Machina Sapiens Criminalis: Can AI entities be held criminally responsible?

By Farrid Assaf SC

Introduction

In 2017, Saudi Arabia granted citizenship to a humanoid robot named Sophia and thus it became the first robot in the world to receive citizenship.[1] Sophia, who can communicate in real-time using a combination of voice and human-like facial expressions and exhibits characteristics of neocortical functioning[2] remains 'alive' to this day and has found use as a 'social robot' taking care of the sick and elderly.[3] Sophia's makers plan to mass-produce similar type robots by the end of the year. Further, during the last year significant steps have been taken in the advancement of autonomous driving technology including the introduction of 'robotaxi' services to the public around the world.[4] It is little wonder therefore that recent years have seen a considerable increase in research dedicated to examining whether criminal liability can or should be attributed to non-biological entities. Indeed, one commentator has suggested that a new subject be added to criminal law, in addition to humans and corporations, to which he refers as *machina sapiens criminalis*.[5] The attribution of criminal responsibility to such entities is a vexed and intensely perplexing issue. Any attempt at resolution or examination entails consideration of profound philosophical and jurisprudential conundra. The purpose of this article is to explore at a high level some of the questions that may fall for consideration in any debate as to whether non-human entities, and in particular artificial intelligence (AI) entities, may be held criminally responsible.

Is the application of criminal law to AI entities necessary?

It is trite to observe that the rationale for criminal law is essentially its use as a mechanism for the preservation of social order. To criminalise a certain kind of conduct is to declare that it is a public wrong that should not be done, to institute a threat of punishment in order to supply a pragmatic reason for not doing it, and to censure transgressors of such declared wrongs.[6] In addition to deterrence however, the aims of criminal law also include its power to express retribution and moral condemnation.[7] In the context of criminal corporate responsibility, the Australian Law Reform Commission recently observed that in order for corporate criminal responsibility to have a distinct purpose it must be possible to apply concepts of retribution and denunciation to a corporation.[8] Similar observations have been made in respect of possible AI criminal responsibility.[9] Accordingly, imposition of criminal responsibility upon either a corporation or an AI entity would require the relevant transgressor capable of moral blameworthiness.[10] Attributing moral blameworthiness to a non-human entity however is exceedingly complex. In the context of corporations at least, the balance of opinion favours the view that a corporation *itself* is to be regarded as a 'blameworthy moral agent'[11] which can act from 'moral positions' and hence act 'wrongly.'[12] Hence, when in 1909, the United States Supreme Court held for the first time that a corporation should be held criminally responsible it noted:

> If, for example, the invisible, intangible essence of air, which we term a corporation, can level mountains, fill up valleys, lay down iron tracks, and run railroad cars on them, it can intend to do it, and can act therein as well viciously as virtuously.[13]

Likewise, in Australia, Federal Parliament has sought to criminalise corporate conduct in a number of disparate areas. In 2019 for example, the ALRC reviewed 25 Commonwealth statutes and identified 2,898 criminal offences as potentially applicable to corporations.[14] The ALRC observed in that regard that Australia's corporate criminal responsibility regime forms a small part of the broader system of corporate regulation which seeks to promote compliance and ensure that corporate entities adhere to the norms of conduct prescribed by Parliament.[15] One can immediately see the parallels between criminalisation of the activities of fictitious legal entities with non-biological entities capable of autonomous activity.

Three models of AI entity criminal responsibility

In one of the more influential works in this area[16], Professor Gabriel Hallevy of Ono Academic College in Israel posits three possible models for 'virtual' criminal responsibility all aimed towards identifying when the conduct of an AI may satisfy both the actus reus and mens rea requirements of a crime. First, the perpetration-by-another responsibility model; second the natural-probable-consequence model and third the direct responsibility model.

Perpetration-by-Another Virtual Responsibility Model

The first model does not consider the AI entity as possessing any human attributes and the entity is considered a mere innocent agent. [17] In Australia, the doctrine of 'innocent agency' is a means by which the common law attaches criminal liability to a person who does not physically undertake some or all of the elements of the offence with which he or she is charged.[18] For example, the accused may have induced a young child to do the act which constitutes the actus reus of the crime, or imported drugs via an airline carrier. In that case, the agent is innocent of any wrong doing and the accused is regarded as a principal in the first degree.[19] In the context of an AI entity, the question arises who is the perpetrator-via-another?[20] There are two possible candidates: the first is the programmer of the AI software and the second is the user, or the end-user.[21] Where it can be readily established that the perpetrator has the relevant mens rea then attribution of criminal responsibility in such cases is likely to be straightforward. For example, if a human intentionally or knowingly programs a robot so that it causes harm to a person, the programmer's criminal responsibility can easily be established on the basis of traditional concepts of attribution and mens rea: the programmer commits the criminal act by using the robot – irrespective of its artificial intelligence – as a tool for carrying out the programmer's intention, and she does so with the requisite intent or knowledge.[22] Difficulties arise however according to the level of sophistication and autonomy that the AI agent possesses.[23] This model is therefore likely not suitable when an AI entity commits an offence based on its own accumulated experience or knowledge or when the software of the AI entity was not designed to commit the specific offence but was committed by the AI entity nonetheless.

Natural-Probable-Consequence Virtual Responsibility Model

According to the second model, a person might be held accountable for an offence if that offence is a natural and probable consequence of that person's conduct. The second model of criminal responsibility assumes deep involvement of the programmers or users in the AI entity's daily activities, but without any intention of committing any offence via the AI entity.[24] In the United States for example, the natural-probable-consequence doctrine has been used to impose criminal responsibility upon accomplices in circumstances where one committed an offence which had not been planned by all defendants and which was not part of a conspiracy. The doctrine attributes responsibility to the accomplice to acts of a perpetrator that were a 'natural and probable consequence' of a criminal scheme that the accomplice encouraged or aided.[25] This is similar to the concept of criminal negligence in Australian law.[26] So, for example, under this model the programmers or users of an AI entity may be criminally liable in say, manslaughter by criminal negligence, if all of the requisite elements of negligence are satisfied and the court were to find that the act or omission fell so far short of the standard of care which a reasonable person would have exercised in the circumstances; it involved such a high risk that death or really serious bodily harm would follow and that the degree of negligence involved in the conduct is so serious that it should be treated as criminal conduct.[27]

The question however remains as to the extent, if any, of criminal responsibility of the AI entity itself. Hallevy suggests two possible outcomes. First, if the AI entity acted as an innocent agent, without knowing anything about the criminal prohibition then no criminal responsibility would be attributed to the AI. If however the AI entity did not act merely as innocent agent then the AI entity itself should be held criminally liable in addition to the liability of the programmer or user.[28]

The Direct Virtual Responsibility Model

In this model, criminal responsibility is ascribed to the AI entity *itself* if both actus reus and mens rea elements could be proved against the entity. Accordingly, this model does not assume any dependence of the AI entity on a specific programmer or user and instead focuses on the AI entity itself.[29] Hallevy argues that most AI algorithms are capable of analysing permitted and forbidden[30] and that provided the requisite elements of actus reus and mens rea are established, the criminal responsibility of an AI entity according to the direct responsibility model is not different from the relevant criminal responsibility of a human.[31]

Possible criminal responsibility of autonomous driverless vehicles

In a separate paper[32] Hallevy examines potential criminal liability in the context of autonomous driverless vehicles. Leaving aside the first two models above which are contingent upon human intervention, Hallevy argues that it is theoretically possible to be able to establish the mens rea element of an offence committed by an autonomous driverless vehicle and that there is no reason for it to be exempt from criminal liability.[33]

Rocco Fazzari

While that may be theoretically possible, a significant practical problem would be trying to prove the mens rea element of the AI entity to the requisite criminal standard. One significant practical obstacle in that regard is the lack of transparency to AI decision making systems and processes. For example, machine learning algorithms commonly used today are capable of learning from massive amounts of data, and once that data is internalised, they are capable of making decisions experientially or intuitively like humans.[34] It can often be difficult, if not impossible, to determine how an AI that has internalised massive amounts of data is making its decisions.[35] The implications of this opacity for determining any mens rea element are obviously significant and far-reaching.[36] Commentators have suggested two possible (but ultimately poor) solutions to the lack of transparency. The first is to regulate the degree of transparency that AI must exhibit and the second to impose strict liability for harm inflicted by AI. Both solutions are problematic, incomplete, and likely to be ineffective levers for the regulation of AI[37] and would in particular likely lead to a stifling of innovation in AI research and development.

Can AI entities be punished?

Possibly one of the most challenging aspects of the imposition of criminal liability on an AI entity is punishment. While the literature acknowledges and analyses the problem no satisfactory solution has been proffered. Monetary penalties are unlikely to be of any utility since an AI entity is unlikely to own any property in the traditional sense. Physical destruction or impairment of the AI entity may be akin to corporal punishment or even the death penalty but is unlikely to have a comparable effect on the robot, at least as long as it is not imbued with a will to live.[38] In short, AI entities are likely incapable of understanding the meaning of punishment and therefore, cannot draw a connection between anything 'done to them' and their prior fault.[39]

Conclusion

This brief article has sought to raise awareness of some of the possible challenges that may arise in the not too distant future regarding criminal responsibility and AI entities. One approach of course is to simply do nothing and instead continue to treat an AI malfunction as simply bad luck[40] as has been done in the past.[41] Such an approach however is likely to become politically unpalatable as AI technology becomes increasingly more sophisticated and widespread. There are, already numerous examples of 'corporate algorithmic harm' that warrant further research and inquiry. For example, a lender's automated platform approving mortgages in a fashion that has a discriminatory racial impact but might also have a business justification; competing retailers' pricing algorithms setting prices at matching, super-competitive levels and a delivery company's self-driving truck striking a jaywalking pedestrian.[42] The author considers that a coordinated and considered approach needs to be taken towards the issues identified as has been the case with the Federal Government's recent examination of corporate criminal responsibility mentioned above. [43]

BN

ENDNOTES

[1] Nora Osmani, 'The Complexity of Criminal Liability of AI Systems' (2020) 14(1) *Masaryk University Journal of Law and Technology* 53, 59-60, referring to Wootson, C. (2017) *Saudi Arabia, Which Denies Women Equal Rights, Makes a Robot a Citizen.* [online] Available from: https://www.ndtv.com/world-news/saudi-arabia-which-denieswomen-equal-rights-makes-a-robot-a-citizen-1768666 [Accessed 20 January 2019].

[2] Priya Persaud, 'Protecting against Ultron: Exploring the Potential Criminal Liability of Self-Programming Deep Learning Machines' (2020) 72(2) *Rutgers University Law Review* 577, 584.

[3] Michelle Hennessy, 'Makers of Sophia the robot plan mass rollout amid pandemic', Reuters, 25 January 2021 Available from: https://www.reuters.com/article/us-hongkong-robot-idUSKBN29U03X].

[4] William Smith 'Robotaxis: the future of transport', *Technology Magazine* [online] Available from: https://www.technologymagazine.com/ai/robotaxis-future-transport

[5] Nora Osmani, 'The Complexity of Criminal Liability of AI Systems' (2020) 14(1) *Masaryk University Journal of Law and Technology* 53, 58, referring to Hallevy, G. (2015) *Liability for Crimes Involving Artificial Intelligence Systems.* Springer International Publishing, p. 21.

[6] A Ashworth and J Horder, *Principles of Criminal Law* (7th ed, 2013) 22, as quoted in David Brown et al, *Criminal Laws: Materials and Commentary on Criminal Law and Process of New South Wales* (The Federation Press, 6th ed, 2015) 63.

[7] Lawrence Friedman, 'In Defence of Corporate Criminal Liability' (2000) 23 *Harvard Journal of Law and Public Policy* 833 at 834.

[8] *Corporate Criminal Responsibility – Discussion Paper 87*, ALRC, November 2019 at [2.31].

[9] See for example Dafni Lima, 'Could AI Agents Be Held Criminally Liable: Artificial Intelligence and the Challenges for Criminal Law' (2018) 69(3) *South Carolina Law Review* 677, 688-9.

[10] Dafni Lima, 'Could AI Agents Be Held Criminally Liable: Artificial Intelligence and the Challenges for Criminal Law' (2018) 69(3) *South Carolina Law Review* 677, 686-7.

[11] Brent Fisse and John Braithwaite, *Corporations, Crime and Accountability* (Cambridge University Press, 1993) 35-6.

[12] Sylvia Rich, 'Corporate Criminals and Punishment Theory' (2016) 29 *Canadian Journal of Law & Jurisprudence* 97, 109.

[13] *New York Central and Hudson River Railroad Company v United States* 212 US 481 (1909) 493.

[14] *Corporate Criminal Responsibility – Discussion Paper 87*, ALRC, November 2019 at [3.10].

[15] *Corporate Criminal Responsibility – Discussion Paper 87*, ALRC, November 2019 at [1.13].

[16] Gabriel Hallevy, 'Virtual Criminal Responsibility' (2010) 6(1) *Original Law Review* 6.

[17] Gabriel Hallevy, 'Virtual Criminal Responsibility' (2010) 6(1) *Original Law Review* 6, 11.

[18] *Pinkstone v R* [2004] HCA 23; 219 CLR 444 per McHugh and Gummow JJ at [59].

[19] *Pinkstone v R* [2004] HCA 23; 219 CLR 444 per McHugh and Gummow JJ at [59].

[20] Gabriel Hallevy, 'Virtual Criminal Responsibility' (2010) 6(1) *Original Law Review* 6, 12-13.

[21] Gabriel Hallevy, 'Virtual Criminal Responsibility' (2010) 6(1) *Original Law Review* 6, 11.

[22] Sabine Gless, Emily Silverman and Thomas Weigend, 'If Robots Cause Harm, Who Is to Blame: Self-Driving Cars and Criminal Liability' (2016) 19(3) *New Criminal Law Review* 412, 425.

[23] Dafni Lima, 'Could AI Agents Be Held Criminally Liable: Artificial Intelligence and the Challenges for Criminal Law' (2018) 69(3) *South Carolina Law Review* 677, 690.

[24] Gabriel Hallevy, 'Virtual Criminal Responsibility' (2010) 6(1) *Original Law Review* 6, 13.

[25] See for example LaFave, *Principles of Criminal Law*, 3rd ed, 2000, p 636 and *People v Prettyman* 926 P 2d 1013 (Cal 1996) at 1015, 1019 endorsing liability of the confederate for crimes committed as a 'natural and probable consequence' of the crime originally aided and abetted.

[26] See for example a discussion of the elements of manslaughter by criminal negligence in *Justins v Rina* [2010] NSWCCA 242 at [223].

[27] *Justins v Rina* [2010] NSWCCA 242 at [223].

[28] Gabriel Hallevy, 'Virtual Criminal Responsibility' (2010) 6(1) *Original Law Review* 6, 16.

[29] Gabriel Hallevy, 'Virtual Criminal Responsibility' (2010) 6(1) *Original Law Review* 6, 17.

[30] Gabriel Hallevy, 'Virtual Criminal Responsibility' (2010) 6(1) *Original Law Review* 6, 19.

[31] Gabriel Hallevy, 'Virtual Criminal Responsibility' (2010) 6(1) *Original Law Review* 6, 21.

[32] Gabriel Hallevy, 'Unmanned Vehicles: Subordination to Criminal Law under the Modern Concept of Criminal Liability' (2011) 21(2) *Journal of Law, Information and Science* 200, 208.

[33] Gabriel Hallevy, 'Unmanned Vehicles: Subordination to Criminal Law under the Modern Concept of Criminal Liability' (2011) 21(2) *Journal of Law, Information and Science* 200, 210.

[34] Yavar Bathaee, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation' (2018) 31(2) *Harvard Journal of Law & Technology* 889, 891.

[35] Yavar Bathaee, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation' (2018) 31(2) *Harvard Journal of Law & Technology* 889, 891.

[36] Yavar Bathaee, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation' (2018) 31(2) *Harvard Journal of Law & Technology* 889, 892.

[37] Yavar Bathaee, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation' (2018) 31(2) *Harvard Journal of Law & Technology* 889, 893.

[38] Sabine Gless, Emily Silverman and Thomas Weigend, 'If Robots Cause Harm, Who Is to Blame: Self-Driving Cars and Criminal Liability' (2016) 19(3) *New Criminal Law Review* 412, 423-4.

[39] Sabine Gless, Emily Silverman and Thomas Weigend, 'If Robots Cause Harm, Who Is to Blame: Self-Driving Cars and Criminal Liability' (2016) 19(3) *New Criminal Law Review* 412, 423-4.

[40] Dafni Lima, 'Could AI Agents Be Held Criminally Liable: Artificial Intelligence and the Challenges for Criminal Law' (2018) 69(3) *South Carolina Law Review* 677, 694.

[41] See for example *Payne v ABB Flexible Automation*, No. 96-2248, 1997 U.S. App. LEXIS 13571, at *2-3 (8th Cir. June 9, 1997), where a factory worker, Michael Payne, was killed by the actions of an AI robot. The court ruled in favour of the manufacturer of the AI machine, finding that the design of the AI was not a product defect and that Payne should not have approached the AI without proper safety precautions however the court failed to discuss potential criminal liability of the AI machine.

[42] Mihailis E. Diamantis, 'The Extended Corporate Mind: When Corporations Use AI To Break the Law' (2020) 98(4) *North Carolina Law Review* 893, 896-7.

[43] *Corporate Criminal Responsibility – Discussion Paper 87*, ALRC, November 2019.