

The Recognition of Legal Personhood for Autonomous AI Systems

Anna Moraiti

Abstract: The design and deployment of autonomous AI systems poses significant legal challenges, not least in the area of criminal law, as their inherent complexity and opacity hinders the connection of a harm with the acts or omissions of numerous involved agents, that may be natural or corporate persons. This paper explores the proposition of a new model of criminal liability, which would recognise certain highly sophisticated AI systems as proper subjects under criminal law, given the potential difficulty, or even inability, to identify any other culpable agent. Proponents of this model often make a comparison with the model of corporate criminal liability, arguing that since corporations are being subjected to criminal sanctions, the foundation for an expansion of the scope of application of substantive criminal law to other types of artificial agents is established. However, this paper argues that the relevant claim is not self-evident, as criminal responsibility implies autonomy, whereas the inverse is not necessarily the case. Personhood for the purposes of criminal law has been linked to the ability for reflection, that is, the ability to distinguish right from wrong and freely choose the former, as only someone who could choose otherwise may be blamed for committing a crime.

In this context, this analysis briefly traces the notion of criminal personhood, and comments on the application of its aims and rationales to the case of autonomous AI systems. Moreover, it emphasises that different branches of law may espouse different substantive definitions of autonomy, in accordance with the varying goals that are pursued in each individual case. In order to illustrate this point, this contribution makes a contrast with the content of legal personhood under civil law, which can be regarded as a 'cluster property', meaning that an entity may qualify as a legal person not because of its qualities, but because it unites a number of incidents. The different aims underpinning these two branches of law inevitably point towards different potential legal regimes, as well as towards a nuanced approach when it comes to the conceptualisation of legal personhood for autonomous AI systems.

1. Introduction

Are modern AI systems more accurately described as products, tools, or objects, or as persons?¹ This general question cannot be answered unequivocally. Although in everyday usage the distinction between a person and a thing is a fundamental one, and appears to be clearly delineated for practical purposes, legal personhood is nuanced and blurs the lines of this classification. It is a socially constructed concept, and can be conferred for purposes that are deemed to be important for the function of the legal system.² Ultimately, one cannot provide a meaningful answer to the question posed above without taking into account the nature of legal personhood and the specific purposes served by the conferral of personhood under each branch of law.

The more specific issue that is concisely explored in this paper is whether AI systems that qualify as being autonomous may be equipped with onerous legal personhood, in a manner similar to corporations, and thus be legitimately held liable for harm stemming from their operation.³ In fact, both corporate criminal liability and the debated criminal liability of AI systems raise doctrinal questions that are directly related to the understanding of the culpability principle. These questions cannot be ignored in view of the rapid developments registered within modern risk societies, and the need to identify and punish culpable conduct in cases of severe harm.⁴

Before delving into the exploration of this comparison, it is necessary to provide a concise definition of the basic terms related to AI systems that are vested with a degree of autonomy. In fact, there is no generally accepted definition of the term 'AI', and this broad concept may refer to a number

-
- 1 The question explored in the present contribution forms part of the author's doctoral dissertation. The author would like to thank Assoc. Prof. Niovi Vavoula for her valuable comments on the working draft of this article, which were provided in the context of the workshop 'Navigating the Future Frontiers of AI Law and Governance', organised by the University of Luxembourg on 17–18 October 2024 (supported by the Luxembourg National Research Fund PRIDE19/14268506).
 - 2 David J Gunkel, *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond* (MIT Press 2023) 18–19.
 - 3 Visa AJ Kurki, 'The Legal Personhood of Artificial Intelligences', in *A Theory of Legal Personhood* (Oxford University Press 2019) 176; Simon Chesterman, *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law* (Cambridge University Press 2021) 115.
 - 4 Ulfrid Neumann, 'Structure and Regulatory Content of the Principle of Guilt in Criminal Law' (2020) *Penal Chronicles* 481.

of learning techniques and uses. In most modern analyses, the use of this term refers to systems that operate on the basis of machine learning methods.⁵ Machine learning can be defined as the science of detecting reliable patterns in analysed data, with the goal of producing accurate predictions.⁶ It is attached to systems with very high computational power and renders them capable of improving over time from their own experience, which is based on enormous training data sets.⁷

The common feature of such systems is that they appear to have developed some degree of agency, in the sense that they can interact with and effect changes in their environment while modifying their internal states without being guided by human interventions.⁸ In this respect, the autonomous or emergent properties of an AI system appear as its complexity increases, leading to unpredictable courses of behaviour and outputs. Autonomy itself is an elusive concept, as it is not a binary classification, but it can more accurately be described as existing on a spectrum.⁹ The attributes that one may take into account — and each of them in various degrees — are the adaptability to changing conditions, the potential for gaining new knowledge, the freedom to act without human involvement and the ability to replace human actors.¹⁰ This means that different systems may qualify as autonomous without always being vested with all of its features.

Lastly, it is necessary to clarify the distinction between weak and strong AI. The term ‘strong AI’ generally denotes a hypothetical system that has either attained or surpassed human intelligence, which means that to qualify

5 Carla L Reyes, ‘Autonomous Corporate Personhood’ (2021) 96 *Washington Law Review* 1453, 1463.

6 Nello Cristianini, *The Shortcut: Why Intelligent Machines Do Not Think Like Us* (First edition, CRC Press 2023) 42.

7 Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (Alfred A Knopf 2017) 82–.

8 By contrast, in rule-based AI it is relatively obvious why the system behaves in one way instead of another, as its programmers have previously set guidelines that are followed by the system; Crina Grosan and Ajith Abraham, *Intelligent Systems – A Modern Approach* (Springer 2011) 149–.

9 Karni Chagal-Feferkorn, ‘When Do Algorithmic Tortfeasors That Caused Damage Warrant Unique Legal Treatment?’ in Woodrow Barfield (ed), *The Cambridge Handbook of the Law of Algorithms* (Cambridge University Press 2020) 474.

10 For an account of various concepts of autonomy, see Giovanni Sartor and Andrea Omicini, ‘The Autonomy of Technological Systems and Responsibilities for Their Use’ in Nehal Bhuta and others (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press 2016).

as such, the system must have developed an equally broad or even broader problem-solving capacity and adaptability to external conditions. In the latter case, one would refer to the concept of Artificial General Intelligence (hereinafter: AGI), or Superintelligence.¹¹ The term ‘weak/narrow AI’ denotes non-conscious systems that are based on machine learning methods and are trained on big data. They possess the ability to execute specific tasks, surpassing human capacities while doing so, but without claiming either a general problem-solving capacity or an infinite adaptability to changing conditions.¹²

In the following, autonomous systems are understood as falling under the broad and approximative category of weak AI. Of specific interest are robotic systems that pose a threat for human bodily integrity and life, so this category may include applications such as autonomous vehicles, care robots, and other humanoid robots that are powered by AI and interact with the physical environment. Therefore, for present purposes, the term will encompass robotic systems that employ any machine learning method and intuitively seem to serve as something more than a simple tool, when their performance in the physical world acts as a substitute for actions that have been typically undertaken by natural persons and creates a risk of physical harm against humans.¹³

The present analysis proceeds by summarising some arguments that have been advanced in favour of a regime that would directly punish AI systems (section 2). These arguments are then discussed in light of the content that is uniquely associated with criminal personhood, the goals that are specifically served by criminal justice systems, and the nature of intelligence and morality that is presently exhibited by AI systems (section 3). Next, the

11 See Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014) 22, who defines Superintelligence as ‘any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.’

12 Having created such systems is, of course, no small achievement, and it develops far reaching implications for contemporary societies. See, in this direction, Yuval N Harari, *Homo Deus: A Brief History of Tomorrow* (Harper 2017) 308. As the author notes, ‘[h]umans are in danger of losing their value, because intelligence is decoupling from consciousness. Until today, high intelligence always went hand in hand with a developed consciousness. [...] However, we are now developing new types of non-conscious intelligence that can perform [...] tasks far better than humans. For all these tasks are based on pattern recognition, and non-conscious algorithms may soon excel human consciousness in recognising patterns. This raises a novel question: which of the two is really important, intelligence or consciousness?’

13 Reyes (n 5) 1480.

potential of legal personhood for civil law purposes is briefly discussed (*section 4*) and some final thoughts with respect to this comparative perspective are articulated (*section 5*).

2. Central Arguments in Favour of Punishing Autonomous AI Systems

There are various examples that one could invoke in order to illustrate how severe physical harm could be inflicted through the operation of hardware systems that are powered by AI. Notably, in 1981, an industrial robot caused the death of an employee at a motorcycle factory based in Japan. The robot killed the engineer by pushing him against adjacent machinery as it had not been deactivated properly. This example has been associated in the relevant literature with the discussion on whether, and if so, under which conditions, AI systems that qualify as being autonomous should be punished for the harm caused by their malfunction.¹⁴ In a different setting, a police robot operating autonomously could illegally cause bodily harm to participants in a protest.¹⁵ A robot deployed as a guard within a prison could seriously injure an inmate because it misinterpreted their interaction or miscalculated the level of risk involved.¹⁶ Similarly, a care robot could fail to inform the patient under its supervision to take their medication at the right time, which may be found to be causally linked to the patient's death. The question linked to such scenarios is whether an AI system that has been programmed so as to weigh the expected benefits and drawbacks of different courses of action could be punished on occasion, when its 'choice' of the option that is associated with the highest expected value contravenes with legal provisions.¹⁷

Gabriel Hallevy is one of the most prominent supporters of a new model of criminal liability for AI systems, and the first scholar to extensively approach this issue.¹⁸ In his work, he conceptualises a third subject of punishment (in addition to humans and corporations) that he refers to

14 Gabriel Hallevy, *When Robots Kill: Artificial Intelligence under Criminal Law* (North-eastern University Press 2013) 38.

15 Kamil Mamak, *Robotics, AI and Criminal Law: Crimes Against Robots* (Routledge 2023) 126.

16 Hallevy (n 14) 18.

17 Kurki, 'The Legal Personhood of Artificial Intelligences' (n 3) 180–181.

18 For a complete development of his views, see Gabriel Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems* (Springer International Publishing 2015).

as 'delinquent thinking machine' or 'machina sapiens criminalis.' Such a machine need not be so technologically advanced so as to reach the capabilities of AGI.¹⁹ In Hallevy's view, current criminal law doctrine can be directly applied to narrow AI systems because the latter exhibit rational qualities and are designed so as to imitate the function of the human mind. Additionally, he argues that since corporations have been subjected to criminal sanctions for centuries, the paradigm for such an expansion of the scope of application of substantive criminal law is already sufficiently developed.²⁰

More specifically, the author asserts that not only AI systems, but also robotic systems that are not powered by AI are capable of acting or omitting for criminal law purposes. The requirement of an act is met by any movement registered in the physical world, and this may stem either from the autonomous operation of the system or from its manipulation by a human operator.²¹ Hallevy advances a thin definition of the criminal act as he notes that 'criminal law considers as an act any material performance with a factual-external presentation, whether willed or not.'²² If a legal duty to act that is specifically addressed to the machine exists, an omission can be verified each time that the system fails to undertake the required action.²³ When it comes to the establishment of the mental element, the author claims that as AI systems can perceive the conditions of their environment through cameras, microphones and sensors they can make assessments, decide based on evaluations and develop goal-driven conduct that constitutes the artificial equivalent of what is defined as human intentionality for criminal law purposes.²⁴ Similar considerations are put forth with regard to criminal negligence, as Hallevy supports that a machine could potentially take unreasonable risk in a way that largely resembles human risk-taking.²⁵

Essentially, Hallevy compares the way that machine learning functions with the role that practical experience assumes in the internalisation of legal values by humans — and, by extension, corporate persons. In his view, the system is perpetually making connections between learnt and novel factors in its environment and is generalising based on all the data processed, thus

19 *ibid* 24.

20 *ibid* 39.

21 *ibid* 61–62.

22 Hallevy (n 14) 52.

23 Hallevy (n 18) 63.

24 *ibid* 97.

25 *ibid* 127.

accumulating useful experience. Having argued in favour of this new model of criminal liability, the author subsequently addresses the potential purposes of punishment that could be served by such a regime and notes that since AI systems cannot experience suffering or intimidation, the purposes of retribution and deterrence would not be fulfilled, but rehabilitation and incapacitation could be advanced as the system would learn how to refine and enhance its decision-making processes and the imposition of criminal sanctions would ensure that causing further harm is practically impossible.²⁶

However, it must be noted that the author applies a different definition for strong AI compared to the one advanced in the context of this contribution. Hallevy defines strong AI as a system that merely works on the basis of assessing probabilities, which means that perhaps he conflates the notions of strong and weak AI, the latter covering the category of present-day, machine learning systems that are currently being regulated. But even if the definition is intentionally constructed this way, the author hastily bridges the relevant doctrinal gap by directly comparing the decision-making processes of machine learning systems and the way that humans and corporate persons weigh different factors and make evaluative choices. Whereas indeed, as he notes, both humans and machine learning systems can learn inductively, this is not the only way in which the human mind can function. A fundamental feature of human reasoning rests on the ability not only to understand general concepts and abstract principles, but also to apply them in concrete and diverse factual contexts.²⁷

3. The Misplaced Comparison Between Corporations and AI Systems

3.1 The Risk of Performing an Empty Imputation

Both corporate criminal liability and the debated criminal liability of AI systems constitute manifestations of the same broader phenomenon, which is the liability shift that seems to take place when accountability for harmful outcomes is diverted to a complex network, made up of many natural and

²⁶ *ibid* 211; Hallevy (n 14) 160–161.

²⁷ For the elaboration of this argument see Hallevy (n 18) 100.

artificial agents.²⁸ It would not be implausible to revise criminal law doctrines in order to allow for the criminal liability of AI systems: this regime could apply for a more or less extensive array of scenarios and it could be combined with the personal liability of the producers or professional users of the system.²⁹ Drawing from the model of corporate criminal liability, the doctrinal tool that may seem relevant is *respondeat superior*, which imputes mental states of the corporate employees to the corporate person as long as their conduct falls within the scope of their employment duties.³⁰ Similarly, it could be that the mental states of the producers or users of the AI system would be imputed to the latter, which would be recognised as a person for criminal law purposes, vested with a type of 'electronic personhood'. In this manner, it could be sanctioned by fines or confiscation of assets.³¹

However, one could seriously question whether the recourse to such a measure would fulfil the unique purposes that are associated with the function of criminal law, as well as whether the paradigm of corporate criminal liability can be a direct point of reference for such a significant expansion of the scope of substantive criminal law. Hallevy seems to underestimate the conceptual difficulties involved:

'Artificial intelligence technology has the capability to fulfill the requirements of criminal law with no single change of these definitions. [...] How come that these very definitions used by criminal law systems around the world for centuries were adequate and almost non-questionable, and suddenly they became "shallow"? If they are good enough for humans and corporations, why would not they be good enough for artificial intelligence technology?'³²

The fundamental difference in this case is that corporations are composed of individuals, who are indirectly targeted by sanctions imposed against the corporation, whereas AI systems constitute a manifestation of human

28 Susanne Beck, 'Corporate Criminal Liability' in Markus D Dubber and Tatjana Hörnle (eds), *The Oxford Handbook of Criminal Law* (Oxford University Press 2014) 568.

29 Chesterman (n 2) 116; Kurki, 'The Legal Personhood of Artificial Intelligences' (n 2) 179.

30 Abigail H Lipman, 'Corporate Criminal Liability' (2009) 46 *American Criminal Law Review* 359, 360.

31 Chesterman (n 3) 123.

32 Hallevy (n 18) 105.

creativity.³³ In addition, it should be noted that the diverse individuals or corporations that may be at fault in cases of harm associated with the function of AI systems may not be necessarily affected by sanctions exclusively imposed against these systems. In Kelsen's terms, 'the sanction is applied to individuals, because human beings only can be the objects of the sanction, the victims of forcible deprivation of life and freedom. But the sanction is applied to them not individually, but collectively. That a sanction is directed against a juristic person means that collective responsibility is established of the individuals who are subject to the total or partial legal order personified in the concept of juristic person.'³⁴

For criminal law purposes, the aggregate theory of the corporate person claims an important position; under this theory, the corporation does not possess an independent identity, but is seen as a symbol for the collaboration of individuals, who constitute the fundamental unit of interest.³⁵ The intense interaction and the dynamics developed between the corporate person and its members — at a structured, permanent and organised level — constitute an essential element of this theory of corporate legal personhood, which is missing in the case of AI systems.³⁶

Whereas it is possible to pierce the corporate veil and punish the directing minds behind the corporate person, the clear-cut choice in most future cases would be to attribute blame either on the system itself, or on its operators.³⁷ This would be so because an AI system, as such, is too opaque to be referred to as a symbol for the collective efforts of the persons who have collaborated in its production and use, especially taking into account that these collaborations typically involve multiple corporations and/or individual freelance programmers. Would it be meaningful at all to impose criminal sanctions on such a type of agent? In other words, the so-called 'problem of many hands' may exist in corporate structures as well, but it is considerably exacerbated in the case at hand. Conferring criminal agency to AI systems, and rendering them the sole addressees of criminal sanctions, would allow powerful individuals holding key positions within corporate structures to avoid taking the blame for harmful outcomes. This

33 Chesterman (n 2) 124, 126.

34 Hans Kelsen, *General Theory of Law and State* (Transaction Publishers 2006) 106.

35 Beck (n 28) 576.

36 Other scholars have rightly observed that this theory cannot be readily invoked in the case of AI systems; see Chesterman (n 3) 118.

37 In this direction, see David Runciman, *The Handover: How We Gave Control of Our Lives to Corporations, States and AIs* (Profile Books 2023) 67.

has occasionally been the case with sanctions that have been exclusively directed against corporate persons.³⁸

3.2 The Risk of Altering the Nature of Criminal Responsibility

In most jurisdictions, criminally responsible persons may be humans, qualifying as autonomous and reasons-responsive, and the corporations that the latter can form. In both cases, one deals with agents that can challenge the State's exclusive exercise of legislative power.³⁹ According to the rule of law account of legal personhood, only entities that are capable of genuinely exhibiting wrongful conduct can be regarded as subjects of criminal punishment. Corporate persons may be perceived as belonging to this category only if they are regarded as a symbol for the union of members who possess legal personhood or, in other words, as a genuine collectivity of individuals.⁴⁰

It would not be possible to hold AI systems liable under such a high standard of criminal responsibility for two reasons. First, because this standard requires that the subject of punishment understand what criminal wrongdoing entails, the impact of their conduct on the state's exclusive authority to legislate, and its significance for the stability of the social order.⁴¹ While modern AI systems have some computational knowledge on their status and their internal states, this is not meaningful enough for accountability mechanisms to be activated with regard to their proper conduct.⁴² And second, because their conduct is jointly determined by programming policies and the autonomous development of the system following machine

38 *ibid* 270; Chesterman (n 3) 121; Joanna J Bryson, Mihailis E Diamantis and Thomas D Grant, 'Of, for, and by the People: The Legal Lacuna of Synthetic Persons' (2017) 25 *Artificial Intelligence and Law* 273, 285–.

39 Malcolm Thorburn, 'In Search of Criminal Law's Person', *The Criminal Law's Person* (Hart 2022) 101.

40 *ibid* 108.

41 *ibid* 115–116, 117. See, in a similar direction, Beatriz A Ribeiro and others, 'Metacognition, Accountability and Legal Personhood of AI' in Henrique Sousa Antunes and others (eds), *Multidisciplinary Perspectives on Artificial Intelligence and the Law*, vol 58 (Springer International Publishing 2024) 178–179. The authors argue that AI systems cannot be held accountable because they lack strategic and monitoring capacities, which are necessary in order to understand criminal wrongdoing and reflect upon the nature of one's conduct.

42 Ribeiro and others (n 41) 180.

learning methods, which does not allow them to be considered as a genuine collectivity. Whatever emergent properties they may have, these do not originate in the interaction of individual members; and while this may not be inconsistent with the recognition of legal personhood in other domains, it is substantive for criminal law purposes.⁴³ Therefore, the claim that the definition of intent for criminal law purposes is narrower compared to the conceptions supported in other disciplines (e.g., in philosophy and psychology) seems to oversimplify the content and purpose of criminal personhood.⁴⁴

The argumentation in favour of the recognition of criminal personhood for AI systems is moved by two considerations: first, by efficiency considerations, as humanising them would, allegedly, contribute in their smooth incorporation to many sectors of societal activity and in the alleviation of fears regarding the possibility of unforeseen outcomes.⁴⁵ Second, by the need to ensure at all times that accountability is maintained, especially in cases of extensive harm that are of interest to the society at large and may present evidentiary difficulties as to the identification of culprits. Both aims already exist behind the concept of corporate personhood, and this certainly favours the comparison between the two cases.⁴⁶ However, they do not suffice to set aside the crucial differences between the criminal law's person (which includes corporate persons) and modern AI systems.⁴⁷ As Paul Burgess accurately frames this issue:

‘The motivators that have traditionally been used to hold power-wielding entities accountable have been human-centered. These may include removal from power, sanctions, fines, imprisonment, or reputational damage. There seems to be no immediate way to make a machine accountable in this way [...] On this basis, the Future Rule of Law would require the inclusion of a human-in-the-loop to *be* accountable. This is not necessary for fear of ineffective technology or malfunction, but

43 In the same direction, see Bert Jaap Koops, Mireille Hildebrandt and David-Olivier Jaquet-Chiffelle, ‘Bridging the Accountability Gap: Rights for New Entities in the Information Society?’ (2010) 11 *Minnesota Journal of Law, Science & Technology* 497, 560.

44 See in this respect Hallevy (n 8) 98ff.

45 Hallevy (n 14) 22; Anna Beckers and Gunther Teubner, *Three Liability Regimes for Artificial Intelligence: Algorithmic Actants, Hybrids, Crowds* (Hart 2021) 26.

46 Reyes (n 5) 1493.

47 In this direction, see Runciman (n 37) 100.

to ensure that the exercise of power has a human actor that is always accountable.⁴⁸

One of Hallevy's basic premises is that criminal law doctrine should not change the course of technological developments, but instead, it should adapt its basic concepts and categories to fit the latest technological developments.⁴⁹ This need is put to the forth as it becomes clear that difficulties in ensuring accountability arise in both human-on-the-loop scenarios, in which the AI system has been delegated a substantive amount of autonomous decision-making and humans can occasionally intervene after the decision has been taken, and human-out-of-the-loop scenarios, in which the possibility of meaningful human intervention is excluded.⁵⁰ Criminal law's function has in former times focused on the identification of blameworthy individuals; in the course of time, it seems to prioritise the resolution of conflicts that implicate the public at large.⁵¹ Proponents of this new model of criminal liability seem to endorse the latter view on the role and function of criminal justice systems.

Against this background, it becomes clear that the answer to the question of transposing or not the model of corporate criminal liability to another type of artificial agent reveals a specific understanding of criminal punishment. Taking into account modern AI systems, it would be difficult to bridge the doctrinal gaps involved without affecting the nature of criminal responsibility. As it has been argued also with respect to criminal sanctions that have been imposed against corporations, the discussed regime may amount to the activation of a tool of last resort for the satisfaction of civil claims.⁵² Such a development would further reinforce the tendency towards using the criminal law as an instrument for compliance goals, whereas in reality, the excessive deployment of criminal provisions renders this branch of law ineffective.⁵³

48 Paul Burgess, 'The Future Rule of Law and AI: The Necessary Evolution of a Concept', *AI and the Rule of Law: The Necessary Evolution of a Concept* (Bloomsbury Publishing 2024) 171.

49 Hallevy (n 18).

50 Burgess (n 48) 157.

51 Beck (n 28) 582.

52 Thomas Weigend, 'Societas Delinquere Non Potest? A German Perspective' (2008) 6 *Journal of International Criminal Justice* 927, 944.

53 *ibid* 945.

While this may be so, it is also true that the criminal law, constituting the most powerful instrument of state coercion, should adapt, to a certain extent, to technological developments so as to continue to fulfil its double protective function. A balance achieved between two parameters — the protection of citizens' fundamental legal interests, on the one hand, and the protection against the arbitrariness of State power, on the other hand — would legitimate new legislative initiatives in this respect.⁵⁴ Such a balance can be achieved when new criminal law provisions aim at punishing culpable conduct that is related to technological developments without setting aside fundamental criminal law categories, such as the culpability principle.⁵⁵

3.3 On the Morality and Intelligence of AI Systems

Admittedly, modern AI systems develop a certain degree of moral agency, as they can produce effects in their environment and these effects often have moral consequences for human interests.⁵⁶ If rationality, intentionality, and communicative capacity have been identified as essential features of personhood, then a thinking AI system could qualify as a person.⁵⁷ Such a speculative scenario has been eloquently described in Ian McEwan's novel, titled *Machines Like Me*. One of the main characters of this story is a humanoid robot named Adam, who is equipped with a sense of morality by design as well as a personality that is initially shaped by his owners but is then further developed by his interaction with the environment and other people. The book largely describes the robot's decision-making processes and practical choices, that are determined by the interaction between his inbuilt moral code and his personal inclinations.⁵⁸ It is shown that the

54 Agapios Papanefytou, *Criminal Liability of Legal Persons: A Construction of Criminal Immunity for Corporate Offenders?* (Nomiki Vivliothiki 2012) 130.

55 *ibid* 137.

56 Mark Coeckelbergh, *AI Ethics* (MIT Press 2020) 50.

57 Susanna Kim Ripken, 'Philosophical Dimensions of the Corporate Person', *Corporate Personhood* (Cambridge University Press 2019) 62.

58 Ian McEwan, *Machines Like Me* (Nan A Talese/Doubleday 2019) 73. As explained by the narrator and one of the two owners of the robot, '[s]oftware engineers from the automobile industry may have helped with Adam's moral maps. But together we had contributed to his personality. I didn't know the extent to which it intruded on, or took priority over, his ethics. How deep did personality go? A perfectly formed moral system should float free of any particular disposition.'

clash of such powerful drives in self-conscious artificial agents could lead to unpredictable outcomes that may severely harm human interests.

However, the notion of AGI does not correspond to a tangible possibility as of yet. Modern AI systems rely on the recognition of probabilities, and require a massive amount of training data in order to extract patterns from such probabilities. It is true that they are extremely powerful, as they can process enormous amounts of data in minimal time frames, functioning between abstraction and enumeration. Nonetheless, at no point during their operation do these systems develop a literal understanding of the meaning behind this data.⁵⁹ In this respect, one should not overestimate the value of the knowledge produced by machine learning algorithms, as this can be proved erroneous because of false correlations, with material factors having been misevaluated, while irrelevant factors may have been considered as significant for the purposes of the system's task.

The discussion on AI systems' potential for legal personhood is, to a large extent, driven by their perceived potential to develop intelligent behaviour. In fact, intelligence should not be conceived in narrow terms: there is more than one kind of intelligence, and as many AI researchers emphasise, this is not an exclusively human feature.⁶⁰ The Turing test, introduced in 1950, relies on a benchmark that refers to human intelligence, as it assesses the system's ability to exhibit a level of performance in cognitive tasks that so resembles human performance, that the human evaluator cannot discern whether the task is executed by a human or a machine. However, it is noteworthy that this test was designed at the time with the goal of providing an operational definition of intelligence, and it is no longer referred to as an accurate measure for AI systems' intelligent conduct.⁶¹

At present, the registered developments in the field of AI call for a reference to more neutral definitions of intelligence. In general terms, one may neutrally define 'intelligence' as the ability to develop purposeful, goal-oriented behaviour. The ability for such behaviour is demonstrated by the adaptation to new situations and task environments, where different agents may co-exist.⁶² In other words, an AI agent can be regarded as intelligent

59 Nello Cristianini, *The Shortcut: Why Intelligent Machines Do Not Think Like Us* (CRC Press 2023) 19, 36.

60 Coeckelbergh (n 56) 64; Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press 2021) 6.

61 AM Turing, 'Computing Machinery and Intelligence' (1950) LIX (236) *Mind* 433–.

62 Cristianini (n 5) 5; Beatriz A Ribeiro and others, 'Metacognition, Accountability and Legal Personhood of AI' in Henrique Sousa Antunes and others (eds), *Multidis-*

simply by means of a limited, task-oriented resilience, which could include the consideration of legal provisions during the process.⁶³ Direct comparisons between human and artificial agents are not necessarily fruitful. As AI agents cannot process emotions and abstract concepts, humans cannot detect patterns in data sets or perform large-scale computations; yet, applying this definition, both agents behave intelligently.⁶⁴ Different entities may set goals that are fundamentally different, both in nature and complexity. And it is noteworthy that autonomous AI systems are already capable of producing results that in many respects are beyond human capabilities and comprehension.⁶⁵

This standard of intelligence is thus in principle met by modern, narrow AI systems, but it falls short of the high standard associated with criminal agency, which entails the ability to perform purposeful conduct as a response to the right or wrong reasons. Furthermore, any elaboration on future, fully-fledged AI agents falls against the impossibility (at least with reference to current data) of conceptualising a full artificial moral agent, that may completely diverge from the typical example of a human moral agent.⁶⁶ As David Runciman observes in this respect, '[i]t is not clear whether AGI would require anything we would recognise as consciousness to function. Perhaps it will be possible for an AI system to move across the full range of intelligent reflection without being self-reflective in anything like a human sense.'⁶⁷

In view of the above, the next logical question is whether there are alternative measures that should be prioritised over the creation of another fiction under criminal law, or, in other words, whether the legal personhood of AI systems can be constructed in a different way, for different legal purposes. This question is concisely addressed in the following section.

ciplinary Perspectives on Artificial Intelligence and the Law, vol 58 (Springer International Publishing 2024) 172–173.

63 Kurki, 'The Legal Personhood of Artificial Intelligences' (n 3) 181.

64 *ibid* 12, 61.

65 Beckers and Teubner (n 45) 40.

66 Cristianini (n 6) 3–4.

67 Runciman (n 37) 84.

4. The Recognition of Legal Personhood for AI Systems Under Civil Law

4.1 On the Bundle Theory of Rights

In the context of this analysis, it is neither possible nor material to elaborate on the content of the many and diverse conceptions of legal personhood. It is noteworthy, though, that there is no definitive set of properties for the formulation of this definition, and this recognition may largely depend on the finding that an entity occupies an important social standing.⁶⁸ As it has been noted already, the notion of personhood is not identical for all branches of law. Therefore, defining the specific purposes that are served by such recognition in each case amounts to a fundamental and essential task.

More specifically, in order to approach the personhood of artificial entities, it should be taken into account that this may exist along a spectrum.⁶⁹ Different jurisdictions hold different conceptions of personhood, and these can be further modified in the course of historic time.⁷⁰ According to the bundle theory of legal personhood, that is advanced by Visa Kurki, an entity can qualify as a legal person regardless of its intrinsic qualities.⁷¹ This theory places the emphasis on the entity's capacity to hold legal entitlements and burdens, and legal personhood may be granted regardless of more substantive considerations.⁷² Legal personhood is conceived of as a cluster property — not a clear-cut, binary classification — and thus it is formed of many different incidents which may be combined in different ways, without any one of them being individually decisive for this recognition.⁷³

68 Chesterman (n 3) 119; Gunkel (n 2) 19; Nadia Banteka, 'Legal Personhood and AI: AI Personhood on a Sliding Scale' in Ernest Lim and Phillip Morgan (eds), *The Cambridge Handbook of Private Law and Artificial Intelligence* (Cambridge University Press 2024) 621.

69 Ripken (n 57) 92.

70 Penny Crofts, 'Crown Resorts and the Im/Moral Corporate Form' in Elise Bant (ed), *The Culpable Corporate Mind* (Bloomsbury Publishing 2023) 61.

71 Visa AJ Kurki, 'The Incidents of Legal Personhood', *A Theory of Legal Personhood* (Oxford University Press 2019) 91.

72 Burgess (n 48) 167.

73 Kurki (n 52) 93–94, 121. As the author explains, '[...] because of the cluster-property nature of legal personhood, a binary question about X's personhood or nonpersonhood may not always be appropriate. One must often rather ask about the respects in which X is a legal person. Corporations are legal persons for most purposes, but not all purposes.' For an argumentation in favour of the recognition of personhood along a spectrum, see also Banteka (n 68). More specifically, the author argues that

These incidents of legal personhood can be either passive or active. In general, the passive incidents of legal personhood concern an entity's capacity to benefit from fundamental legal protections, the capacity to own property, to undergo harm from a legal point of view and to claim a victim status under criminal law (legal standing). Active incidents are associated with the entity's capacity to enter into contractual relations (legal competences), as well as the capacity to be subjected to civil or criminal liability (onerous legal personhood). In other words, passive legal personhood is confirmed when an entity becomes entitled to claim-rights (with the status of children constituting the main example), whereas active legal personhood is linked to the exercise of legal competences and the capacity to bear duties (as is typically the case with most adults).⁷⁴

More specifically for the issue at hand, the notion of 'onerous legal personhood' refers to the entity's capacity to hold legal duties and to be held accountable for one's acts. A failure to fulfil such duties is penalised, and sanctions are enforced against this subject.⁷⁵ This is precisely the element that forms the essence of the criminal law's person, and it is connected to other incidents of personhood: for instance, corporations can be held liable under civil and criminal law because of their capacity to own proper funds.⁷⁶ According to the account presented here, an entity can legitimately bear legal duties provided that it is able to act and suffer detriment; additionally, it must be possible to communicate legal requirements to the entity in question.⁷⁷

It is crucial to note that these dimensions of personhood can be realised to various degrees. In other words, an entity may be legitimately held responsible in only certain respects, or be endowed with legal personhood without being considered as directly accountable.⁷⁸

Applying the bundle theory of legal personhood means that AI systems could be endowed with a limited, passive legal personhood, and this for reasons that have nothing to do with the system's worth, value, or reasons-

the scope of legal personhood should be more restricted the more autonomous the AI system actually is, so that its operators continue to monitor its safety parameters and to assume liability for any damages caused.

74 Visa AJ Kurki, 'Who or What Can Be a Legal Person?', *A Theory of Legal Personhood* (Oxford University Press 2019) 138.

75 Kurki, 'The Incidents of Legal Personhood' (n 71) 116.

76 *ibid* 117.

77 Kurki, 'Who or What Can Be a Legal Person?' (n 74) 146.

78 *ibid* 150; Ribeiro and others (n 41) 181.

responsiveness, but rather for the purpose of ensuring an effective and fair allocation of risks.⁷⁹ This is a conclusion that is wholly consistent with the considerations exposed in the previous section of this paper, regarding a potential regime under criminal law, and it should be endorsed. The author focuses on the commercial sector and conceives of AI systems as administrators or representatives of legal platforms whose objectives are essentially defined by humans.⁸⁰

This proposal of a limited legal personhood under civil law is both feasible and sensible, and it is endorsed as appropriate to capture the type of legal personhood that could soundly incorporate AI applications to many sectors of activity, promoting the goal of a fair allocation of risks and addressing some of the challenges involved.⁸¹ Indeed, a regime of full legal personhood for AI systems would not be appropriate, as the recognition of active personhood would mean that they would be allowed to own and freely manage their resources.⁸² However, it may be observed that it is not detailed enough so as to account for the different dynamics that are developed in various contexts of their deployment, the concrete risks that are associated with different types of artificial agents and, more specifically, the hybrid nature of action that is inherent in collaborations between humans and AI systems.

4.2 The Notion of ‘Machine Behaviour’ and the Related Proposed Liability Regimes

In view of this need, a noteworthy typology of AI systems that is based on features of machine behaviour has been advanced in the legal field by Anna Beckers and Günther Teubner.⁸³ It is meant to identify and delineate distinct risks that are posed by broader categories of modern AI systems, and to pose the question of personhood in specific socio-technical and legal

79 Kurki, ‘The Legal Personhood of Artificial Intelligences’ (n 3) 189.

80 A legal platform is a bundle of legal positions (a set of rights and duties) that are interconnected and are usually administered by legal acts. They are attached to specific legal persons and exist only in a legal sense. For this definition, see Kurki, ‘The Incidents of Legal Personhood’ (n 71) 96; Kurki, ‘Who or What Can Be a Legal Person?’ (n 74) 135.

81 Gunkel (n 2) 156.

82 See in the same direction Beckers and Teubner (n 28) 10ff.

83 See Beckers and Teubner *supra* (n 45).

contexts.⁸⁴ These are the risk of autonomous decision-making (individual machine behaviour), the risk of association with human agents (hybrid human-machine behaviour) and, finally, the risk of systemic interconnectivity with other agents (collective machine behaviour). This distinction, at the same time, delineates the extent to which legal personhood could be recognised for certain AI systems under civil law.

This classification could be relevant for other branches of law, notably for criminal law purposes, for two main reasons. First, because a clear classification generally precludes the tendency of constructing separate regimes for each AI application. It may be that, depending on the context, a harm caused by the failure of a care robot could have been also caused by other types of autonomous AI applications, if the type of machine behaviour exhibited is essentially the same. And second, because it may contribute to shaping the duty of care of manufacturers or professional users of the system, by taking into account the degree of unpredictability that is associated with the system's conduct.⁸⁵

Individualised software agents constitute the first type of AI agent, associated with the ability and risk presented by the development of autonomous decision-making. Examining individual machine behaviour, one may identify certain intrinsic properties of the algorithm, and the related risk is created by a single source code. Human-machine associations or 'digital hybrids' are created by the close interaction between humans and machines. As the two actors merge, the risks associated with their combined action cannot be broken down into individual contributions, potentially stemming from either side. This is what is identified as the association risk, inherent in hybrid machine behaviour. In some respects, digital hybrids' behaviour is functionally similar to the acts of corporate persons and the difficulty in ensuring accountability is particularly acute. Lastly, interconnected systems or 'multi-agent crowds' exhibit a systemwide behaviour that is formed by the interconnection of multiple AI systems in a network. The latter collectively act like swarms, and this is precisely the in-

84 A similar approach on the question of legal personhood for artificial entities is also applied by Reyes, who examines the case of autonomous corporations; see *supra* (n 4) 1466ff.

85 While it may be possible to identify fault in cases of individual machine behaviour, the same task becomes much more complicated when one is confronted with collective action that is exhibited by multiple, interconnected agents. This is an important issue that would require a separate analysis, and cannot be discussed in the context of this contribution.

terconnectivity risk that is posed by collective agents. In this particular case, the identification of liable parties becomes almost impossible. In view of the features of collective machine behaviour, the requirement of foreseeability cannot be upheld, as it would not have been possible for a human actor to intervene in order to prevent the harm caused.

Having delineated this typology, Beckers and Teubner observe that digital autonomy in the context of civil law can be defined by taking into account the system's ability for developing goal-oriented action and, at the same time, the degree of its participation in communicative processes. For this definition, the philosophical contours of personhood — that base this conferral on the entity's reasons-responsiveness and capacity of freedom — are not authoritative. In their analysis, the defining feature for the recognition of personhood is not the ability for consciousness but for communication.⁸⁶ Indeed, it can be observed that their definition of autonomy chimes with the rationales of civil liability, i.e., with the need to ensure a fair distribution of risks, a fair compensation for damages, and the reduction of financial loss.

Accordingly, the authors propose three distinct liability regimes for autonomous AI systems. Firstly, they analyse the potential legal regime that could apply to AI systems that qualify as 'individual actants'. Such systems are often used in the framework of 'digital assistance', that is in order to conclude contracts in representation of a human principal or organisation (smart contracts).⁸⁷ Because the algorithm in this case acts as a representative and performs legally binding action, which on occasion may cause damage, and precisely in view of the fact that the system can autonomously undertake action that exceeds its principal's intentions, a regime of limited legal personhood constitutes, in their opinion, the most suitable option. The rules of vicarious liability would apply in this case, so these AI systems would be regarded as vicarious agents.⁸⁸ The relevant rules can provide for a reasonable allocations of risks: the principal must assume the risk of hacking because they appear as the party that controls the digital agent, unless it is obvious that the digital agent has produced a result that exceeds

86 Beckers and Teubner (n 45) 58.

87 The term 'digital assistance' is defined as the social relation between human users and digital agents.

88 *ibid* 48, 55.

the delegated decision-making power or the contracting party could have reasonably known that the digital agent was hacked.⁸⁹

The second case relates to the potential legal regime that could apply for digital hybrids. The authors argue that a separate legal capacity should be attributed to this new type of collective actor, that can be regarded as a network. As in this context it is no longer possible to distinguish among individual contributions, these actors must be treated as hybrids precisely in view of the emergent collective properties that appear in their combined input.⁹⁰ They can be compared to corporate actors for two reasons: first, because they do not act on their own behalf, but for the association, and second, because there could be cases of conflict of interest involving the members and the association.⁹¹ However, this comparison remains approximative because corporate actors are characterised by more clearly-structured relations. Therefore, digital hybrids cannot be perceived as genuine collective actors. The hybrid is identified as the source of the harmful action and, subsequently, liability is distributed among the members who control or benefit from its activity (i.e., the operators, owners, manufacturers and deliverers).⁹² The victim's claim is facilitated because they only need to prove that the damage was caused by the hybrid.

The third and last regime examined is that of multi-agent crowds. In this case, it is not possible to refer to either individual or collective decision-making and action, or to any kind of purpose, which means that multi-agent crowds cannot be vested with any type of personhood.⁹³ The interconnectivity risk that appears in this constellation relates to the impossibility of either foreseeing or explaining its outcomes. For instance, the Flash Crash of 2010 was probably produced by such collective machine behaviour, and in general no individual liability can be discerned in similar cases of malfunctions in high-frequency trading.⁹⁴ There is a growing consensus on the opinion that the operation of inter-connected systems does not leave room for the clarification of causality issues and the identification of blameworthy individuals.

89 *ibid* 64.

90 *ibid* 90.

91 *ibid* 94–95.

92 *ibid* 105.

93 *ibid* 111–112.

94 Orlando Cosme Jr, 'Regulating High-Frequency Trading: The Case for Individual Criminal Liability' (2019) 109 *Journal of Criminal Law and Criminology* 365.

In view of this state of affairs, the solution lies in the establishment of risk pools that would provide compensation for damages sustained through adequate funds. In the authors' view, this is a measure to be prioritised because, strictly speaking, it is not possible to comprehend or effectively control interconnected machine operations. This resort to sectoral fund-based schemes is already known from other fields of application, in which extended damages have been caused by complex factors. This is the case with environmental disasters and financial crises, both exacerbated by the activities of corporate actors. For instance, the 'Superfund programme', developed in the US during the 1980s as a federal response for the prevention and mitigation of environmental harm, provides an interesting model for consideration.⁹⁵ This regime involves strict, joint and several liability for potentially responsible parties (site owners, facility operators, transporters or anyone who generates hazardous substances), as well as a federal trust fund in case no such parties can be identified. In this context, the observance of typical practices within the relevant field cannot exempt a party from liability.⁹⁶ In order to reasonably limit the extent of the damages covered, the fund-based regime discussed by the authors would have to remain a subsidiary remedy, activated, for instance, in cases of loss that are not covered by private insurance schemes.

In conclusion, it can be noted that even though the nuanced approach presented by Beckers' and Teubner's scheme accounts for the dangers that are inherent in autonomous AI systems and chimes with the specific aims of civil liability, it cannot as such be authoritative for the recognition of criminal personhood and the establishment of a potential criminal liability against them.⁹⁷ The authors themselves somehow allude to a similar

95 Alexander E Farrell, 'Overview of the Superfund Program' in Gregg P Macey and Jonathan Z Cannon (eds), *Reclaiming the Land - Rethinking Superfund Institutions, Methods and Practices* (Springer) 25-26. This programme was launched by the Comprehensive Environmental Response, Compensation, and Liability Act in 1980 and the Superfund Amendments and Reauthorization Act in 1986. Along with other legislative acts, it aims at providing a structured response to the risk of harm that is created by hazardous substances and compounds, a risk posed both for human health and the environment.

96 *ibid* 41.

97 The nuanced nature of this approach is also praised by Gunkel (n 2) 157, even though ultimately the author does not espouse it for referring to the institution of robot slavery. However, it is noteworthy that Beckers and Teubner do not directly refer to the relevant arguments that exist in the literature, and they simply recognise that the legal status of AI systems can be reasonably compared to the provisions regarding slavery

observation, when they note that ‘each social system creates for algorithms its own criteria of personhood.’⁹⁸ That having been noted, however, any proposition that entails the recognition of even a limited legal personhood to an artificial entity should be examined with caution, as experience has shown that the contours of such personhood typically expand in the course of time.⁹⁹

5. Concluding Remarks

This paper intends to illustrate that while both corporations and AI systems constitute powerful and highly opaque agents that may be delegated extensive decision-making powers in a variety of social contexts, their comparison under the lens of substantive criminal law may seem reasonable at first, but it can be neither self-evident nor absolute. Whereas the recognition of criminal personhood for certain autonomous AI systems is now presented as a plausible and efficient legal development, AI systems cannot be regarded as genuine collectivities, and thus cannot be meaningfully included within the class of criminally responsible subjects. While this may be so, the prospect of a limited legal personhood for civil law purposes — introduced with the goal of ensuring a fair allocation of risks and economically efficient mechanisms — may be endorsed without particular doctrinal problems. If AI systems ever come to develop a kind of general agency that would render direct accountability mechanisms meaningful, a different approach may have to be advanced, but in that case the primary difficulty would not be legal but technical, as merely understanding AGI would already pose an enormous challenge.

under Roman law. This statement is not, in itself, sufficient to justify the rejection of their conceptual framework.

98 Beckers and Teubner (n 45) 147.

99 Chesterman (n 3) 141.

