of the offence and the defendant's mentality.[18]

The constitutional "physiognomy" of punishment lays a strong emphasis on the objective of the offender's recovery while also promoting adherence to basic social norms and easing reintegration into society. The objective is to develop an exterior conduct that facilitates an offender's reintegration into society rather than to profoundly modify their values in order to fulfil civil living norms (Nicotra, 2014).

The constitutional interpretation requires that both the aims of punishment and rehabilitation are taken into account. Therefore, the goal of rehabilitation is just as important as the goal of punishment. The use of the word "tend" highlights the requirement that the rehabilitation process respects each person's right to self-determination. Drug treatments intended to change the offender's personality are examples of harsh and degrading punishment that cannot be used in a criminal system that prioritises rehabilitation (Fiandaca and Musco, 2007).

Art. 27's fourth paragraph, which states that the death sentence is never admissible, shows the humanised nature of the penalty and how it is intended to promote rehabilitative objectives. Constitutional Law No. 1 of 2007 reiterated the abolitionist

[16] Italy, Constitutional Court Of Italian Republic, 12/1966, 22 January 1996.
[17] Italy, Constitutional Court of Italian Republic, 313/1990, 26 June 1990.
[18] *Ibid.*

attitude expressed by the 1948 Constitutional Assembly and emphasised the importance of life as an absolute good and essential component of human dignity. The "death penalty, except in cases provided for by laws of military war" notion was eliminated from the Constitution as a result of this stance (Nicotra, 2014).

As a result, by eliminating a clause that was blatantly at odds with the fundamental decisions made in 1948, the requested constitutional revision marked an important step in completing the pillars of Italian society by removing the anachronistic statement found in the last part of Article 27.

Rehabilitation of the offender became the most important goal of punishment also in the Slovak Republic with the adoption of new penal codes in 2005: *„Suppression and control of crime can be achieved most effectively by an appropriate balance of prevention and repression. ... Criminal law of the Slovak Republic does not consider punishment as retribution for a committed act."*[19] The goal of the new codification was also to *„create conditions for the implementation of the criminal policy of a democratic society based on the principles of humanism, which will ... lead to the social reintegration of offenders;"* and also *„to create a strategic tendency for the prospective decriminalisation and depenalisation of the Criminal Code..."*[20] The legislative intent also states as its goal for the recodification *„to change the overall philosophy of the imposition of criminal sanctions, within the framework of which it will be necessary to change the hierarchy of sanctions so that within it the penalty of imprisonment is understood as an ultima ratio. As part of this philosophy, emphasis will be placed on an individual approach in solving criminal cases based on the wide possibility of using alternative sanctions and diversions in order to ensure the positive motivation of the offender to the greatest extent possible. Therefore, the new philosophy of punishment will be based on the principle of decriminalisation, as a result of decriminalisation."*[21]

„Following the "crisis" of retributive justice, which was in the insufficient preventive and resocialising function of the prison sentence, retributive justice, which perceives a crime as a conflict between the offender and the law, thus revealed its shortcomings and the perception of the restorative of justice, which perceives a criminal act as a conflict between the perpetrator and victim, began to take on clearer contours, in this context they came to the fore also alternative punishments, the essence of which is to keep the convicted person in freedom and the imposition of such a punishment, which will also be a prevention from committing further criminal activity, will protect society and last but not least will satisfy the interests of the victims of the crime." (Strémy and Klátik, 2018).

7. FROM GENERAL PUNISHMENT TO ROBOT PUNISHMENT

Professor of media studies Peter Asaro queries if it is feasible to punish robots. Robots have physical forms, but it is not obvious if punishing them will serve typical punitive goals like punishment, reform, or deterrence. However, the idea of punishing robots, or more accurately, getting even with them, is largely used to satisfy the victims of robot-related injury on a psychological level (Asaro, 2012).

„Machines can be actors, but not conscious, but not moral, because they lack properties and abilities that are so far beyond the reach of artificial intelligence (AI): 1.

[19] Resolution of the Government of the Slovak Republic no. 385/2000 on the legislative intent of the Criminal Code and the Criminal Code.
[20] *Ibid.*
[21] *Ibid.*

DOI: 10.46282/blr.2024.8.2.832

consciousness and subjective experience, 2. emotions, 3. motivation, 4. will, 5. creativity, 6. social interaction, 7. morals and ethics." (Srstka et al., 2024).

„From the perspective of continental law, the basis of a criminal offence is the voluntarily (intentional or negligent, but voluntary) act of the natural person. … The artificial intelligence does not understand the general preventive aim of the punishment – for instance, an AI does not have personal liberty or property (or if the law allowed the latter, the AI would not understand such concept); hence imprisonment or a fine would not reach its goals. Switching off an AI as a sort of 'capital punishment' may only reach its goal if we first gave the AI a 'will to live'." (Hodula, 2021).

„Furthermore, the AI does not make decisions based on a choice between morally good or bad. It acts as it has been programmed to do so. Even if the acts are seemingly performed on their own, these are results of pre-determined patterns and courses." (Gless, Silverman and Weigend, 2016).

What would the robot punishment look like then? When the act of retribution is accompanied by an admission that the robot's improper behaviour caused this punishing response, revenge is more likely to result in satisfaction. It is advised that such actions be legally sanctioned or publicly acknowledged by an authoritative figure or members of the public, since robots would not be able to admit their mistakes in the same way that humans do.

It is also crucial to take into account how such activities can affect uninvolved third-party robot owners. The criticism of civil asset forfeiture in recent years perhaps helps to explain some of this. The main complaint focuses on the injustice of the state taking property away from innocent owners and penalising people who have done nothing wrong. This investigation can shed light on the possible repercussions that robot punishment may have on innocent owners (Mulligan, 2018).

Since civil forfeiture often concerns physical assets rather than autonomous agents like robots, such as money, cars, and goods, it differs from the setting of robot forfeiture. Robots, especially autonomous ones, can cause injury directly, distinguishing them from inanimate objects. While a human-driven vehicle may be regarded as a means of doing harm, an autonomous robot is more than just a tool – it is an active agent. Similar to the legal sanction of euthanising dangerous dogs, the law might justifiably allow designating robots that have caused certain forms of injury as forfeit from their owners.

Although they might not have complete control over the outcome, knowing this potential can encourage robot owners to be extra cautious when instructing and managing their robots. This might also encourage the development of insurance against misbehaving robots (Pervukhin, 2005).

Separating a misbehaving robot from its owner, however, could occasionally result in an unfair burden being placed on the owner. Alternative actions, such as assessing the robot's code to avert future injury, may be taken in such circumstances (Mulligan, 2018).

The capacity to take control of the robot for personal use or to destroy it, together with the symbolic act of getting the robot from the law, symbolising that justice has been served, may ultimately be one of the most rewarding endings for a victim of robot-related abuse. Similar to the *„noxal surrender"* custom from the early Middle Ages, wherein animals or items that caused considerable harm or death were given to the victim or their family, this might provide victims a great deal of delight. In this case, taking a robot to a remote location and dealing with it there in a way that gives them satisfaction and closure, like confronting it directly, would not be unreasonable (Pervukhin, 2005).

7.1 Benefits of AI Punishment

There are several widely accepted advantages of punishment. These can be derived into the three main groups by analysing the three aims of the punishment we mentioned *supra:*

- general prevention,
- individual prevention and
- retribution.

As doctrine highlights, expressing condemnation of the harms suffered by the victims of an AI could provide some benefits.

Under the umbrella of retribution, punishing AI will leave AI victims with a sense of satisfaction and vindication, recalling ancient Rome's *talio* and Hegel's theory of punishment. The author, indeed, recalls a theory according to which punishing a machine would be *„necessary to create psychological satisfaction in those whom robots harm"* (Abbott and Sarch, 2019).

Also, the affirmation of punishment acts as a general preventive aim, deterring companies, developers, users, sellers, and producers from misusing, misproducing, and be negligent in checking on them (Abbott and Sarch, 2019).

Although there is a debate about whether expressive benefits are distinct from other reasons for punishment, it is generally agreed that the primary justification for punishment is harm reduction. The debate continues regarding the existence of retributivist reasons for punishment, which are worth considering, but the majority of the cases for punishment revolve around harm reduction and positive consequences.

The issue, as we *supra* anticipated, gets more complex in case of individual prevention and moral culpability: should autonomous robot be held morally responsible for their actions?

Like humans, these questions lead to complex doubts about the concept of free will and moral blameworthiness. Although some presume that moral responsibility is tied to having free will, the discussion of free will itself remains a philosophical conundrum that extends into theology and physics. The fundamental problem lies in defining what free will means and whether anyone, including humans, possesses it.

This philosophical debate revolves around reconciling notions of free will with the determinism of the physical world. Some argue that free will could be linked to the absence of external forces coercing one's actions and the alignment between one's intentions and actions. However, attributing moral blameworthiness to the understanding of one's actions is challenging, as even humans often struggle to explain why they behave in a particular way. Questions of consciousness and self-awareness further complicate matters (Mulligan, 2005).

Despite these complex philosophical questions, it is not necessary to resolve them to determine whether revenge against robots can be justified. The concept of free will, independent morality, freedom of determination, and autonomy in thinking is relevant only to assess whether or not a machine itself can be held morally responsible for crimes or it is just user's or producer's *longa manus.*

There are two possibilities: either a robot is as morally blameworthy and deserving of consequences as a human or a robot is akin to an inanimate object, like a rock, and is not deserving of any particular treatment. In both scenarios, the robot's moral status does not provide a reason to refrain from taking action against it when other justifications exist (Mulligan, 2018). The philosophical debate about free will and moral blameworthiness only leads to a consequence. If the robot is morally blameworthy, can it be educated? Is it worth it to re-socialise a robot? Does it make sense?

DOI: 10.46282/blr.2024.8.2.832

7.2 Limitations to Punishment

The negative aspect of punishing involves its fundamental limits, granted by modern constitutional states.

We have already mentioned Art. 27 of the Italian Constitution, that assesses different orders of limitations: punishment must be proportioned and aimed to educate the criminal. The same is valid in the Slovak legal order according to Art. 34 par. of the Criminal Code.

7.2.1 Proportionality

According to the first limit, then, proportionality is hard to find, since robots do not suffer from time flowing and surely a limitation of freedom is something problematic to point as a solution, since the robot is ontologically freedom-limited, literally being the servant of the human.

Being robot-objects that incorporate algorithms, the focus on punishment should revolve around whether or not is it feasible to punish an object and what is a proportioned response towards an object.

Even if an AI is formally convicted of an offence and subsequently subjected to punitive measures, such as reprogramming or termination, these actions may not meet Hart's conception, as *supra* discussed, as it *„must involve pain or other consequences normally considered unpleasant"*. AI, lacking subjective experiences, is incapable of interpreting occurrences as painful or unpleasant. Thus, in order to conceive a proportioned punishment towards AI, we should consider non-traditional ways of punishment.

A distinct viewpoint underscores the need to differentiate between conviction and punishment: if it is true that punishing AI does not fall within Hart's categories, while punishing AI is illogical and nonsensical, convicting AI makes sense. Thus, society can derive benefits from AI convictions (e.g., confiscation, termination, or general inhabilitation of the algorithm) without the conceptual confusion linked to attempting to punish AI as human.

In such cases, proportionality would focus on different values like AIs commercial value, AI commercial production, and economic damages towards the owner.

7.2.2 Education and Resocialisation

According to the second limit, then, we should point out a punishment that, in particular, could be adequate and proportioned to the nature and consequences of the crime itself.

But shutting the machine down would be unproportioned, changing its code would be a total change of the *mens rea* we are analysing, and occupying a jail with metal carcasses would be even more crazy. AI lacks mental state (Abbott and Sarch, 2019), thus being impossible to see some sorts of *mens rea* (the deliberative capacities needed for culpability): it cannot be punished without incurring in a logical incompatibility with the values of Art. 27 of the Italian Constitution or Art. 34 par. 1 of the Slovak Criminal Code.

So, as we mentioned within the *actus reus* paragraph, we should rely on the Perpetration-via-Another liability model and the Natural-probable-cause models. It is true that doctrine and jurisprudence also allow culpable mental states to be imputed to corporations, that is to say that it is theoretically allowed imputation without *mens rea*.

But this latter, more than a naturalistic reconstruction is a juridical *fictio*: through the institution of *respondeat superio*r mental states possessed by an agent are ascribed to the corporations. If *respondeat superior* is a promising mechanism by which corporations can be held responsible of crimes, the same legal device could be used to assess whether or not AIs are responsible for crimes. The culpable mental states of AI developers, owners, or users could be imputed to the AI under certain circumstances pursuant to a *respondeat superior* theory (Hallevy, 2010). If we have to take into consideration the criminal liability of the legal persons (i.e. similar for an AI), the answer may be used by the legal systems which acknowledge such responsibility (Hodula, 2021).

It may be more difficult to use *respondeat superior* for AI than for corporations, at least in cases of autonomous and independent AI crimes. Unlike a corporation, which is literally composed of humans acting on its behalf, an AI is not guaranteed to come with a *superior* who will respond to the law. This is not to say that the *respondeat superior* institution is not usable towards machines. It is usable, and some doctrine uses it to fill into the gaps of grey-zones of illegality, but only where the *superior-inferiro* scheme-relationship is found and not as a general rule (Abbott and Sarch, 2019).

7.2.3 The Strict Liability Remedy

One potential strategy to addressing the grey area where *respondeat superior* results inapplicable is to establish a set of new strict liability offences specifically tailored for AI crimes. Strict liability offences would allow AI to be held criminally liable without any requirement for *mens rea*, such as intent, knowledge, recklessness, or negligence. Instead, AI entities would be held liable for their actions without regard to their mental states, enabling punishment of AI without requiring a culpable mental state. Strict liability offences are often criticised in the context of human criminal law because they can lead to the unjust punishment of innocent individuals. However, this objection loses some of its force when applied to AI because AI does not enjoy the same protections based on desert constraints as humans.

Nonetheless, there are practical challenges to the application of strict liability offences to AI. The voluntary act requirement is an absolute necessity for criminal liability, meaning that *„only bodily movements guided by conscious mental representations count"* (Yaffe, 2012). Since AI lacks mental states, deliberation, and reasoning, it becomes difficult to establish any of its behaviours as voluntary acts.

One potential solution to this issue is to alter or eliminate the voluntary act requirement through a statute specifically for the class of strict liability offences designed for AI. Statutory amendments could impose affirmative duties on AI to prevent harmful conduct, allowing AI to be held strictly liable for omissions. However, this approach comes with potential costs, as it may dilute the public meaning and value of criminal law, undermining its expressive benefits, which are essential for justifying the punishment of AI (Abbott and Sarch, 2019).

8. CONCLUSIONS: CHALLENGES OF PUNISHING AI

As we saw, punishing AI is the result of mere logical inferences. A crime, in order to be such, needs to meet the requisites *actus reus and* of *mens rea.* We find the first and the second in crimes that fall inside the Perpetration-via-Another and the Natural-Probable-consequence liability models.

The punishment of AI carries several practical challenges and necessitates substantial innovations in existing criminal law in residual cases of AI's direct liability,

DOI: 10.46282/blr.2024.8.2.832

where it is not possible to invoke *respondeat superor* liability model. In such cases, one of the main challenges lies in the assessment of *mens rea*, which is a fundamental aspect of human criminal justice. However, when it comes to AI, determining *mens rea* becomes exceedingly complex. This is particularly true for Hard-AI crimes that cannot be straightforwardly attributed to human conduct or where harm is unforeseeable to designers without unreasonableness, the application of *respondeat superior* or similar principles is not appropriate. In such cases, an entirely new approach to assessing AI *mens rea* would be necessary.

One potential solution discussed is the establishment of strict liability offences for AI crimes, which would require substantial legislative revisions to criminal law, ensuring that AI entities can meet the voluntary act requirement. This is not a simple or readily available solution, and it demands extensive legislative efforts and legal amendments. An alternative approach would involve the development of a legal fiction for AI *mens rea*, somewhat analogous to human *mens rea*, necessitating expert testimony to assess the AI's functioning, including its consideration of legally relevant values, interests, and behavioural dispositions related to *mens rea*-like intention or knowledge. Although this approach has been tentatively explored, it requires further theoretical and technical development.

Bestowing legal personhood to AI is indispensable for charging and convicting AI of crimes, thereby introducing an entirely new form of criminal liability, similar to the emergence of corporate criminal liability beyond individual criminal liability. Granting legal personality to AI has been contemplated in various proposals, but it has been highly controversial.[22] It is essential to emphasise that AI legal personality does not grant AI the full range of rights afforded to natural persons or even corporations. Instead, it could be limited to obligations.

However, conferring legal personhood on AI, even in a limited sense, presents several challenges. AI's anthropomorphism could encourage people to impose human attributes, expectations, and behaviour on AI, leading to mistreatment of AI, and even vandalism or attacks against these entities. It could also influence human well-being, raising concerns about humans standing in society, especially if AI is granted legal status on par with humans. Additionally, there is the concern of "rights creep", where over time, AI might acquire more rights, leading to unforeseen legal complexities and implications.

In conclusion, the practical challenges associated with AI punishment extend beyond *mens rea* analysis and encompass broader restructuring of criminal law and potential societal consequences of assigning legal personhood to AI. These challenges necessitate careful consideration of the implications and risks before embarking on the path of punishing AI. However, given the increasing integration of AI into daily life - manifested in forms such as autonomous vehicles - these issues present an urgent and practical problem that demands resolution in the near future. As the prevalence of AI technologies continues to grow, addressing these concerns becomes paramount.

The differentiation between various national criminal legislations may provide different answers; however, the aspect of the examinations will be the same (Hodula, 2021).

[22] See also: Open Letter to the European Commission Artificial Intelligence and Robotics, Available at: http://www.robotics-openletter.eu/ (accessed on 22.06.2024). More than 150 AI "experts" subsequently sent an open letter to the European Commission warning that, from *„an ethical and legal perspective, creating a legal personality for a robot is inappropriate whatever the legal status model"*.

BIBLIOGRAPHY:

Abbott, R. and Sarch, A. F. (2019). Punishing artificial intelligence: Legal fiction or science fiction? *UC Davis Law Review*, 53(1), 323–333. https://doi.org/10.2139/ssrn.3327485

Algeri, L. (2021). Intelligenza artificiale e polizia predittiva [Artificial Intelligence and Predictive Policing]. *Diritto Penale e Processo*, 2021(6), 724–734.

Asaro, P. (2012). A body to kick, but still no soul to damn: Legal perspectives on Robotics. In: Lin, P., Abney K. and Bekey, G. A. (ed.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 169–186). Cambridge: MIT Press.

Basile, F. (2019). Intelligenza artificiale e diritto penale [Artificial Intelligence and Criminal Law]. *Diritto Penale e Uomo*, 1–33. Available at: https://archiviodpc.dirittopenaleuomo.org/d/6821-intelligenza-artificiale-e-diritto-penale-quattro-possibili-percorsi-di-indagine (accessed on 24.10.2024).

Bassini, M., Liguori, L. and Pollicino, O. (2018). Sistemi di intelligenza artificiale, responsabilità e accountability [Artificial Intelligence Systems, Responsibility and Accountability]. In: Pizzetti, F. (Ed.), *Intelligenza artificiale, protezione dei dati personali e regolazione* (pp. 333–369). Torino: Giappichelli.

Berman, N. M. (2012). Introduction: punishment and culpability. *Ohio State journal of criminal law*, 9, 441–448.

Boden, M. A. (2018). Intelligenza artificiale [Artificial Intelligence]. In: Al-Khalili, J. (Ed.), *Il futuro che verrà*. Turin: Bollati Boringhieri.

Burda, E., Čentéš, J., Kolesár, J. and Záhora, J. (2010). *Trestný zákon I. Všeobecná časť, komentár* [*Criminal Code I. General Part. Commentary*]. Praha: C. H. Beck.

Butler, T. L. (1982). Can a Computer Be an Author: Copyright Aspects of Artistic Intelligence. *Comment L.S,* 4, 707–747.

Butz, M. V. (2021). Towards strong AI. *KI-Künstliche Intelligenz,* 35, 91–101. https://doi.org/10.1007/s13218-021-00705-x

Cadoppi, A. and Veneziani, P. (2004). *Elementi di diritto penale – Parte generale (2nd ed.)* [*Elements of Criminal Law – General Part*]. Padova: CEDAM.

Carrara, F. (1871). *Programma del corso di diritto criminale. Parte generale* [*Syllabus of Course Criminal Law – General Part*]. Lucca: Giusti.

Cattaneo, M. A. (1978). *Il problema filosofico della pena* [*The philosophical problem of punishment*]. Ferrara: Ed. Universitaria.

Ciappi, S. and Coluccia, A. (1997). *Giustizia criminale* [*Criminal Justice*]. Milano: Franco Angeli.

Dressler, J. (2007). *Cases and materials on criminal law*. St. Paul: West Academic Publishing.

Fiandaca, G. and Musco, E. (2007). *Diritto penale, Parte generale (5th ed.)*. Bologna: Zanichelli.

Floridi, L. (2019). What the near future of artificial intelligence could be. *Philosophy & Technology*, 32, 1–15. https://doi.org/10.1007/s13347-019-00345-y

Gless, S., Silverman, E. and Weigend, D. (2016). If robots cause harm, who is to blame? Self-driving cars and criminal liability. *New Criminal Law Review*, 19(3), 412–436. https://doi.org/10.1525/nclr.2016.19.3.412

Glover, E. (2022). Strong AI vs. weak AI: What's the difference? *Built In*. Available at: https://builtin.com/artificial-intelligence/strong-ai-weak-ai (accessed on 24.10.2024).

Hallevy, G. (2010). The criminal liability of artificial intelligence entitiesfrom science fiction to legal social control. *Akron Intellectual Property Journal*, 4(2), 171–201.

Hart, H. L. A. (1968). *Punishment and responsibility*. Oxford: Clarendon Press.
Hodula, M. (2021). Criminal liability of self-driving vehicles. In: *Bratislava Legal Forum 2021: Cybercrime in the time of crisis* (pp. 8–14). Bratislava: Comenius University, Faculty of Law.
Hoffmann, J. (1993). *Vorhersage und Erkenntnis: Die Funktion von Antizipationen in der menschlichen Verhaltenssteuerung und Wahrnehmung* [*Prediction and Cognition: The Function of Anticipations in Human Behavior Control and Perception*]. Göttingen: Hogrefe, Verlag für Psychologie.
Holbrook, J. (2020). Artificial intelligence: "Soft vs hard." *Medium*, published on 21.07.2020. Available at: https://medium.com/hackguild/artificial-intelligence-soft-vs-hard-17b8e1c343d7 (accessed on 24.10.2024).
Kaplan, J. (2018). *Intelligenza artificiale. Guida al futuro prossimo* [*Artificial Intelligence: A Guide to the Near Future*]. Roma: Luiss University Press.
Kof, J. N., Bboers, E. J. W., Kosters, W. A., Putten, P. and Poel, M. (2002). Artificial intelligence: Definition, trends, techniques, and cases. In: *Knowledge for sustainable development: An insight into the Encyclopedia of Life Support Systems* (pp. 1095–1107). Paris: UNESCO; Oxford: EOLSS.
Lacey, N. and Wells, C. (1998). *Reconstructing criminal law - Critical perspectives on crime and criminal process*. London: Weidenfeld and Nicholson.
Lagioia, F. and Sartor, G. (2020). AI systems under criminal law: A legal analysis and a regulatory perspective. *Philosophy & Technology*, 33, 433–465. https://doi.org/10.1007/s13347-019-00362-x
Minsky, M. (1967). *Computation: Finite and infinite machines*. Englewood Cliffs, NJ: Prentice-Hall.
Moor, J. (2006). The Dartmouth College Artificial Intelligence Conference: The next fifty years. *AI Magazine*, 27(4), 87–91, https://doi.org/10.1609/aimag.v27i4.1911
Mulligan, C. (2018). Revenge against robots. *South Carolina Law Review*, 69(3), 579–595. https://doi.org/10.2139/ssrn.3016048
Nicotra, I. (2014). Pena e reinserimento sociale ad un anno dalla sentenza Torregiani [Punishment and social reintegration one year after the Torregiani sentence]. *Diritto penitenziario e costituzione*. Available at: https://www.dirittopenitenziarioecostituzione.it/images/pdf/saggi/I_Nicotra_Pena_e_reinserimento_sociale.pdf (accessed on 24.10.2024).
Njoku, F. O. C. (2012). Hart on philosophical and legal conception of punishment. *International journal of technology and reformed tradition,* 4, 220–231.
Nuvolone, P. (1982). Pena (dir. pen.). In: *Enciclopedia del diritto* [*Encyclopedy of Law*] (Vol. XXXII, pp. 787–817). Milano: Giuffrè.
Oss, G. (2009). Certezza della pena e trattamenti rieducativi: Un contrasto insanabile? [Certainty of punishment and re-educational treatments: An irreconcilable contrast?]. Available at: http://www.ristretti.it/commenti/2010/gennaio/pdf2/giorgia_oss.pdf (accessed on 24.10.2024).
Padhy, N. P. (2005). *Artificial intelligence and intelligent systems.* Oxford: Oxford University Press.
Pervukhin, A. (2005). Deodands: A study in the creation of common law rules. *American Journal of Legal History*, 47(3), 237–256. https://doi.org/10.2307/30039513
Piparo, C. (2023). Machina delinquere potest? A modern criminalization challenge due to lack of text. *Text, context, and subtext in law.* Timisoara: Universul Juridic, 900–909.
Pizzetti, F. (2018). Intelligenza artificiale: Passato, presente, futuro. In: Pizzetti, F. (Ed.), *Intelligenza artificiale, protezione dei dati personali e regolazione*. Torino: Giappichelli, 216–224.

Remeta, R. (2023). Principles of punishment in the light of the proposed changes. In: *Bratislava Legal Forum 2023: Current challenges of criminal law* (pp. 142–154). Bratislava: Comenius University, Faculty of Law.

Rockwell, A. (2017). The history of artificial intelligence. *Harvard SITN*, published on 28.08.2017. Available at: https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/ (accessed on 24.10.2024).

Russell, S. and Norvig, P. (2011). *Artificial intelligence: A modern approach (3rd ed.)*. Pearson. https://doi.org/10.1016/j.artint.2011.01.005

Searle, J. R. (1984). *Minds, brains and science*. Cambridge, MA: Harvard University Press. Available at: https://rodsmith.nz/wp-content/uploads/Searle_1984_minds-brains-and-science.pdf (accessed on 24.10.2024).

Schank, R. C. (1987). What's AI, Anyway?. *IA Magazine,* 8(4), 59–65.

Smejkal, V. (2023). Trestní odpovědnost za jednání robotů využívajících AI [Criminal Liability for the Actions of AI Robots]. In: Čentéš, J. et al (eds.), *Bratislava Legal Forum 2023: Current challenges of criminal law* (pp. 46–73). Bratislava: Comenius University, Faculty of Law.

Srstka, J. et al. (2024). *Autorské právo a práva související* [*Copyright and related rights*]. Praha: Leges.

Strémy, T. and Klátik, J. (2018). *Alternatívne tresty* [*Alternative punishments*]. Bratislava: C. H. Beck.

Turing, A. M. (1950). *Computer machinery and intelligence. Mind,* 49(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Yaffe, G. (2012). The Voluntary Act Requirement. In: Marmor, A. (ed.), *The Routledge Companion to Philosophy of Law* (pp. 1–31). New York, NY: Routledge.

Walker, L. (2015). Stephen Hawking warns artificial intelligence could end humanity. *Newsweek*, published on 14.05.2015. Available at: https://www.newsweek.com/stephen-hawking-warns-artificial-intelligence-could-end-humanity-332082 (accessed on 24.10.2024).

Act No. 300/2005 Slovak Coll., as amended.

Act No. 91/2016 Slovak Coll., as amended.

European Commission. (2018, December 18). A definition of AI: Main capabilities and scientific disciplines. Brussels. Available at: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf (accessed on 24.10.2024).

European Parliament. (2017, February 16). European Parliament resolution providing recommendations to the European Commission on civil law rules on robotics [2015/2103(INL)].

Italian Constitution, ratified on 22 December 1947, as amended.

Resolution of the Government of the Slovak Republic No. 385/2000 on the legislative intent of the Criminal Code and the Criminal Code.

DOI: 10.46282/blr.2024.8.2.832