Who should be held Criminally Responsible in an AI Advancing World?

Jaylee Vella

This article by **Jaylee Vella** was previously submitted as an assignment for CVL1024 and is being published with the author's permission. The article aims to examine whether Artificial Intelligence (AI) could be held liable should an AI entity accomplish criminal activities. The author examines the components which give rise to a criminal action. Afterwards, she analyses the mental element of a crime, known as the mens rea, in regards to AI through an examination of 'crimes with AI' and 'crimes by AI' respectively, as well as addressing current regulatory responses which aim to allow a wider range of crimes involving AI. Finally, the author concludes by suggesting the implementation of an hybrid liability model, where AI is granted legal personhood whilst following any human involvement within any crimes committed.

TAGS: Criminal Law, Philosophy of Law, Artificial Intelligence, Meaningful Human Control, Mens Rea, Criminal Liability

Jaylee Vella is a current third-year Bachelor of Laws (Hons) student, as well as the current President of Għaqda Studenti tal-Liġi. Although her interests in the various sectors of the law are vast and continuously evolving, she is developing a particularly keen interest for litigation, especially in the criminal sector.

OLJ
ONLINE LAW JOURNAL

1. Introduction

Nowadays, Artificial Intelligence (AI) has become a driving force for technological advancement. This widespread innovation is providing a range of challenges that often conflict with criminal law, specifically with regards to the attribution of criminal responsibility should an AI entity accomplish a criminal activity. This article seeks to understand the requirements for criminal liability, if these autonomous systems meet them, and if so, whether they could be held liable or not.

Firstly, the basics of AI must be understood. General AI aims to replicate human intelligence whilst specific/narrow AI is more focused on carrying out a particular task, like for example chatbots. The latter is more prevalent, with general AI not expected to emerge until at least 2040.[1] Notably, narrow AI has been found to consistently outperform humans[2] in its task, hence the explosive popularity of its usage as it facilitates both our private everyday lives and professional workplace. In response, the European Commission takes a risk-based approach to differentiate between different autonomous systems. At the top there is unacceptable risk that prohibits eight practices, followed by high risk which includes employment of AI in medical sectors, navigation and weapons, then limited risk and finally minimal risk.[3]

2. Understanding Crime when it comes to AI

One must understand the meaning and components of crime. To give rise to a criminal action, one must possess both a material element – the actus reus, as well as the mental element – the mens rea. When it comes to AI, whilst some authors have pointed out that the material element could be difficult to establish[4], others say it can be easily carried out because the 'factual-external representation' of the crime can be seen both through physical robots and through those without a physical presence, like chatbots.[5]

Apart from this, to be able to understand the mental element of a crime, we must make a distinction between bots who operate on a human operator's command and those who have their own volition, which will be referred to under 'crimes with AI', and 'crimes by AI'[6] respectively.

3. Crimes with AI

In this case, the system is used as 'nothing but a tool in the criminal hands of a human agent'[7], a means to an end acting as an innocent agent, or if possessing enough mens rea, an accomplice to the crime.

---

[1] Alice Giannini, Jonathan Kwik, 'Negligence Failures and Negligence Fixes. A Comparative Analysis of Criminal Regulation of AI and Autonomous Vehicles' (2023), 34 (1), Criminal Law Forum, 47
[2] ibid, 46
[3] European Commission, 'Regulatory Framework for Artificial Intelligence' (European Commission, 2025) <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> accessed 27 April 2025
[4] Giannini (n 1), 48
[5] Francesca Lagioia, Giovanni Sartor, 'AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective' (2020), 33 (3), Philosophy and Technology, 440
[6] Giannini (n 1), 49
[7] Daniele Amoroso, Benedetta Giordano, 'Who Is to Blame for Autonomous Weapons Systems' Misdoings?', Elena Carpanelli and Nicole Lazzerini (eds), Use and Misuse of New Technologies (Springer International Publishing 2019), 217

The offensive behaviour could have been pre-programmed by software developers or requested by the users of the system[8]. Either way, the AI system's actions on their own could not have led to a crime. Therefore, here the responsibility of the offence shifts from the AI to the human, and because the AI does not satisfy both components of a crime, it cannot be held liable.
One approach to this could be treating the AI as animals, seeing as when animals cause harm, the humans are legally responsible to them, and they compensate the harm through civil obligation[9]. However, this then opens the discussion as to whether the human intentionally let the AI commit the crime, or whether the offence was a result of negligence. In the first case, the criminal is the human operator, but in the second case, responsibility is not specified; it could be any individual from the beginning of the production, until the actual commencement of the crime, whether through negligence or even recklessness, which could be described as 'consciously ignoring the risk'.[10]

This chain of production of an AI system is complicated, which leads to another phenomenon termed the 'responsibility gap', where humans avoid responsibility of the crime as their involvement with the system may be regarded as insignificant to the crime. This stems from the fact that there were 'too many hands' involved in the production and manufacturing of the system, so much so that focalizing the blame on someone proves difficult.[11]
The concept of 'Meaningful Human Control' aims to address this problem.[12] It aims to pinpoint some form of legal responsibility with the reassurance that that every system should have human authority over it. This provides that there were individuals who took an informed and conscious decision that led to that outcome of the offence.[13] For instance, in the French Road Act, autonomous cars need to have a failure system that informs the individual in the car that they should take over the driving due to malfunctions/difficulties.[14] If the individual fails to act on this delegated work, he is then held responsible.

The problem with this is that humans have been proven to be disoriented once thrown back into control[15] after relying on such AI tools and having conceded their authority over it for a long time.[16] An individual won't second guess an automated driving system, and will therefore be caught off guard, with little to no time to intervene if de facto human operation is needed. Humans acting as a 'fail safe' seems practical to be able to pin-point responsibility, but it is very difficult for them to react to unforeseen circumstances with reasonable decisions.[17] In fact, in the United Kingdom (UK), many owners of such cars treat the system to be capable of doing much more than they are programmed for, and this often results in cases of human negligence.[18]

[8] Lagioia (n 5), 453
[9] ibid
[10] Giannini (n 1), 50
[11] Giannini (n 1), 58
[12] Guido Acquaviva, 'Crimes without Humanity? Artificial Intelligence, Meaningful Human Control, and International Criminal Law' (2023), 21 (5), Journal of International Criminal Justice, 986
[13] ibid 989
[14] Giannini (n 1), 70
[15] A.J. Lohn, 'Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance' (2020) 6, <http://arxiv.org/abs/2009.00802>, accessed 27 April 2025
[16] ibid
[17] ibid
[18] Giannini (n 1), 73

4. Crimes by AI

This category involves crimes where AI systems themselves possess their own mens rea and so their inner calculations result in a performance without the help of a human being. Similarly to humans, some AI systems can demonstrate situational awareness[19] through perception of their surrounding environment by sensing, interpreting, filtering, and comprehending their current situation[20]. Furthermore, if the system can anticipate a future event on this basis and devise its own plans, it can easily fulfil the criteria of volition. Consider a medical bot who realises that certain treatment is too expensive, and so autonomously decides to not provide the medication to the patient to reduce costs[21]. Here, the system makes an intentional value-based judgment with a desirable goal from the beliefs of the environment around it.[22]

This elevates the AI systems to possess the same intelligence that humans do, however, the only thing that separates the two is the lack of morality. An AI that commits a crime will not be able to feel guilt when punished and will therefore lack the motivation to abide by criminal law, whose aim is to deter actions that go against norms.[23] Considering this, one can mention how psychopaths often also lack certain morality yet still acknowledge norms. This can lead to the conclusion that morals need not be required for someone to commit a crime with enough reason-responsiveness to be considered of possessing mens rea.[24]

Despite proving that AI systems can possess both elements of a crime, no liability can be given to them unless the state recognises them as a legal person.[25] Drawing on Corporate Criminal Liability, AI can be granted both rights and responsibilities, including possession of both assets and liabilities. This would allow for such systems to pay their own compensation independently, creating a framework for deterrence as they realise, they will be affected from their actions.[26] When it comes to loss of life from the decision of an AI system, a reprimanding action that could be taken is partial or complete deactivation of the AI system. Whilst morally the AI system lacks remorse, logically it will realise that it will suffer a loss if it does that action again, and so the system will look to preserve its own existence.

5. Regulatory Response

The first general attempt to have an AI legal framework was done by Singapore.[27] Despite wanting to criminalise AI offences, their main aim was more to "encourage AI innovation" through the introduction of AI regulation. In fact, both the Singapore Penal Code Review Committee and Singapore Academy of Law's reports offered a wider view of offences, not just limited to a specific field.[28] However, both initiatives were found to consist of major lacunas, and so new action was taken through the introduction of the "User in Charge" doctrine. This phenomenon cleared the previously blurred line of the AI system having a mens rea, as it put responsibility on the AI user to determine the actions of the system and intervene when necessary.[29]

---

[19] Lagioia (n 5), 441- 443
[20] ibid
[21] Lagioia (n 5), 457
[22] ibid 446
[23] Lagioia (n 5), 449
[24] ibid 450
[25] ibid 455 - 456
[26] ibid 459
[27] Giannini (n 1), 60
[28] ibid 61
[29] ibid 66

The UK Law Commission Report similarly adopted the "User in Charge" principle. The only difference is that the UK version only refers to those in charge of automated vehicles, whilst the Singaporean one has a wider application.[30] The UK created a threshold of levels that would determine whether the self-driving car has enough autonomy to be held responsible in an accident, or whether the user should have taken over the car to prevent said accident.[31] It is argued that for this to work, the car user would have to have a special drivers' license that proves he is able to respond appropriately if needed.[32]

On a broader scale, there is the more recent AI Act, which is said to have phased implementation and is part of a wider pack of policy measurements. It assures that before AI systems are put on the market, they are registered in the European Union database with authorities overseeing market surveillance and deployers and providers ensuring post-market monitoring. This allows for less crimes to be committed by AI themselves due to human oversight.

Furthermore, the European Committee on Crimes Problems has been tasked with drafting laws on criminal liability regarding the use of AI. This is expected to be done by end of 2025, and it aims to encourage member states to adapt their current criminal law and allow for a wider range of offences that involve AI.

6. Conclusion

This paper aims to answer the difficulties that exist when it comes to attributing criminal liability to AI systems. As the systems continue to evolve, these questions grow more pressing, and the situation is more complex than it seems at first glance. These complexities lead to struggles from various countries to accurately adapt their legal systems to AI. Central to this however is the constant role of humans, either their help in the offence, their negligence that led to the offence or their qualities that were learnt by AI systems and used to commit the crimes.

Despite this, the autonomy of AI systems pushes us away from assigning responsibility to humans. AI being granted legal personhood may offer as a future solution, but it still comes with ethical challenges. In fact, a hybrid liability model can be introduced to possibly give AI legal personhood but also involve human oversight and regulatory structures to be followed. This is to ensure that future frameworks still protect human rights, societal values and the foundational principles of criminal law.

[30] ibid
[31] ibid 72
[32] ibid 79



OLJ

GrtSL ONLINE LAW JOURNAL