

## Criminal Regulation of AI Systems: A Primer

A. (Alice) Giannini LLM\*

### 1. Introduction: the general issue

The idea of artificially intelligent (AI) systems committing crimes is by no means new.<sup>1</sup> In truth, ‘bad robots’ that rebel against humans and seize control of them, as well as machines that ‘go mad’ and behave erratically, have been the subject of science fiction for decades.<sup>2</sup> One could argue that Asimov’s three laws of robotics are nothing but the most famous attempt at regulating forms of AI misconduct.<sup>3</sup> Nevertheless, unlike the most popular science fiction novels, nowadays AI has truly penetrated our lives, and is here to stay.

The most recent advancements in AI techniques have allowed for the development of systems that are capable of unsupervised, unpredictable, and autonomous actions. Machine learning (ML)<sup>4</sup> techniques allow algorithms to draw lessons from their past behaviour and teach themselves new behavioural patterns. Therefore, using such techniques might enable algorithmic misbehaviour without any human involvement.

Abbott and Sarch, among the most prominent authors engaged on the topic, argue that AI systems raise an *irreducibility* problem for criminal law.<sup>5</sup> With this expression they refer to situations in which it might be extremely difficult, if not impossible, to reduce a crime committed by an AI system to the actions of a single human being.<sup>6</sup> Irreducibility derives from four characteristics possessed by AI systems: autonomy, i.e., the ability to cause a harmful event without the system being directed to do so by a human agent; opacity, i.e., the impossibility – especially when dealing with more advanced systems such as those based on *deep learning*<sup>7</sup> – to obtain an explanation of how the system, starting from the input (A), obtained the harmful output (B); complexity, i.e., the fact that the creation of an AI system is often the result of the contributions of numerous individuals, developed over a long period of time, as well as the fact that the system may have been trained on heterogeneous and open source databases; and finally

\* Alice Giannini is a Joint PhD Candidate in Criminal Law at the University of Florence and at the University of Maastricht. The supervisors of her PhD research are prof. Michele Papa (Full Professor of Criminal Law, University of Florence) and prof. André Klip (Full Professor of Criminal Law, Criminal Procedure and the Transnational Aspects of Criminal Law, Maastricht University).

1 There is no universally agreed upon definition of AI. Yet, the aim of this paper is not to discuss the issue of defining AI. For the purposes of this paper, it will be assumed that AI has a twofold meaning, following the definition developed by the High-Level Expert Group set up by the European Commission. For a deeper understanding of the issue see, *inter alia*, P. Wang, ‘On Defining Artificial Intelligence’, *Journal of Artificial General Intelligence*, Vol. 10 (2) 2019, p. 1-37. For a general overview of AI definitions, see the research conducted on 55 documents by S. Samoili et al., *AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence*, EUR 30117 EN, Luxembourg: Publications Office of the European Union, 2020.

2 L. Floridi, ‘Should we be afraid of AI?’, *aeon.co*. Available at: [aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible](https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible).

3 (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

4 K. Hao, ‘What is machine learning?’, *MIT Technology Review*, Nov. 17, 2018. Available at: [www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-anotherflowchart-:~:text=What is deep learning%3F,amplify—even the smallest patterns.](https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-anotherflowchart-:~:text=What%20is%20deep%20learning%3F,amplify—even%20the%20smallest%20patterns.)

5 R. Abbott & A. Sarch, ‘Punishing Artificial Intelligence: Legal Fiction or Science Fiction’, *UC Davis Law Review*, Vol. 53, 2019, p. 330.

6 *Ibid.* Drawing on the literature in moral and legal philosophy and ethics of technology, F. Santoni de Sio & G. Mecacci, ‘Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them’, *Philosophy & Technology*, Vol. 34, 2021, identify four responsibility gaps, including the ‘culpability gap’.

7 Deep learning is a subdiscipline of machine learning.

unpredictability, i.e., the ability for the system to undertake activities not anticipated by its original programming.<sup>8</sup>

One thing is certain: the questions that criminal legal scholars are doomed to ask themselves – in the aftermath of yet another advance in technology – are the same. If something goes wrong with these complex systems, should criminal law care? If yes, how?

Indeed, machines are ‘inducing some problems that are specific to criminal law (...) we have to determine whether the behaviour of robots falls within the loopholes of the system, necessitating the intervention of lawmakers at both national and international levels’.<sup>9</sup> It follows, that it will become increasingly challenging to determine, with a legally acceptable degree of certainty (above reasonable doubt), whether the harm caused on a protected legal interest as a result of an action deployed by an AI system can be attributed to a human agent involved in the causal chain of events.<sup>10</sup> In such cases a *responsibility gap* arises. Leaving the debate between ‘techno-pessimists’ and ‘techno-optimists’ on whether this moral responsibility gap can be bridged in the background, the research focuses on its legal counterpart, namely the ‘liability gap’.

## 2. Structure of the research

The main research question of the analysis is the following: To what extent is a theoretical framework of criminal law for liability of non-human agents needed and feasible? Similarly to what happened with animals, or with corporations, this question concerns whether a new type of responsibility for acts of non-humans should be constructed and, if yes, how it should look. In other words, the research will investigate whether a *sui generis* criminal normative framework should be created in order to address the above-described liability gap.

To begin with, the research provides an extensive overview of the existing scholarly debate on criminal liability connected to AI systems. The chapter examines from a critical perspective the literature published in three languages: English, German, and Italian. The authors are classified into three categories (sceptics, moderates, and expansionists). The chapter then highlights the most recurrent questions in the debate.

The literature review delivers the foundation for the following analysis, which will be undertaken both in an *actus reus* and a *mens rea* perspective. Moreover, in order to answer the main research question, it will be neces-

sary to touch upon models of corporate liability. This analysis is mandated by the fact that a very conspicuous number of scholars draws upon the analogy with corporations in order to base their arguments on criminal liability for AI systems. If criminal economic sanctions towards corporations ultimately affect the natural persons behind the legal entity (for their organizational guilt), one should reflect on whether the same punishments could apply to a situation where there is no human involvement, hence directly on the intelligent agent. The research will briefly stretch its frontiers to discuss the topic of ‘algorithmic corporate misconduct’<sup>11</sup> and of ‘corporate algorithmic harm’.<sup>12</sup> In other words, the research will tackle criminal liability for corporations employing AI-systems when such algorithms are involved in misbehaviour.

## 3. Matters of *actus reus*

The issue of whether AI conduct fulfils the *actus reus* element of an offense is sometimes discarded easily in relevant literature. When discussing *actus reus*, specifically the issue of whether an AI system can ‘act’, the real game changer is how one conceives the concept of conduct from a criminal law standpoint. In this regard, as it was argued, it is paramount how much one ‘normatively charges’<sup>13</sup> the concept of action: it might be difficult to conceive an ‘algorithmic’ *actus reus* if acting in a criminally relevant way involves ‘goal determination’,<sup>14</sup> as opposed to just a movement of the ‘body’ (or of part of it).<sup>15</sup>

Some scholars swiftly abandon the doctrine which submits that a criminal act needs to be the result of a willed bodily movement in favour of a strict causalistic perspective.<sup>16</sup> Others believe that AI systems are not capable of committing an act wilfully. They defend the idea that the ability for an agent to understand norms is a prerequisite for its ability to act, and that the voluntariness of an act implies that the agent must be able to act otherwise.

Moving on to the causal nexus, we find the term ‘failures of causation’,<sup>17</sup> which is an effective expression coined by Ugo Pagallo to describe the fact that AI agents break down the classic cause and effect analysis linked to matters of legal causation. The problem of ascertaining causality between the AI system and the harmful event aris-

8 Abbott and Sarch 2019, p. 330.

9 U. Pagallo, *The Laws of Robots: Crimes, Contracts and Torts*, Berlin: Springer 2013, p. 45.

10 S. Beck, ‘Die Diffusion strafrechtlicher Verantwortlichkeit durch Digitalisierung und Lernende Systeme’, *Zeitschrift für Internationale Strafrechtsdogmatik*, Vol. 2, 2020, p. 44.

11 M.E. Diamantis, ‘The Extended Corporate Mind: When Corporations Use AI to Break the Law’, *N.C.L. Rev.*, Vol. 97, 2020.

12 *Ibid.*

13 S. Gleß & T. Weigend, ‘Intelligente Agenten und das Strafrecht’, *ZSTW*, Vol. 126 (3) 2014, p. 572.

14 *Ibid.*

15 *Ibid.*

16 Such as Hallevy & Sartor. See G. Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems*, Berlin: Springer, 2015, p. 25 and F. Lagioia & G. Sartor, ‘AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective’, *Philosophy & Technology*, Vol. 33 (3) 2020, p. 5.

17 U. Pagallo 2013, p. 73.

es because this kind of explanation is linked to the possibility for one to fully dominate (etiologically) a certain event, something that with some AI systems may not be possible. In other words, we can demonstrate that a specific system starting from specific inputs produced an output, but we might not be able to explain why, and how.

There are multiple factors that lead to failures of causation. One of them is the ‘many hands problem’.<sup>18</sup> This expression was created in the field of philosophy and moral responsibility and is used to refer to the fact that the development of AI systems is often the result of the combination of actions by numerous individuals, and the outcome of a long, and complex, chain of individual efforts. Hence, the question that raises is how to isolate the one (human or algorithmic?) factor that led to harm.

Other relevant causes of failures of causation are the ‘black box’ problem<sup>19</sup> and the fact that algorithmic systems might interact in a complex and dynamic ecosystem, which leads to the amplification of the risks of unpredictable dangerous outcomes.<sup>20</sup>

## 4. Matters of *mens rea*

The discussion on AI and *mens rea* follows two main directions.

238

The first direction regards the possibility of conceiving ‘guilty’ AI systems. When discussing direct liability of AI systems, one needs first to address the issue of conceiving said systems as *addressee* of criminal sanctions (i.e., as possessing criminal capacity). Admittedly, legal subjectivity is a prerequisite of liability. The discussion on the issue has taken on an interdisciplinary dimension and is currently happening concurrently in the fields of contract, tort, and criminal law.

The lion’s share of the debate on ‘robotic *mens rea*’ is undertaken by legal philosophers rather than criminal legal scholars. On the one hand, we find the ‘front of robotic liberation’,<sup>21</sup> to which Chopra and White belong.<sup>22</sup> These authors argue that sooner or later robots will be susceptible to the command of the criminal norm and will therefore be reprehensible through punishment. On the other hand, there are those who contend that we cannot speak of robotic culpability because AI systems

lack self-awareness (i.e., they are not conscious of being conscious), free will, and moral autonomy.

The second direction of the debate concerns the liability of the human agent from time to time involved, the so-called ‘human-behind-the-machine’. The research does not consider issues related to cases in which the system is used as a *tool* to commit the crime. Rather, it will focus exclusively on whether human agents could be considered criminally liable for crimes committed by AI systems according to a negligence standard.

## 5. Conclusion

One thing is certain: criminal law is outside its comfort zone when it comes to innovation. AI works as a stress test to classical criminal law constructs such as the concept of act, the causal link, and culpability. It is in this perspective that the question of ascribing criminal liability for artificial intelligence misbehaviour – ‘AIs going bad’ – takes on new relevance.

18 I. van de Poel, L. Royakkers & S.D. Zwart, *Moral Responsibility and the Problem of Many Hands*, London: Routledge 2018.

19 The Black Box Problem can be defined as: ‘an inability to fully understand an AI’s decision-making process and the inability to predict the AI’s decisions or outputs’. Y. Bathaee, ‘The Artificial Intelligence Black Box and the Failure of Intent and Causation’, *Harvard Journal of Law & Technology*, Vol. 31 (2) 2018, p. 905.

20 Council of Europe Study DGI (2019)05, *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*, p. 67.

21 U. Pagallo 2013, p. 54.

22 S. Chopra & L. F. White, *A Legal Theory for Autonomous Artificial Agents*, Ann Arbor: Univ. of Michigan Press 2011.