

Derecho de la neurología: Libro blanco para el desarrollo de la
Carta de Derechos Digitales en el ámbito de la neurotecnología
Beatriz Gregoraci, Jorge Castillo Abella y Francisco Velasco Caballero (dirs.)
ISBN: 979-13-87913-93-3
Madrid, 2026
pp. 641-662
DOI: 10.37417/derecho-de-la-neurotecnologia/14
Marcial Pons Ediciones Jurídicas y Sociales
Editado bajo licencia Creative Commons Attribution 4.0 International License

CAPÍTULO XIV

«NO FUI YO, FUE LA MÁQUINA»: TEORÍA DEL DELITO, RESPONSABILIDAD E INTELIGENCIA ARTIFICIAL

Susana AIRES DE SOUSA
Universidade de Coimbra (Portugal)¹
susanaas@fd.uc.pt

1. INTRODUCTION: DELEGATION TO THE MACHINE

The factory of the future will have only two employees, a man and a dog. The man will be there to feed the dog. The dog will be there to keep the man from touching the equipment.

This is a well-known quote included in Warren Bennis's book *On Becoming a Leader* (1989). This statement can lead us to many discussions and topics: security and innovation challenges in companies, new management models, the automation of the production process.

¹ Profesora asociada, investigadora integrada, Facultad de Derecho, II, Univ Coimbra; <https://orcid.org/0000-0003-4808-2466>.

The main ideas of this text were previously developed in a paper published in portuguese, «Não fui eu, foi a máquina: teoria do crime, responsabilidade e inteligência artificial», in Anabela MIRANDA RODRIGUES (coord.), *A inteligência artificial no direito penal* (coord.), Coimbra, Almedina, 2020, pp. 59-93. The present version revises the text and introduces some new elements.

But none of these is the topic I intend to explore. That statement contains a prediction that, today —fast-forwarding from 1990 to now— can be considered true: in today's factories, not only does automation dominate, but so does the delegation of functions and decision-making to machines. These decisions and actions, traditionally made by humans, now fall under the responsibility of autonomous systems.

Expanding the focus beyond factories to today's world, we see that this is neither a futuristic nor a science-fiction scenario. If we think about it, autonomous AI systems are present in our daily lives and influence the decisions we make². While they are far from the catastrophic depictions often seen on screen, it is certain that their use is already integrated into our routines —whether through information and communication systems or even advisory functions (technical or economic); whether on the internet, in computers, or on smartphones; in medical contexts with diagnostic systems or surgical robots; in the stock market sector through algorithmic trading; or in the transportation sector via autonomous vehicles (cars, shuttles, boats or drones).

Among the various applications of artificial intelligence systems, we can identify several key functions: a *descriptive* and *advisory* role that suggests what to do; a *diagnostic* function that identifies a specific event; a *predictive* function capable of anticipating uncertainties; a *decision-making* function that can make and implement choices³; and even a *creative* function that can present innovative and unexpected solutions.

Delegating these functions to machines raises issues of accountability when their decisions or actions cause harm to protected legal interests. In other words, what happens when things go wrong? The advice was flawed or biased, leading to a stock market crash; the car fails to identify a pedestrian and runs them over; the surgical robot has a technical failure that damages healthy tissue; or the household robot mistakes a pet turtle for a piece of plastic. If there is harm to interests protected by criminal law, who is held responsible for the algorithmic damage? The challenges for the theory of criminal offenses begin⁴.

² Cf. *White Paper on Artificial Intelligence - An European approach to excellence and trust*, European Commission, Brussels, 19/2/2020; also BANTEKA (2021), «Artificially intelligent persons». *Houston Law Review*, vol. 58, p. 537; FLORIDI (2017), «Robots, Jobs, Taxes, and Responsibilities». *Philosophy and Technology*, 30, p. 1. SIMMLER, SIMMLER & MARKWALDER (2019), «Guilty Robots – Rethinking the nature of culpability and legal personhood in an age of artificial intelligence», *Criminal Law Forum*, Springer, p. 2.

³ About the first four functions, GIUFFRIDA (2019), «Liability for AI Decision-Making: some legal and ethical considerations». *Fordham Law Review*, vol. 88, p. 440.

⁴ It is important to emphasize that connections between AI and criminal law in a broad sense are numerous. These connections are evident, for instance, in procedural criminal law, as technology provides new possibilities for criminal investigations and the collection of relevant

When harm or risk is caused by human action, legal theory is well established to address the issue. However, when harm or danger to legal interests (*Rechtsgüter*) is caused or connected to an autonomous machine, the situation becomes more complex.

2. CHALLENGES ON THE GENERAL PART OF CRIMINAL OFFENSES

2.1. Context

Defining artificial intelligence is, in itself, a challenge⁵. There are numerous definitions, which have evolved in complexity—both conceptually and technologically—far beyond Turing's initial vision of AI as merely imitating human behavior to the extent that a third party could not distinguish between machine and human responses⁶. AI encompasses a broad range of systems based on complex computational programming with an immense data analysis capacity and a high degree of autonomy that far surpasses human ability⁷.

Particularly relevant, as it raises the most difficult questions regarding liability, are AI systems based on machine learning. Such an AI system does not merely calculate the best option from pre-fed data. It does not simply perform an analysis beyond human reach (such as choosing from countless possible moves in a strategic game like chess). Instead, it searches for, analyzes, and compares patterns to develop a model that determines its optimal move⁸. Rather than merely retrieving the best-known option from its vast database, the algorithm, fuelled by data, continuously adjusts itself to minimize errors. A relatively consensual definition of machine learning in the literature describes it as a set of statistical tools and processes applied to a broad data base through algorithms, from which a model, rule, or procedure is derived to predict new information and, potentially, make a decision

information. Likewise, in terms of legal consequences, AI can assist prosecutors and judges in risk assessment, whether during the investigation phase or in judicial proceedings, such as determining precautionary measures, setting penalties, or granting parole.

⁵ About this difficulty see *The one hundred year study on artificial intelligence*, published by Stanford University, 2016, p. 12. For a more in-depth analysis, with various possible criteria and definitions, TURNER (2019), *Robot Rules. Regulating Artificial Intelligence*, Palgrave Macmillan, p. 7 ff.

⁶ Cf. TURING (1950), «Computing, Machinery and Intelligence». *Mind*, vol. LIX, núm. 236, 1950, p. 433 ff.

⁷ Cf. BATHAE (2018), «The artificial intelligence black box and failure of intent and causation». *Harvard Journal of Law & Technology*, vol. 31, p. 898.

⁸ Cf. BATHAE (2018), *op. cit.*, p. 898. For example, the Alpha Go programme developed by the company Deep Mind and later acquired by Google, which is capable of creating its own moves that are qualified, in the eyes of a professional player, as creative.

based on it—either directly by the machine or as support for a final decision made by a human⁹.

The methodology used to teach the machine may vary, from a more or less supervised and trained technique to achieve a particular result, to reinforcement learning. This is where the metaphor of «learning» comes into play. The system is designed to adjust its parameters, become more efficient, and create a model capable of generating new information. This «learning» process relies on complex mathematical and statistical systems enabled by computational advancements. Since the specific mechanisms behind a (complex) machine's response are often inaccessible or inexplicable even to its programmers, the literature refers to this as the black box problem.

2.2. The black box problem and causality

The black box problem refers to the opacity and lack of transparency in some of the more complex AI models regarding how they process input data and reach a given output. In other words, algorithmic data processing follows a complex structure that obscures the process leading to a specific (opaque) outcome, despite its high precision in determining or generating new information. The opacity increases with the complexity and accuracy of the machine learning models used. In some cases, the current state of technological development does not allow us to determine exactly how a system arrived at a particular outcome—be it a predictive judgment, a recommendation, or a decision¹⁰. This lack of transparency has garnered significant attention, not only from AI researchers working on transparency techniques but also in legal scholarship¹¹, where some advocate the recognition of a «right to explanation» for algorithm-driven decisions¹².

A comprehensive study on the legal implications of this lack of transparency—especially regarding legal categories underpinning liability assessments, such as cause and intent—was conducted by Yavar Bathaee. As this scholar highlights, AI based on machine learning algorithms, including techniques like deep neural networks or support vector machines, can be as difficult to comprehend as the human brain¹³.

⁹ Cf. BANTEKA (2021), *op. cit.*, p. 544.

¹⁰ This opacity also leads to difficulties in proving the causal mechanism in court, even for machine learning experts. On the procedural and evidential dimension of this black box problem, GLESS (2020), «AI in the courtroom: a comparative analysis of machine evidence in criminal trials». *Georgetown Journal of International Law*, vol. 51, p. 211.

¹¹ SAMEK, WIEGAND & MÜLLER (2017), available in <https://arxiv.org/abs/1708.08296>.

¹² GOODMAN (2017), «European Union Regulations on Algorithmic Decision Making and a «Right to Explanation»». *AI Magazine*, 38(3), p. 50 ff.

¹³ Cf. BATHAE (2018), *op. cit.*, p. 891.

Neuroscience has long sought to map human decision-making, a challenging task despite significant advancements¹⁴. Similarly, complex AI systems present an analogous issue: it is not possible to map the precise process that led to a given output. For instance, a deep neural network operates through a complex, multidimensional architecture of artificial neurons that learn from experience, interconnecting and processing vast amounts of information.

Naturally, these relatively recent systems have been enabled by technological advancements. However, the more opaque, complex, and autonomous they become, the more inaccessible the exact mechanism behind their outputs remains. Some scholars have identified this as a challenge to establishing causal links within predominant legal doctrines in criminal law. These concerns are particularly pronounced in normative-causal theories, which emphasize the predictability of results—where causality is assessed based on the normal course of events and experiential rules, as exemplified by the theory of adequate causation (*Adäquanztheorie*)¹⁵. A legal-normative criterion that hinges causality on the foreseeability of harm encounters significant difficulties when dealing with unpredictable or disruptive occurrences where no prior generalization exists. In an (illogical) way, the «duty of care» arises for the human agent only after their actions have taken place.

2.3. The failure of the subjective element (*mens rea*)

The unpredictability and dynamic nature of complex computational systems raise doubts regarding the subjective attribution of harm, which is a core requirement in criminal offenses. From the perspective of a human involved in the manufacturing, programming, or use of an AI system, the machine's intervention renders the harmful outcome unforeseeable.

The opacity of the system and the unpredictability of the harmful outcome make it difficult to prove either that the human foresaw the result or that such a possibility even existed. The machine objectively determines an outcome by processing information through an algorithm, based on a specific model, without any «intention» or «negligence» in the legal sense. As noted in the literature, «machines and software programs do not have intent»¹⁶—only people act with a specific intention, purpose, or goal.

¹⁴ AIRES DE SOUSA (2017), «Neurociências e processo penal: verdade *ex machina*», *Estudos em homenagem ao Senhor Prof. Doutor Manuel da Costa Andrade*, vol. II, Boletim da Faculdade de Direito da Universidade de Coimbra, p. 883 ff.

¹⁵ About this theory and the scholars who developed it, AIRES DE SOUSA (2014), *A Responsabilidade Criminal pelo Produto e o Topos Causal em Direito Penal*. Coimbra, Coimbra Editora, 2014, p. 383 ff. and 454 ff.

¹⁶ Cf. BATHEE (2018), *op. cit.*, p. 906.

Intent is thus attributed to human actors who create, program, or use the machine. However, as previously discussed, «intelligent programs» are capable of producing certain results through processes that were not initially (*ex ante*) accessible to humans. During these processes, harm may occur—harm that was neither desired, anticipated, nor even foreseeable to the humans involved.

It is crucial to distinguish between the final outcome—potentially defined by a human input—and the process leading to that outcome—which remains hidden or opaque to human understanding (the *black box*). When such a process harms protected legal interests, who bears responsibility? This issue becomes particularly critical when it involves harm to legal interests protected under criminal law.

Since criminal norms, particularly in economic contexts, primarily typify offenses as intentional acts, it becomes nearly impossible to attribute subjective liability when harm occurs via complex AI systems. An example is the use of AI in high-frequency trading in the stock market, which is valued for its speed and efficiency but can also be exploited for market manipulation strategies (e.g., spoofing). In these cases, proving the subjective element required for the crime of market manipulation is nearly impossible. The U.S. legal experience confirms this difficulty, given the challenge of proving intent in fraud-based offenses. The subjective element, designed as a «gatekeeping test»¹⁷ for liability (particularly in fraud cases), becomes an obstacle when harm is caused by autonomous AI systems.

2.4. Humans and machines as actors

The attribution of liability is further complicated by the multiplicity of actors—both human and non-human—involved. AI-driven technological advances have given rise to new concepts describing a new industrial and corporate reality, often referred to as the Fourth Industrial Revolution (Industry 4.0). This revolution is fuelled not only by advancements in computing but also by the networked communication of systems (the Internet of Things¹⁸). In such a scenario, machines and systems communicate and interact autonomously, making production processes more efficient and even predicting errors before they occur. The traditional production model is shifting, with decision-making

¹⁷ BATHAE (2018), *op. cit.*, p. 921, develops these and other ideas regarding artificial intelligence.

¹⁸ One possible scenario in the relatively near future concerns a technologically developed road system in which automated driving takes place via communication between systems, with little or limited human intervention. In the sense of road transport becoming part of the so-called «internet of things», see HILGENDORF (2017), «Automated driving and the law». *Robotics, Autonomics and the Law*, Nomos, p. 173.

increasingly being delegated to machines and to AI-driven systems (the factory of the future is being designed today).

This complexity increases when multiple human actors are involved, such as AI programmers, data collectors, algorithm trainers, software developers, machine manufacturers, system owners, and those integrating them into production¹⁹. With such a diverse range of contributors, liability attribution becomes increasingly complex when something goes wrong. Who should be held accountable?

3. APPROACHING LEGAL SOLUTIONS

3.1. Liability of natural and legal persons

In its classical formulation, the theory of crime is based on the attribution of an event deemed wrongful—either for harming or endangering a legal interest—to human conduct. However, frameworks for assigning criminal liability do not necessarily require that such harm be caused by the immediate action of a human agent. Legal doctrine and legislators have developed and codified forms of liability that extend the traditional model of responsibility, both in terms of the subject held accountable and the nature of the incriminated conduct.

From an agency perspective, criminal responsibility has been extended to legal persons. Also, even from the perspective of the (human) agent, the attribution of a criminal act to an individual person no longer requires their direct physical execution of the act. Individuals can be held accountable in cases of joint participation, even if the unlawful act was carried out entirely or partially by another party—such as in cases of indirect perpetration, instigation, or co-perpetration (as provided for in Article 26 of the Portuguese Penal Code). Lastly, most legal systems have developed and codified liability frameworks for omissions (*Garantstellung*), holding individuals accountable for failing to act when they had a legal duty to prevent a harmful outcome.

In certain cases, the issue of criminal liability for harming a legal interest involving an AI system does not pose significant challenges. This is particularly evident when the system is designed, deployed, or manipulated with the primary intent of facilitating criminal activity. Examples include the use of advanced technological systems for the transportation and trafficking of illegal substances, the creation of trading algorithms to manipulate market information, or the development of software to gain unauthorized access to or destroy third-party data. Additionally, there are documented instances of AI manipulation

¹⁹ Cf. GIUFFRIDA (2019), *op. cit.*, p. 443.

for illicit purposes, such as the deliberate introduction of elements that interfere with an autonomous vehicle's ability to recognize traffic signals, thereby intentionally causing an accident.

Criminal law, in its typical scope, already encompasses such conducts. In such cases, liability will be attributed to the perpetrator, and at a theoretical level, one might even consider classifying the case as indirect perpetration —particularly in scenarios where the harmful act is ultimately carried out by an instrumentalized victim (such as a malicious program that, once downloaded by an unsuspecting user, destroys specific data or information)²⁰.

However, as previously noted, the challenge of assigning liability increases in cases where the harm —attributed to the machine— cannot be traced back to any human intent or foreseeability, whether on the part of the programmer, manufacturer, or user²¹. The difficulty is particularly evident when there was no human intention or predictability regarding the unlawful outcome during the programming or use of the machine. In such cases, liability for the harmful result could only be assigned to an individual —such as the programmer, manufacturer, or user— under negligence-based liability, and only if a violation of an objective duty of care can be established.

Even so, significant obstacles remain, including the challenge of defining duties of care in areas of scientific innovation that lack robust regulation, or identifying precisely who should be held accountable when harm results from a complex interplay of contributions —from programmers and data providers to software owners and users. As a result, it is difficult to pinpoint liability in relation to any single individual's conduct²².

It is worth noting that many AI systems are developed and implemented within corporate settings. Consequently, legal persons that produce, market, or use AI systems may also bear liability, provided that legal provisions allow for corporate criminal responsibility. However, even here, challenges arise in assessing the conditions necessary to hold the corporate entity accountable. Specifically, when harm arises from an autonomous decision made by a machine —such as a robot failing

²⁰ In this context, some literature refers to figures close to the principle of “superior responsibility” as a way of attributing responsibility to the individual. This would be the case of the person who uses the machine to carry out a crime (*perpetration-by-another*) or the person who, knowing or foreseeing that the crime will be carried out by someone acting under their command, does not prevent it from happening (*command responsibility model*). Cf. KING, AGGARWAL, TADDEO & FLORIDI (2020), «Artificial Intelligence crime: an interdisciplinary analysis of foreseeable threats and solutions». *Science and Engineering Ethics*, p. 109 ff. Also HALLEVY (2010), «The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control». *Akron Intellectual Property Journal*, vol. 4, issue 2 (2010), p. 179.

²¹ Cf. PAGALLO (2013) *The Laws of Robots*. Springer, p. 69.

²² Cf. *Ibid.*, p. 70.

to detect a worker and fatally crushing them²³— it may be challenging to establish the necessary criteria under either a direct liability framework or a vicarious liability model. Under the latter—where a corporate entity is held accountable for the actions of its representatives or those in leadership positions, a model primarily followed in Portuguese law²⁴— it may be challenging to attribute the harm through the actions of those specific individuals. On the other hand, under a model based on corporate fault, the autonomy of AI systems introduces an element of unpredictability that is difficult to predict, detect or mitigate within the corporate environment²⁵.

Thus, corporate liability for harm caused by so-called «intelligent» machines is not exempt from obstacles—on the contrary, it presents significant challenges. The complexity is likely to increase as more tasks traditionally performed by human workers are transferred to or delegated to autonomous systems. As Mihailis E. Diamantis highlights in an article dedicated to this issue²⁶, algorithmic decision-making in corporate environments necessitates a reassessment of the foundations of corporate liability. Specifically referring to the U.S. legal system, Diamantis argues that, since the same harmful task may be carried out by a human in some companies and by an autonomous system in others, corporate liability frameworks should be expanded (through the concept of «the corporate mind») to include algorithms among the entities capable of binding a company to criminal responsibility²⁷.

3.2. Machines (AI systems) liability

The perceived insufficiency of classical legal frameworks for attributing criminal liability has been a key driver behind theoretical proposals advocating for *electronic legal personality* in civil law and, consequently, the direct criminal liability of machines as a response to the *responsibility/accountability gap*.

²³ This is a real example described by WENG, CHEN, & SUN (2009), «Toward the Human–Robot Co-Existence Society: On Safety Intelligence for Next Generation Robots». *International Journal of Social Robotics* (2009), p. 273. Also on this case, HALLEVY (2013), *When Robots Kill*. Northeastern University Press, 2013, p. 38. In 1981, in Japan, an engineer entered a restricted area of the production sector to service a robot; having failed to switch it off completely, he was pushed and crushed by its hydraulic arm.

²⁴ AIRES DE SOUSA (2023), «Portuguese Report on traditional criminal law categories and AI» *Revue Internationale de Droit Pénal*, vol. 94, issue 1, p. 325.

²⁵ AIRES DE SOUSA (2024), «The importance of AI for financial crimes». *Revue Internationale de Droit Pénal*, vol. 95, issue 1, 2024, p. 176.

²⁶ DIAMANTIS (2020), «The extended corporate mind: when corporations use AI to break the law». *North Carolina Law Review*, vol. 98, 2020, p. 893 ff.

²⁷ *Ibid.*, p. 893 ff.

Indeed, the technological possibility of strong AI—capable of autonomously learning and making decisions—has echoed in legal literature as a basis for the potential liability of artificial intelligence²⁸. The subjectivization/humanization of AI is undeniable and dates back to its very origins, with its creation and development guided by processes that imitate the human mind. This personalization is, in part, induced by the narrative surrounding artificial intelligence, but also by language itself, which, through metaphor, makes complex computing systems comprehensible to those without specialized technical knowledge. AI systems become «humanized» through language, acquiring human-like qualities and attributes: they are described as intelligent, capable of learning and acting, making autonomous, and sometimes even creative or spontaneous, decisions.

Thus, it is not entirely unexpected that some have proposed granting legal personality to AI, making it a legal entity capable of holding rights and responsibilities²⁹. A significant push in this direction came from the European Parliament in its Resolution of February 16, 2017, which contained recommendations to the European Commission regarding future civil legislation on robotics and AI. Among its suggestions, point 59(f) of the report proposed the creation of «a specific legal status for robots,» opening the possibility of granting the most sophisticated autonomous robots the status of «electronic persons»³⁰. This recommendation has been interpreted in legal literature as a suggestion that so-called «strong AI» could be legally recognized—at least in civil law—as an electronic person, enabling the liability of robots that can make autonomous decisions and independently interact with third parties³¹. In other words, this would apply in cases where holding natural persons accountable is complicated by the autonomy and independence of artificial intelligence mechanisms. However, it should be noted that this recommendation sparked significant controversy³²,

²⁸ With interesting examples and a summary of the historical evolution of the arguments against and in favour of a legal personality for AI, TURNER (2019), *op. cit.*, p. 173 ff.

²⁹ About the topic, BECK (2013), «Über Sinn und Unsinn von Statusfragen – zu Vor – und Nachteilen der Einführung einer elektronischen Person», HILGENDORF & GÜNTHER (coord.), *Robotik und Gesetzgebung*, Nomos, 2013, p. 239 ff.; GLESS, SILVERMAN & WEIGEND (2016), «If robots cause harm, who is to blame? Self-driving cars and criminal liability». *New Criminal Law Review*, vol. 19, N. 3, p. 416 ff.; SULLIVAN & SCHWEIKART (2019), «Are current tort liability doctrines adequate for addressing injury caused by AI». *AMA Journal of Ethics*, vol. 21, N. 2, p. 163 f.; CHEN & BURGESS (2019), «The boundaries of legal personhood: how spontaneous intelligence can problematize differences between humans, artificial intelligence, companies and animals». *Artificial Intelligence and Law* 27, p. 73 ff., TURNER (2019), *op. cit.*, p. 183; GIUFFRIDA (2019), *op. cit.*, p. 88.

³⁰ https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_PT.pdf.

³¹ BANTEKA (2020), *op. cit.*, p. 22.

³² See, for example, the reaction of specialists in artificial intelligence and robotics, industrial leaders, experts in the field of medicine, law and ethics in an open letter addressed to the

and the European Parliament later distanced itself from it —whether explicitly or implicitly— in subsequent discussions and documents.

Some forms of compensation and damage repair mechanisms for harm caused by AI will need to be created or rethought within civil law (for example, through insurance frameworks or compensation funds) without necessarily requiring the recognition of electronic personality³³. However, some of the arguments supporting the need to grant legal personality to AI have also been extended into criminal law, advocating for the direct liability of machines for the harm they cause³⁴.

The expansion of criminal agency and liability to AI systems has gained increasing attention in legal scholarship, based on the premise that such systems could, in theory, fulfill all the objective and subjective elements necessary for criminal attribution—within a social framework designed to maintain communal expectations³⁵. For instance, Gabriel Hallevy advances the seemingly straightforward argument that any entity meeting the prerequisites for criminal liability should be held accountable, irrespective of whether it is a natural person, a collective entity, or an artificial system³⁶. In his view, extending criminal law to autonomous and intelligent machines would not require fundamental alterations to the core principles of criminal responsibility³⁷, as the actions of AI could still be examined through the established lenses of external conduct (*actus reus*) and mental state (*mens rea*)³⁸.

However, this almost clinical, purely formalist approach to the fundamental categories of criminal law directly clashes—insurmountably—with the ethical foundations of legal wrongdoing and culpability in legal systems such as Portugal's.

Accepting a strictly formalist conception, in which wrongdoing is reduced to the mere causation of harm, would conflate responsibility with causality³⁹. More specifically, within the framework of crimi-

European Commission <http://www.robotics-openletter.eu/>. Also, in the literature on the subject, FLORIDI (2017), *op. cit.*, p. 4; BANTEKA (2020), *op. cit.*, p. 22.

³³ GLESS, SILVERMAN & WEIGEND (2016), *op. cit.*, p. 414, núm. 7 and 8.

³⁴ Cf. HALLEVY (2010), *op. cit.*, p. 171 ff.; (2013), *op. cit.*, p. 38 ff. and 84 ff. (a critical review of this work can be found at CHARNEY (2015), «Can Androids plead automatism? A review of when robots kill: artificial intelligence under the criminal law by Gabriel Hallevy». *University of Toronto Faculty of Law Review*, 69, p. 69 ff.); (2019), *op. cit.*, p. 8 ff. Also, with a critical perspective, GLESS, SILVERMAN & WEIGEND (2016), *op. cit.*, p. 419 ff.; CAPPELLINI (2018), «*Machina delinquere non potest? Brevi appunti su intelligenza artificiale e responsabilità penale*», *Criminalia*, p. 499 ff.

³⁵ Cf. SIMMLER & MARKWADER (2019), *op. cit.*, p. 15 ff.

³⁶ HALLEVY (2010), *op. cit.*, p. 199.

³⁷ HALLEVY (2015), *op. cit.*, p. 61.

³⁸ HALLEVY (2015), *op. cit.*, p. 61 f. and p. 82 ff.

³⁹ PAGALLO (2013), *op. cit.*, p. 38 f: "only a subject capable of acting intentionally, with moral understanding and free will is responsible".

nal law theory, the capacity to act unlawfully cannot be equated with an algorithm's ability to make choices—otherwise, free will would be mistaken for pre-programmed (albeit uncontrolled) decision-making designed to achieve a specific task⁴⁰. Criminal liability presupposes the existence of a conscious, moral agent capable of understanding the normative significance of legal rules and recognizing themselves as a subject bound by the obligations they impose. Without this foundation, criminal wrongdoing is severed from its ethical underpinnings, and culpability—understood as the basis for blame—is rendered meaningless both in the assessment of the act and in the determination of punishment⁴¹. Autonomous AI systems do not meet these requirements, and even if technology were to evolve toward machines capable of moral consciousness, a prior discussion on the ethical and moral boundaries of such development would be essential.

Recognizing machines as criminal agents would mark a radical transformation of punitive law, primarily driven by a symbolic purpose. This conclusion also emerges from other arguments invoked in favor of AI's criminal liability, many of which are similarly consequentialist or symbolic in nature. One such argument is the supposed general deterrence effect (intimidation) on AI system creators and programmers⁴². However, the superficiality of this argument is evident, revealing its inherent weakness: punishing, banning, or even «sentencing» a machine to «death» (i.e., deactivation) does not reach its creator. Otherwise, criminal offenses such as the possession of illegal weapons or drug trafficking would make little sense, as they would merely entail holding the object of the crime accountable. In reality, such punishment would serve a purely symbolic purpose.

Among the many arguments supporting the direct liability of AI, it is also common to compare AI to legal persons, arguing that if abstract entities like corporations can be criminally liable, so too can AI⁴³. However, this comparison is flawed. The supposed equivalence in responsibilities does not hold, as it is based solely on the appearance that both involve a structured, non-human system producing a harmful outcome. However, the substantive difference is significant: while corporations have their own legal personality, they remain fundamentally human constructs, with their decisions guided by human actors in pursuit of corporate objectives. When corporations engage in unlawful activities, they can be held liable precisely because they are ultimately controlled by people. In contrast, the proposal to hold AI systems

⁴⁰ GLESS, SILVERMAN & WEIGEND (2016), *op. cit.*, p. 420.

⁴¹ GLESS, SILVERMAN & WEIGEND (2016), *op. cit.*, p. 423 ff.

⁴² ABBOTT & SARCH (2019), «Punishing artificial intelligence: legal fiction or science fiction». *University of California, Davis*, vol. 53, p. 346 ff.

⁴³ HALLEVY (2015), *op. cit.*, p. 103.

accountable is based on the opposite premise —their autonomy (i.e., the absence of human control) in making «decisions». Thus, the intended equivalence fails: at the core of corporate actions are always natural persons, and corporate legal personality derives from a human foundation.

Another argument found in legal literature suggests that holding AI directly liable would make it easier to compensate victims. However, criminal law should not be directly used for this purpose when the legal system already provides specific instruments expressly designed for victim compensation (such as civil liability mechanisms, insurance frameworks, or compensation funds in the AI context).

Furthermore, the doubts and obstacles surrounding the recognition of AI's legal personality make it difficult —out of systematic coherence— to recognize AI as a criminal agent. This would mean introducing into criminal law something that was not recognized at the constitutional or civil law levels⁴⁴.

A critical analysis of AI's criminal liability would require a more extensive discussion than the one presented here, especially since such a notion fundamentally contradicts the humanist nature and ethical essence of criminal law as a branch of public punitive law. Many of the principles and foundations of criminal liability would be either undermined or outright dismantled by this proposal. However, one unavoidable paradox emerges from the idea of AI's criminal liability: designed to address the accountability gap caused by machine autonomy, criminal liability for AI actually results in a justice deficit by shifting responsibility away from individuals (programmers or users) and corporations. In attempting to resolve the accountability gap, AI becomes a scapegoat, allowing those truly involved to evade responsibility. This artificial notion of liability creates an illusion of accountability —a mere functionalist construct in which «dysfunctions» (whether behaviors, situations, or consequences) are subjected to «liability». This formal and symbolic justification sacrifices substantive justice by shifting responsibility away from those who, as legal persons (natural or corporate), have the capacity to be accountable under the law. The ability to gather new information and make choices does not grant AI systems any moral capacity —and that remains the ultimate and insurmountable limit to their liability⁴⁵.

This raises a crucial question: is AI's decision-making truly an expression of autonomy, or merely an unpredictable process of pre-programmed selection?

⁴⁴ KING, AGGARWAL, TADDEO & FLORIDI (2020), *op. cit.*, p. 108 f.

⁴⁵ BATHAE (2019), *op. cit.*, p. 930.

3.3. «Intelligent» products liability

Holding artificial intelligence mechanisms liable for damages could represent a significant barrier to technological development, making it incompatible with the recognized benefits and advantages of AI. At the same time, there is a concern that strict regulations could hinder market access for new technology companies, placing them at a competitive disadvantage compared to already established firms.

While a total ban on the implementation and development of complex AI systems seems undesirable, it is possible to consider a legal framework based on product liability. This would apply to a product that, despite having special characteristics, in some situations poses a proven inherent risk due to its unpredictability and lack of control. We believe that shifting the focus from the «person» to the «product» and its inherent risks could open up a range of new solutions worth examining in the context of so-called «intelligent» products. Among the previously debated proposals, three deserve special attention in the AI context: liability based on type/nature of the product; the precautionary principle and a dynamic (responsive) regulation⁴⁶.

3.3.1. *Responsibility for the type/nature of the product*

Legal solutions must be open to consider the different types of «intelligent products», mainly from the perspective of their autonomy and of the risks they pose: deterministic products (e.g., deterministic robots programmed for specific tasks) must be dealt with differently from cognitive robots or self-driving cars which are programmed to take autonomous options in complex situations. It may be helpful here to consider the distinction between responsibility for the «type» of product and responsibility for the «process or method of production»⁴⁷.

Typical (traditional) product liability cases deal mostly with a responsibility for the method or process of production or distribution, mainly when problems in manufacturing or in the distribution process generate a defective product. That can happen also with intelligent products: for instance, if the software of the car is malfunctioning

⁴⁶ In this part, regarding these proposals, we closely follow our article, «The importance of AI for financial crimes», *Revue Internationale de Droit Pénal*, vol 95, Issue 1, 2024, p. 172 ff.

⁴⁷ About this distinction, BRICOLA (1978), «Responsabilità penale per il tipo e il modo di produzione». *La Responsabilità dell'impresa per i Danni all'ambiente e ai Consumatori*. Giuffrè Editore, 75 ff.; PIERGALINI (2004), *Danno da Prodotto e Responsabilità Penale*, Giuffrè Editore, p. 40 ff.; AIRES DE SOUSA (2014), *op. cit.*, 106 ff.

and causes an accident, therefore imposing a recall of the product over safety issues⁴⁸.

However, having in mind AI products, we need to consider another form of responsibility which attends to the specific nature of the product. This kind of responsibility is usually mentioned as «responsibility for the type» (in the sense of nature, qualities and risks) of product and goes beyond the fabrication moment, considering the need of regulation at every level: conception, fabrication and use.

Disrupting technology, due to its nature, demands a higher level of minimization of the risks, especially if they are partially predictable. As reported by COMEST (UNESCO) the rapidly increasing presence of cognitive products affects human behaviours and induce social and cultural changes, while also generating issues related to safety, privacy and human dignity. In order to accomplish a better knowledge of AI products and minimize unpredictable harms, their design and conception must comply with the rules that govern our society. This has been and can be called the *design test* or «Value Sensitive Design approach», which takes place in the moment of conception: AI will only pass this test if by design, fundamental principles such as democracy, rule of law and compliance with fundamental rights and principles —such as human dignity, human autonomy and privacy— are incorporated in AI, from the outset of program development⁴⁹. In some cases —those where machines are used in dangerous contexts, such as autonomous vehicles, autonomous weapons, military robots, surveillance products— due to its dangerous or intrusive nature, self-regulation may be short or not enough. Companies must know what they should comply with.

3.3.2. *Precautionary principle and standard measures: balance between law and innovation*

If the risks of AI systems are uncertain, precautionary measures can be established in order to bring some level of trust to the digital world. The uncertainty problem has been addressed in the past by the precautionary principle⁵⁰ and by measures that have this principle as foundation (as well as by the ethical principle of responsibility). Environmental protection, food safety and public health pro-

⁴⁸ Cf. <https://www.reuters.com/business/autos-transportation/tesla-recalls-over-475000-electric-vehicles-2021-12-30/>

⁴⁹ World Commission on the Ethics of Scientific Knowledge and Technology, *Report of COMEST on robotics ethics*.

⁵⁰ Cf. *The Precautionary Principle*. United Nations Educational Scientific and Cultural Organization, COMEST, 2005; FISHER, JONES & VON SCHOMBERG (eds.) (2006), *Implementing the Precautionary Principle. Perspectives and Prospects*, Edward Elgar.

tection are examples of contexts where the precautionary principle has been used.

This principle, based on ethical responsibility⁵¹, «means that when human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that harm». It can be used as a regulation tool in conception, development and use of intelligent products. Two of the main characteristics of the measures decided upon this principle are reversibility and flexibility as the risks become more controlled and manageable.

This principle can be applied for example to AI applications that may be considered dangerous or harmful until proven safe, such as the use of autonomous vehicles in crowded areas or the use of autonomous weapons. It can also be helpful in clarifying some of the duties of the producer and provider. In fact, some level of regulation seems essential for the recognition of legal standards and duties of care such as the duty of monitoring a new and innovative product in order to have a minimum control of the machine. For example, the nature and possible risks of some AI products may demand an intensified and extended duty of supervising and monitoring the product after its commercialization in order to ensure that the product can be trusted —a level of trust adequate to modify the risk level to acceptable and tolerable. In some cases, a lower degree of the machine's autonomy must be considered as well.

The question of the duties associated to the production of potentially harmful products has been addressed within the topic of responsibility for the product. One of the main discussions is related to their extent: should these duties be considered in a wide or in a narrow perspective?

A permanent duty of monitoring and surveillance of the product may sound excessive. In a narrower perspective that duty needs to be restrained to cases where the potential danger of the product is predictable at the moment the product is placed in the market. Some intelligent products may be in this last case, imposing a duty of surveillance. When this duty is not fulfilled, and a harmful event occurs, the producer —corporation or individual person— may be held liable.

3.3.3. *Responsive regulation: the sandbox approach*

AI technologies may be tremendously disruptive and unpredictable. Depending on the sector, they can cause a systemic risk, demanding

⁵¹ CORTINA (2004), «Fundamentos filosóficos del principio de precaución», in *Principio de precaución, biotecnología y derecho*. Editorial Comares, p. 3-16.

a responsive and dynamic regulation. Precautionary measures can be dynamic in order to achieve a more responsive and flexible regulation. An example of dynamic regulation is the sandbox approach which aims to create a space, an environment for responsible AI development. This approach was first applied in the financial sector but soon expanded to other sectors⁵². In fact, sandbox approaches have been applied specifically in the financial and securities market. Regulatory authorities such as the *Financial Conduct Authority*⁵³ of the UK create spaces that allow new business companies to test their ideas and intelligent products. The aim of a sandbox is to create a «safe space for testing innovative products, services or business models in a controlled environment»⁵⁴. A sort of testing in the real world that allows a better knowledge of the risk of the product and therefore more promising for the development of future and adequate regulation, with proper safeguards. The new products are open to discussion, to debate and to control.

3.4. «AI crimes» (and the special part of the criminal code)

As the risks of some AI products become more known and predictable, some countries have advanced with specific crimes, related to AI, in order to protect legal interests (*Rechtsgüter*), e. g., privacy and personal data, economic and property interests, or justice fairness. We briefly present some examples.

In the context of a reform of the justice system⁵⁵, article 33 of the French law 2019-222 from the 23rd of March 2019 (*Réforme pour la justice*) has modified article L10 of the *Code de Justice Administrative*, criminalizing the use of AI to predict a court decision. The prohibition is aimed at preventing anyone —but especially legal tech companies focused on litigation prediction and analytics— from publicly revealing the pattern of judges' behaviour in relation to court decisions, avoiding what has been called «forum-shopping»⁵⁶. The new law states: «The identity data of magistrates and members of the judiciary cannot be reused with the purpose or effect of evaluating, analysing, comparing or predicting their actual or alleged professional practices»⁵⁷.

⁵² LECKENBY, DAWOUD, BOUVY, & JÓNSSON (2021), «The Sandbox Approach and its Potential for Use in Health Technology Assessment: A Literature Review», *19 Appl Health Econ Health Policy*.

⁵³ Cf. <https://www.fca.org.uk/firms/innovation/regulatory-sandbox>.

⁵⁴ European Parliament, *Regulatory Sandboxes and Innovation Hubs for FinTech*.

⁵⁵ Cf. https://www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000038261761.

⁵⁶ Cf. <https://www.artificiallawyer.com/2019/06/04/france-bans-judge-analytics-5-years-in-prison-for-rule-breakers/>; also ENGSTROM & GELBACH (2020), «Legal Tech, Civil Procedure, and the Future of Adversarialism». *University of Pennsylvania Law Review*.

⁵⁷ «Les données d'identité des magistrats et des membres du greffe ne peuvent faire l'objet d'une réutilisation ayant pour objet ou pour effet d'évaluer, d'analyser, de comparer ou de pré-

Another example comes from the US and is the anti-spoofing provision of the Dodd-Frank Wall Street Reform and Consumer Protection Act, that amended the Commodity Exchange Act's (CEA) «Prohibited Transactions». Under this statute «it shall be unlawful for any person to engage in any trading, practice, or conduct on or subject to the rules of a registered entity that is of the character of, or is commonly known to the trade as, “spoofing” (bidding or offering with the intent to cancel the bid or offer before execution)». A knowing violation of the anti-spoofing provision is a felony, carrying a maximum sentence of 10 years imprisonment and a fine of the greater of \$1 million or triple the violator's monetary gain resulting from the alleged conduct⁵⁸.

Finally, more recently, there has been a discussion about the role of criminal law in preventing the use of deepfakes (*e.g.*, image, video, voice, text) for purposes that may have criminal relevance⁵⁹, such as political disinformation, incitement to hatred, or, on an individual level, the artificial creation of intimate and private (fake) content. This will certainly be a space for the creation of new criminal offenses or the adaptation of existing ones to the specificities of AI, *e.g.*, electoral crimes, crimes against the rule of law or crimes against privacy or honor.

4. CONCLUSION

The growing autonomy and unpredictability of certain artificial intelligence systems challenge traditional legal frameworks for attributing liability, particularly in the realm of criminal law. While some of these challenges may be addressed through existing legal structures, such as civil and criminal product liability, others require novel legal adaptations to account for the computational complexity and evolving nature of so-called «intelligent» systems.

This study has highlighted the specific difficulties arising when AI systems autonomously learn and make decisions that result in unforeseeable harm. The inherent unpredictability of these technologies raises significant concerns regarding risk assessment and legal accountability. Given this uncertainty, precautionary regulatory measures—such as oversight obligations and dynamic regulation (*e.g.*, regulatory sandboxes)—play a crucial role in reducing opacity and increasing public trust in AI systems.

dire leurs pratiques professionnelles réelles ou supposées. La violation de cette interdiction est punie des peines prévues aux articles 226-18, 226-24 et 226-31 du code pénal, sans préjudice des mesures et sanctions prévues par la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés».

⁵⁸ Cf. <https://www.congress.gov/111/plaws/publ203/PLAW-111publ203.pdf>.

⁵⁹ SCHÜNEMANN (2019), «Gefährden fake news die demokratie, wächst aber im strafrecht das rettende auch?». *Goldammer's Archiv für Strafrecht*, vol. 166, núm. 10, 2019.

As knowledge about AI technology advances, legal responses should remain adaptable. The principle of revisability must guide regulatory interventions, ensuring that criminal liability is only considered when risks become sufficiently foreseeable. In this evolving landscape, public authorities must assess whether specific criminal offenses tailored to strong AI are necessary, balancing innovation with the imperative of risk mitigation and legal certainty.

5. BIBLIOGRAPHICAL REFERENCES

- ABBOTT, RYAN & SARCH, Alex (2019), «Punishing artificial intelligence: legal fiction or science fiction». *University of California, Davis*, vol. 53, p. 323-384.
- AIRES DE SOUSA, Susana (2024), «The importance of AI for financial crimes». *Revue Internationale de Droit Pénal*, vol. 95, Issue 1, 2024, p. 163-180.
- (2023), «Portuguese Report on traditional criminal law categories and AI». *Revue Internationale de Droit Pénal*, vol. 94, Issue 1, 2023, p. 319-330.
- (2017), «Neurociências e processo penal: verdade ex machina», *Estudos em homenagem ao Senhor Prof. Doutor Manuel da Costa Andrade*, vol. II, Boletim da Faculdade de Direito da Universidade de Coimbra, p. 883 ff.
- (2014), *A Responsabilidade Criminal pelo Produto e o Topos Causal em Direito Penal*. Coimbra, Coimbra Editora, 2014.
- BECK, Susanne (2013), «Über Sinn und Unsinn von Statusfragen – zu Vor- und Nachteilen der Einführung einer elektronischen Person», en HILGENDORF & GÜNTHER (coord.), *Robotik und Gesetzgebung*, Nomos, 2013.
- BANTEKA, Nadia (2021), «Artificially intelligent persons». *Houston Law Review*, vol. 58, p. 537-596.
- BATHAE, Yavar (2018), «The artificial intelligence black box and failure of intent and causation». *Harvard Journal of Law & Technology*, vol. 31, p. 889-938.
- BRICOLA, Franco (1978), «Responsabilità penale per il tipo e il modo di produzione». *La Responsabilità dell'impresa per i Danni all'ambiente e ai Consumatori*. Giuffrè Editore.
- CAPPELLINI, Alberto (2018), «Machina delinquere non potest? Brevi appunti su intelligenza artificiale e responsabilità penale», *Criminalia*, p. 499-520.
- CHARNEY, Rachel (2015), «Can Androids plead automatism? A review of when robots kill: artificial intelligence under the criminal law by Gabriel Hallevy». *University of Toronto Faculty of Law Review*, 69 (2015), p. 69-72.
- CHEN, Jiahong & BURGESS, Paul (2019), «The boundaries of legal personhood: how spontaneous intelligence can problematize differences between humans, artificial intelligence, companies and animals». *Artificial Intelligence and Law*, 27, p. 73-92.
- CORTINA, Adela (2004), «Fundamentos filosóficos del principio de precaución», in *Principio de precaución, biotecnología y derecho*. Editorial Comares, p. 3-16.
- DIAMANTIS, Mihailis E. (2020), «The extended corporate mind: when corporations use AI to break the law». *North Carolina Law Review*, vol. 98, 2020, p. 894-932.

- ENGSTROM, David Freeman & GELBACH, Jonah B. (2020), «Legal Tech, Civil Procedure, and the Future of Adversarialism». *University of Pennsylvania Law Review*, available at SSRN: <https://ssrn.com/abstract=3551589>.
- FISHER, Elizabeth; JONES, Judith & VON SCHOMBERG, René (eds.) (2006), *Implementing the Precautionary Principle. Perspectives and Prospects*, Edward Elgar.
- FLORIDI, Luciano (2017), «Robots, Jobs, Taxes, and Responsibilities». *Philosophy and Technology*, 30, p. 1-4.
- GIUFFRIDA, Iria (2019), «Liability for AI Decision-Making: some legal and ethical considerations». *Fordham Law Review*, vol. 88, p. 439-456.
- GLESS, Sabine (2020), «AI in the courtroom: a comparative analysis of machine evidence in criminal trials». *Georgetown Journal of International Law*, vol. 51, p. 195-253.
- GLESS, Sabine; SILVERMAN, Emile & WEIGEND, Thomas (2016), «If robots cause harm, who is to blame? Self-driving cars and criminal liability». *New Criminal Law Review*, vol. 19, N. 3, p. 412-436.
- GOODMAN, B. & FLAXMAN, S. (2017), «European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation”». *AI Magazine*, 38(3), 50-57.
- HALLEVY, Gabriel (2013), *When Robots Kill*. Northeastern University Press, 2013.
- (2010), «The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control». *Akron Intellectual Property Journal*, vol. 4, Issue 2 (2010), p. 171-201.
- HILGENDORF, Eric (2017), «Automated driving and the law». *Robotics, Autonomics and the Law*, Nomos.
- LECKENBY, Emily; DAWOUD, Dalia; BOUVY, Jacoline & JÓNSSON, Páll (2021), «The Sandbox Approach and its Potential for Use in Health Technology Assessment: A Literature Review», *19 Appl Health Econ Health Policy* available at: <https://doi.org/10.1007/s40258-021-00665-1>.
- KING, AGGARWAL, TADDEO & FLORIDI (2020), «Artificial Intelligence crime: an interdisciplinary analysis of foreseeable threats and solutions». *Science and Engineering Ethics*, 26, p. 89-120.
- PAGALLO, Ugo (2013), *The Laws of Robots*. Springer, 2013.
- PIERGALINI, Carlo (2004), *Danno da Prodotto e Responsabilità Penale*, Giuffrè Editore.
- SAMEK, Wojciech; WIEGAND, Thomas & MÜLLER, Klaus-Robert (2017), «Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models». *ArXiv abs*, <https://arxiv.org/abs/1708.08296>.
- SCHÜNEMANN, Bernd (2019), «Gefährden fake news die demokratie, wächst aber im strafrecht das rettende auch?». *Goldammer's Archiv für Strafrecht*, vol. 166, núm. 10, 2019.
- SIMMLER, Monika & MARKWALDER, Nora (2019), «Guilty Robots – Rethinking the nature of culpability and legal personhood in an age of artificial intelligence». *Criminal Law Forum*, Springer, 1-31.
- SULLIVAN, Hannah R. & SCHWEIKART, Scott J. (2019), «Are current tort liability doctrines adequate for addressing injury caused by AI». *AMA Journal of Ethics*, vol. 21, N. 2, p. 160-166.
- TURING, Alan, (1950), «Computing Machinery and Intelligence». *Mind*, vol. LIX, núm. 236, p. 433-450.

- TURNER, Jacob (2019), *Robot Rules. Regulating Artificial Intelligence*, Palgrave Macmillan,
- WENG, Yueh-Hsuan; CHEN, Chien-Hsun & SUN, Chuen-Tsai (2009), «Toward the Human–Robot Co-Existence Society: On Safety Intelligence for Next Generation Robots». *International Journal of Social Robotics* (2009), p. 267-282.

Other References

- The one hundred year study on artificial intelligence*, Stanford University, 2016.
- World Commission on the Ethics of Scientific Knowledge and Technology*, Report of COMEST on robotics ethics, available at <https://unesdoc.unesco.org/ark:/48223/pf0000253952>.
- The Precautionary Principle*, United Nations Educational Scientific and Cultural Organization, COMEST, 2005.
- Regulatory Sandboxes and Innovation Hubs for FinTech*, European Parliament available in [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652752/IPOL_STU\(2020\)652752_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652752/IPOL_STU(2020)652752_EN.pdf).

