Panjab University Law Magazine (MAGLAW)
ISBN: 2582-3507, Volume Iv Issue I, pp 68-86

BETWEEN CODE AND CULPABILITY: DECIPHERING THE POSSIBILITY OF AI *MENS REA* FOR CRIMINAL LIABILITY THROUGH JURISTIC PERSONHOOD FOR AI

Raajdwip Vardhan*

ABSTRACT

The rapid evolution of Artificial Intelligence (hereafter AI) challenges traditional legal frameworks, including existing criminal law frameworks of attributing liability. While actus reus in AI-driven offences can be identified, determining mens rea remains complex due to AI's autonomous decision-making and the Black Box Problem. Against this backdrop, this paper examines whether AI can be granted juristic personhood, thereby attributing mens rea directly to it, and explores the feasibility of a strict liability regime to bypass the mens rea requirement in AI-driven offences. Drawing from existing jurisprudence, the study argues that recognising AI as a juristic person is not legally untenable, and thus, it is possible to attribute mens rea to AI if the constitutive elements of its different forms are present, a possibility that cannot be ignored vis-à-vis the hypothetical "Strong AI", thus also allowing criminal liability to be directly applied to AI entities themselves.

Keywords: *Strong AI, mens rea, actus reus, criminal liability, black box problem*

1. INTRODUCTION

Since the dawn of evolution, humanity has continuously pushed the boundaries of discovery and innovation. One of the transformative outcomes of this pursuit was the development of the computing machine. Initially designed for complex calculations, it has now become an indispensable technological asset, evolving far beyond its original purpose. Among the groundbreaking advancements within computation is the creation of Artificial Intelligence (hereafter AI) which has transcended the limits of what was considered possible through computation.

As AI becomes increasingly integrated into various sectors, legal challenges surrounding its role in criminal activities have emerged. Traditionally, accidents involving robots following pre-programmed instructions have been addressed

* Ph.d. Research Scholar, Department of law, North-Eastern Hill, University Shillong

under product liability, holding manufacturers accountable for malfunctions and negligence. However, AI-powered systems capable of autonomous decision-making present a new dilemma: How should the law, respond to a truly intelligent and independent entity?

In the domain of criminal law, establishing guilt requires both actus reus and mens rea, with the former alluding towards the guilty act and the latter alluding towards the guilty mind. While actus reus may be easily attributed to an AI system, determining mens rea poses a complex challenge since an autonomous and independent AI that functions intelligently makes ascertaining the source of the mens rea vis-à-vis the offence committed a difficult proposition.

Against this backdrop, this paper explores the fundamental legal questions involving the determination of mens rea in offences involving AI through a twofold approach. First, it will delve into an analysis of whether AI entities can be granted juristic personhood, thus bringing them within the scope of law. Second, it analyses how mens rea in AI-driven crimes can be interpreted and addressed within the existing criminal law framework. The study aims to advance the evolving discourse on AI accountability and criminal liability by analysing how the concept of mens rea can be interpreted.

2. ARTIFICIAL INTELLIGENCE: A CONCEPTUAL OVERVIEW

The history of AI is traceable practically to developments in computing that happened during and after the 2nd World War, yet, from a conceptual perspective, the origin of the concept of autonomous beings that are artificially created can be traced to the writings of Homer in the 7th century B.C who depicted such automata in both "The Iliad" and "The Odyssey", his two great epics.[1] The actual beginnings, however, would happen during the middle of the 20th century, when Alan Turing published a paper titled "On Computable Numbers With an Application to the Entscheidungsproblem" which conceptualized a hypothetical machine known as the "Turing Machine" that could solve any computational problem through binary codes reducible to 1s and 0s.[2] Turing's initial conceptualization, and progress in computing technology over the next decade, culminated in the early stage of AI development at the Dartmouth Conference of

1 Stephen Cave, Kanta Dihal et.al., (eds.) AI Narratives: A History of Imaginative Thinking About Intelligent Machines 41 (Oxford University Press: Oxford, 2020).

2 Pamela McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* 63 (A.K. Peters Ltd: Massachusetts, 2004).

1956, where John McCarthy coined the term "Artificial Intelligence" and defined it as "the science and engineering of making intelligent machines."[3]

From a definitional standpoint, AI is an elusive concept. This is because AI has permeated and become integrated into various aspects of human society, carrying out a diverse range of tasks. At its most rudimentary level, an AI is representable by algorithms, albeit extremely complex ones, yet, this approach does not take into account the nuances of AI. The comprehensive definitional attempt that has been taken up by Russel & Norvig, who have categorized AI into four different domains based on the kind of task that it is expected to perform – 'Thinking Humanely' which is encapsulated in Richard Bellman's definition as - "The automation of activities that we associate with human thinking, activities such as decision-making, problem-solving, and learning" [4]; 'Acting Humanely' that is captured within George F. Luger's definition of AI as "the branch of computer science that is concerned with the automation of intelligent behaviour"[5]; 'Thinking Rationally' that is best described by Patrick Winston's definition of AI, which states that "AI is the study of the computations that make it possible to perceive, reason, and act"[6]; and, 'Acting Rationally', which is best represented by Poole & Mackworth's definition that "AI is the study of the design of intelligent computational agents"[7], capture the diverse capabilities of modern AI to the greatest possible extent.

Since 1956 when the term AI was conceptualized, it has exhibited exponential growth, made possible by the creation of novel forms of computation that have enabled AI to be applied to various domains of human society. Key technological developments such as "Machine Learning" (hereafter ML), "Deep Learning" (hereafter DL) and "Artificial Neural Networks" (hereafter ANN) have played a crucial role in this proliferation, and a conceptual overview of AI without delving briefly into these developments would be incomplete as they form the foundation upon which modern AI systems operate and evolve. ML is a subset of AI that enables systems to learn from data without explicit programming, and it can be defined as the "ability to learn and enhance from

---

3 Christopher Collins et.al., "Artificial Intelligence in Information Systems Research: A Systematic Literature Review and Research Agenda" 60 *International Journal of Information Management* 02 (2021).

4 Richard Bellman, *An Introduction to Artificial Intelligence: Can Computers Think?* 28 (Boyd & Fraser: San Francisco, 1978).

5 George F. Luger, *Artificial Intelligence: Structures and Strategies for Complex Problem Solving* 01 (Pearson Addison Wesley: Massachusetts, 2008).

6 Patrick H. Winston, *Artificial Intelligence* 05 (Addison-Wesley: Massachusetts, 1992).

7 David L. Poole & Alan K. Mackworth, *Artificial Intelligence: Foundations of Computational Agents* xiii (Cambridge University Press: Cambridge, 2010).

experience automatically without being specifically programmed."[8] ML models identify patterns and improve performance over time, with this adaptability enabling AI to perform complex tasks that are complex in nature, and which, due to their inherent dynamism, involving a lot of elements, cannot be completely programmed through code. ANNs are structures whose creation was inspired by the physical structure of the brain, consisting of layers of interconnected nodes (which function like neurons of the brain) that are responsible for processing and transmitting information and can adjust their 'weights' (numerical values representing the direction and strength of neuron influence) through training,[9] thus mimicking human learning and cognition to some degree. DL is a more advanced form of ML that leverages multiple layers of ANNs to analyse vast amounts of data, and can be defined as "a type of machine learning that uses deep (or many-layered) artificial neural networks - software that roughly emulates the way neurons operate in the brain."[10] Unlike traditional ML, which requires feature extraction by humans, DL models autonomously discover patterns and correlations within data, and therefore, can be deemed to be a method of multi-level data processing akin to the data processing that takes place within the human brain.[11]

However, the use of ML, ANN and DL has given rise to a new phenomenon known as the Black Box Problem. This 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."[12] Another definition of a Black Box can be an AI "which uses data not accessible to the data subject, and/or which deploys algorithms which are either similarly inaccessible or so complex that they cannot be reduced to a series of rules and rule applications comprehensible to the data subject."[13] Thus, this problem refers to AI systems whose decision-making processes are opaque, making it difficult to understand or predict their outputs, due to the use of inaccessible data or highly complex algorithms that cannot be

---

8 Iqbal H. Sarker, "Machine Learning" Algorithms, Real-World Applications and Research Directions" 2 *SN Computer Science* (2021).

9 Enzo Grossi & Massimo Buscema, "Introduction to Artificial Neural Networks" 19 *European Journal of Gastroenterology & Hepatology* 1046-1048 (2008).

10 Gregory Scopino, *Algo Bots and the Law: Technology, Automation, and the Regulation of Futures and Other Derivatives* 35 (Cambridge University Press: Cambridge, 2020)

11 Iqbal H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions" 2 *SN Computer Science* (2021).

12 Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation" 31 *Harvard Journal of Law & Technology* 905 (2018).

13 Frank Pasquale, "Normative Dimensions of Consensual Application of Black Box Artificial Intelligence in Administrative Adjudication of Benefits Claims" 84 *Law and Contemporary Problems* 36 (2021).

easily explained or understood by those affected by its decisions. It arises due to the complexity and opacity of ML and DL mechanisms where vast amounts of data are processed, patterns are identified, and decisions are made by AI without providing a clear explanation of their reasoning. Unlike traditional rule-based algorithms, where the logic and reasoning used are transparent, understandable and comprehensible for humans, ML and DL models function through intricate transformations, thus making it difficult for humans to understand and trace how the AI has arrived at a particular decision.

3. THE LEGAL PERSONHOOD DEBATE: CAN AI BE A 'PERSON' UNDER LAW?

The word 'person' originates from the Latin term "persona", which initially referred to the masks worn by actors in theatrical performances. Over time, the term evolved to denote not just the characters being portrayed but also the actors themselves, thus becoming a term that refers to humans in general.[14] In legal discourse, however, the concept of 'person' is more complex, encompassing both 'natural persons' and 'juristic persons', with the former being human beings who inherently possess rights and obligations, and the latter being an entity that is granted a fictional legal recognition of rights and duties, thereby making it a 'person' before law, without it being a human being. In this vein, Salmond defines 'person' as "any being to whom the law regards as capable of rights or duties. Any being that is so capable is a person, whether human being or not, and nothing that is not so capable is a person, even though he be a man."[15] This definition highlights that legal personhood is not contingent upon human characteristics but rather on an entity's capacity to bear rights and obligations.

Legal systems worldwide have long recognized non-human entities as juristic persons. For example, under corporate law, a corporation is treated as a distinct legal entity, separate from its shareholders, possessing rights and liabilities independent of its members.[16] Centuries before this development, the law furthered the notion of 'person' to institutions such as the Church and universities, recognizing that they could "hold property, sue or be sued, and enter into contracts in its own name, apart from any of the members who were members of or affiliated with the institution" with the additional characteristic being that the property or rights vested in such institutions will continue being

---

14 Eliva Arcelia Quintana Adriano, "The Natural Person, Legal Entity or Juridical Person and Juridical Personality" 4 *Penn State Journal of Law and International Affairs* 366 (2015).

15 John W. Salmond, *Jurisprudence or the Theory of the Law* 334 (Stevens & Haynes: London, 1902).

16 *Salomon v. Salomon & Co. Ltd*., (1897) AC 22.

vested on the juristic entity even after the "death or departure of any of the natural persons associated with the entity."[17] In addition, natural entities such as rivers (the Whanganui River in New Zealand)[18] and animals (Cecilia the Chimpanzee was freed under a habeas corpus petition by a court in Argentina)[19]

Given this established legal framework, there are no fundamental barriers to granting juristic personhood to an AI agent. Scholars argue that "when a legal system confers legal rights and obligations on an entity, it has determined to treat that entity as though it were a person in fact. It is kind of a pretence in which legal systems can decide to engage, regardless of whether an entity really is a person."[20] This understanding suggests that juristic personhood is a legal fiction, not necessarily tied to human attributes, with legal systems being capable of assigning rights and responsibilities to any entity, notwithstanding the presence or absence of biological existence. This perspective reinforces the idea that legal personhood is a construct shaped by societal and legal needs rather than intrinsic human attributes.

The evolution of AI necessitates a serious reconsideration of its legal status. AI systems range from "Weak AI", which performs specific tasks within predefined limits, to "Strong AI", which, in theory, could exhibit human-like cognitive abilities. A notable example of weak AI is IBM's Deep Blue, a chess program that famously defeated the then-world champion Gary Kasparov in 1997,[21] thus heralding the era of chess AI that is significantly stronger than humans in the game of chess. However, despite the advanced abilities that Deep Blue exhibits, they are strictly confined to the domain of chess, an environment which does not necessitate the creation of any rights and obligations from a legal perspective, and therefore, consequently, there is no pressing need to grant it legal personhood.[22]

However, AI is rapidly advancing beyond narrowly defined tasks. The emergence of ML and DL, along with the addition of ANN have enabled AI to

---

17 Margaret M. Blair, "Corporate Personhood and the Corporate Persona" *University of Illinois Law Review* 788 (2013).

18 Miriama Cribb, "Beyond Legal Personhood for the Whanganui River: Collaboration and Pluralism in Implementing the *Tw Awa Tupua Act*" *The International Journal of Human Rights* 02-03 (2024).

19 Steven A. Wise, "A New York Appellate Court Takes a First Swing at Chimpanzee Personhood: And Misses" 95 Denver Law Review 276-277 (2017).

20 Joanna Bryson et.al., "Of, For, and By the People: The Legal Lacunae of Synthetic Persons" 25 *Artificial Intelligence and Law* 276 (2017).

21 Murray Campbell et.al., "Deep Blue" 134 *Artificial Intelligence* 57 (2002).

22 Shubham Singh, "Attribution of Legal Personhood to Artificially Intelligent Beings" *Bharati Law Review* 199 (2017).

learn on its own, refine its knowledge, and automate its functioning in ways that may have previously been unfathomable to even the developers of the AI during its creation. For example, autonomous vehicles make independent decisions in real-world environments, sometimes resulting in accidents or legal disputes. If an AI-driven car causes harm, determining liability becomes complex. While assigning juristic personhood to an autonomous self-driving car is necessarily not the argument here, the point being attempted to be conveyed is that there is a need to reconsider the possibility of granting juristic personhood to AI entities at some point in the future when the autonomy exhibited by AI goes beyond the formalistic limitations of "Weak AI".

The debate extends further when considering the prospect of "Strong AI". Although this merely remains a theoretical concept at present, future advancements could produce AI that not only processes information but also exhibits traits such as intention, conscience and self-awareness. Considering a possibility where AI could replicate all aspects of human cognition, including the emotional and psychological attributes, the only distinguishing factor between such an AI and a human being is its origin, with the former being a product of programming and the latter being 'naturally' born, it can be a natural corollary to grant such AI personhood. In other words, if "Strong AI" were to possess intelligence comparable to that of a human across all measurable parameters, the refusal to grant personhood would be an untenable legal position. As one author asserts, "One cannot, on conceptual grounds, rule out in advance the possibility that AIs should be given the rights of constitutional personhood."[23] This raises profound ethical and legal questions about the criteria for personhood and if one makes an assumption that intelligence, awareness, and moral reasoning are the defining criteria for bestowing personhood within the constitutional parameters to an artificial entity, then denying "Strong AI" such recognition solely based on its artificial origin would be arbitrary.

Assigning personhood may provide a legal mechanism for attributing responsibility, much like corporate personhood limits individual liability within a corporation, and recognizing AI as a juristic person will lead to a scenario where accountability could be structured in a manner that balances technological innovation and progress with legal safeguards for all parties involved. Interestingly, the process of granting 'rights' to AI entities has already begun taking nascent steps – Sophia is the first AI robot to have been granted citizenship by the Kingdom of Saudi Arabia, a 'male' AI named Shibuya Mirai

---

[23] Lawrence D. Solum, "Legal Personhood for Artificial Intelligence" 70 *North Carolina Law Review* 1261 (1992).

being given residency in Japan; and, an AI named Sam was recently made the owner of a bank account in Tokyo.[24] Granting a juristic personhood to AI at present does not necessarily mean equating it with human beings but rather it implies providing a structured legal framework to govern AI behaviour, assign liability, and establish safeguards against potential risks. Just as corporate personhood was devised to allocate legal responsibility in a business context, AI personhood could serve as a mechanism to regulate autonomous agents operating in increasingly complex environments.

To summarise, the legal recognition of AI as a juristic person is an issue that warrants serious consideration. Legal history demonstrates that personhood is a malleable concept, extended to entities beyond human beings based on necessity. AI is no longer a distant science fiction concept but an active and evolving presence in society. As AI systems become more autonomous and integral to daily life, addressing their legal status is imperative. By proactively engaging with the legal questions surrounding AI personhood, we can ensure that the law remains equipped to handle the challenges posed by emerging technologies.

4. THE MENS REA OF JURISTIC PERSONS: AN EXAMINATION OF THE PRESENT JURISPRUDENCE

The determination of liability in criminal law revolves around the twin concepts of mens rea and actus reus. The maxim "actus non facit reum nisi mens rea" roughly translates to, "there can be no crime large or small, without an evil mind"[25], is one of the cornerstones of criminal law. This means that in order to ascertain the criminal liability of a 'person', the convergence of both the actus reus element and the mens rea element is necessary. In other words, to impose criminal liability, "two elements must be proven: first, there must be an actus reus, which is the criminal conduct; second, there must be a mens rea, which is a particular internal mental state"[26] and in the absence of either of these two elements, no liability is imposed in general, under criminal law jurisprudence.

In several legal systems, such as those of France and Germany, the principle of "societas delinquere non-protest" which roughly translates to "a legal entity cannot be blameworthy" has historically prevented the imposition of criminal liability on corporations and other legal entities. This principle is rooted in the

---

24 Lizansha Birla & Raj Pipra, "Legal Identity of Artificial Intelligence" 1 *NFSU Journal of Law & Artificial Intelligence* 44 (2022).

25 Eugene J. Chesney, "The Concept of Mens Rea in the Criminal Law" 29 *Journal of Criminal Law and Criminology* 627 (1939).

26 Jake Feiler, "The Artificially Intelligent Trolley Problem: Understanding Our Criminal Law Gaps in a Robot Driven World" 14 *Hastings Science and Technology Law Journal* 06 (2023).

notion that only natural persons possess the requisite mental state or mens rea for criminal culpability.

Indian law also follows a similar path. While juristic personhood has been given to a number of entities, this recognition has bestowed a limited number of rights and often very few liabilities. For example, while an idol of a deity is considered a juristic person (Ram Lalla was deemed to be a juristic person in "M. Siddiq v. Suresh Das"[27]), its rights are limited to proprietary interests and the ability to sue or be sued. However, deities do not enjoy constitutional or fundamental rights. As a result, an idol, despite being recognized as a juristic person, does not possess constitutional personhood. The absence of mens rea precludes the imposition of criminal liability on deities, as they lack the cognitive capacity to form intent. Interestingly though, although a similar issue also arises concerning corporations, whose juristic personhood has been recognized since the 19th century,[28] existed during the initial stages, jurisprudence has evolved to attribute the mens rea of corporations to the mental state of key individuals who are responsible for running it. In this regard, Lord Denning famously remarked that "a corporation is akin to a human body, and certain individuals such as directors and managers, represent the mind of the company. Thus, when the law requires personal fault for liability, the fault of these individuals can be attributed to the corporation."[29] Under British common law, the act of imposing liability on corporations is deemed to be 'individualistic' where "a company is liable if and only if the offence can be attributed to a controlling officer and not otherwise."[30] In Canadian law, the leading position can be taken from the landmark "R v. Canadian Dredge & Dock"[31] is the primary authority, and it recognises the 'identification doctrine' and states that the 'directing mind' of the company can be held liable for ascertaining culpability for acts of the corporation.

Interestingly, Indian jurisprudence has embraced a perspective similar to Lord Denning's views and the Canadian legal position. In "The Assistant Commissioner v.. M/s. Velliappa Textiles Ltd."[32] the judiciary recognized that while corporations are artificial persons, the mens rea of individuals in charge of their affairs – their "alter egos", can be extrapolated to the corporate entity.

---

27 (2020) 1 SCC 1

28 Elizabeth Pollman, "Reconceiving Corporate Personhood" 4 *Utah Law Review* 1631 (2011).

29 H.R. Bolton (Engg.) Co. Ltd. vs. T.J. Graham, [1957] 1 QB 159.

30 Shouvik Kr. Guha & Abhyudaya Agarwal, "Criminal Liability of Corporations: Does the Old Order Need to Change?" 1 NUJS Law Review 334 (2008).

31 (1985) 1 SCR 662.

32 (2003)11 SCC 405.

Consequently, corporations can be prosecuted for offences requiring mens rea provided that the intent of their decision-makers can be attributed to them.

Thus, while juristic persons like deities and corporations share the characteristic of being non-human legal entities, their treatment under criminal law differs. Deities, lacking cognitive abilities, cannot be held criminally liable, whereas corporations, through the actions and intent of their controlling individuals, can be subjected to criminal prosecution. This reflects a pragmatic approach by the legal system to balance the conceptual limitations of mens rea with the practical need to hold juristic persons accountable. Ultimately, criminal liability for juristic persons is contingent upon human agency, reinforcing the principle that law operates through those who exercise control over the legal entities bestowed personhood by the law.

5. DECIPHERING THE MENS REA OF AI: THE POSSIBILITIES AND CHALLENGES

Famed author Issac Asimov, one of the scions of science fiction literature in the 20th century, had given three fundamental rules of robotics in 1942 - "A robot may not injure a human being or, through inaction, allow a human being to come to harm"; "A robot must obey orders given to it by human beings except where such orders would conflict with the First Law"; and, "A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws."[33] Subsequently, in 1985, he revised these, and gave another rule known as the 'Zeroth Law' which stated – "A robot may not injure humanity, or through inaction, allow humanity to come to harm."[34] In a perfectly utopian world, these rules, when transposed to AI as fundamental directives of functioning, may perhaps be enough to ensure that AI does not cause any harm to humans. However, in today's practical world, this is not the case.

5.1 THE NEED FOR DECIPHERING MENS REA IN AI-CRIMES

The rapid advancements in AI have posed unprecedented challenges to legal systems worldwide. The question of AI's liability has become particularly significant, as AI-driven technologies increasingly perform tasks traditionally reserved for humans. The primary question revolves around whether AI can be held liable for criminal acts, and if so, under what legal framework. This

---

33 Robin R. Murphy & David Woods, "Beyond Asimov: The Three Laws of Responsible Robotics" 24(4) *IEEE Intelligent Systems* 17-19 (2009).

34 Roger Clarke, "Asimov's Laws of Robotics: Implications for Information Technology" 27 *Computer* 58 (1994).

discussion extends beyond just "Weak AI", which functions within programmed constraints, to the realm of "Strong AI", which is hypothesized to possess self-awareness as well as autonomous and independent decision-making capabilities.

It is a settled position of criminal law that both mens rea and actus reus need to be proven for determining liability. The determination of the actus reus element is comparatively simple since the commission of an offence in any way – such as an autonomous AI-driven vehicle harming a pedestrian[35] or a surgical robot governed by AI making an error in surgery resulting in injury or death to the patient[36] or an AI-driven autonomous weapon being responsible for killing humans[37], in each of these examples, the actus reus aspect is clear since it is easily discernible that the AI entity took the action which resulted in the harm caused to the human being. However, the determination of mens rea proves to be more difficult because akin to human counterparts, mens rea forms a part of the abstract mind, and deciphering its presence vis-à-vis a crime is a complex process. In addition, when discussing offences committed by juristic persons, such as AI, the determination of mens rea becomes even more tangled and convoluted primarily due to the fact that, unlike human beings, a juristic person is simply an entity that has been bestowed with legal personhood through the creation of a legal fiction, and thus, cannot be deemed to possess the 'mental element' with regard to the commission of the criminal act. Since mens rea is represented in the form of knowledge and intention at the highest degrees and negligence at lower degrees[38], determining the nature of this intangible and abstract mental element becomes extremely difficult in an entity endowed with juristic personhood.

Yet, unlike other juristic persons such as corporations, rivers, idols and other such elements, an AI does possess the ability to 'think' in some capacity, albeit the same may not be the same as a human. In other words, even though the capabilities of AI are simply dismissed as "just a calculation" instead of being

---

35 Alexandra DeArman, "The Wild, Wild West: A Case Study of Self-Driving Vehicle Testing in Arizona" 61 *Arizona Law Review* 984 (2019).

36 Anoushka Sharma, "Cancer Patient in US Dies After Surgical Robot Burns Holes in Organs" *NDTV World*, 14 Feb, 2024, *available at* <https://www.ndtv.com/world-news/cancer-patient-in-us-dies-after-surgical-robot-burns-holes-in-organs-5056648#:~:text=According%20to%20the%20lawsuit%2C%20Ms,not%20informed%20to%20the%20family> (last visited on 06 March, 2025).

37 "Mohsen Fakhriazadeh: 'Machine-gun with AI' used to kill Iran scientist", *BBC*, 7 Dec, 2020 <*available at* https://www.bbc.com/news/world-middle-east-55214359 (last visited on 05 March, 2025).

38 David C Vladeck, "Machines without Principals: Liability Rules and Artificial Intelligence" 89 *Washington Law Review* 124 (2014).

deemed to be "actual intelligence"[39], and although "AI does not think the way a person does, AI is not conscious or self-aware in the same sense that a person is"[40] it is able to exhibit some cognitive capacity that resembles human intellectual capabilities. This is primarily due to the nature of AI where "high-level reasoning or everything that is hard for us and needs special skills requires much less computation and are more easy to reverse engineer and program, in comparison to low-level sensory motor skills."[41] Undeniably though, AI, unlike other entities that have been bestowed juristic personhood, stands on a different pedestal, as far as its 'thinking' capabilities are concerned, with abilities to process information and reach reasonable conclusions that would otherwise not be possible without some semblance of 'intelligence'.

Against this backdrop, it is necessary to perceive how the notion of mens rea can be determined for crimes involving AI. Crimes involving AI allude to those offences whose commission primarily depends on the involvement of an AI in some capacity, thus making it "necessary" or "essential" for the commission of the crime, therefore implying that although logically possible, the absence of AI would make the crime highly unlikely, thereby highlighting that AI is a primary contributory factor, responsible for facilitating or enhancing the commission of the offence.[42]

5.2 HALLEVY'S DIRECT LIABILITY MODEL AND AI MENS REA

Legal scholars have posited various models of criminal liability for offences involving AI. Gabriel Hallevy, one of the pioneering minds dealing with the intersectionality of AI and criminal liability has proposed three models for determining liability for crimes involving AI. These models delineate the attribution of criminal liability in various scenarios for crimes involving AI, and therefore, necessitate a closer analysis. Among these, two of his theories, the "Perpetration-via-Another Model"[43] and "Natural-Probable-Consequence-Liability-Model"[44] trace mens rea to human counterparts. The former treats the

---

39 Haroon Sheikh, Corien Prins et.al., *Mission AI: The New System Technology* 16-17 (Springer: Switzerland, 2023).

40 Ryan Abbott, *The Reasonable Robot: Artificial Intelligence and the Law* 27 (Cambridge University Press: Cambridge, 2020).

41 Vadim S. Rotenberg, "Morovec's Paradox: Consideration in the Context of Two Brain Hemisphere Functions" 55(3) *Activitas Nervosa Superior* 108 (2013).

42 T.C. King, Nikita Aggarwal et.al., "Artificial Intelligence: An Interdisciplinary Analysis of Foreseeable Threats and Solutions" 26 *Science and Engineering Ethics* 89 (2020).

43 Gabriel Hallevy, "The Criminal Liability of Artificial Intelligence Entities - From Science Fiction to Legal Social Control" 4 *Akron Intellectual Property Law Journal* 179 (2010).

44 *Ibid*, at 181.

AI as a mere instrument, taking the human, either the programmer or the user, as the source of mens rea and thus liable. For example, suppose a programmer designed an AI robot to commit arson when no one was present in a factory at night, assuming that the programming was still available and usable as evidence and in that case, the programmer will be deemed to be the perpetrator of the offence, with the AI merely the tool through which the offence was executed.[45] The latter imposes liability based on negligence and argues that if a reasonable person could foresee an AI committing a crime, but the programmers failed to mitigate such risks, they may be held accountable under accomplice liability, with this model emphasizing negligence, rather than intent, can establish culpability if foreseeable risks were ignored, making the human actor responsible for preventable AI-driven offences.

However, due to novel developments in the domain of AI, primarily on account of the ML and DL models used in modern AI, resulting in the creation of the Black Box Problem, it is difficult to reach a clear consensus on why the AI acted in a particular way. The development of an AI will inevitably have a number of programmers involved in the process, with a multiple people working on the AI for a long period of time. Considering that liability is ascertained with this model, it will be difficult to pinpoint which individual developer is liable for the offence committed by the AI, and whose individual negligence resulted in the offence. In addition, the AI might have been programmed differently, however, over the years, due to its capabilities under ML and DL, it was able to change its directives and ignore the criminality of an action, thus resulting in the offence.[46] This opacity also makes it difficult to pinpoint mens rea liability on either the programmer or the user, since it may not be clear 'beyond reasonable doubt', a cornerstone of criminal liability that mandates "criminal guilt to be proven beyond reasonable doubt"[47], that the programmer or user was responsible for the offence. Although the first model may pass the test of liability, since in that model the AI is primarily considered to be akin to a tool being used by the perpetrator to commit an offence, the second model's applicability becomes a

---

45 Gabriel Hallevy, "I, Robot – I, Criminal – When Science Fiction Becomes Reality: Legal Liability of AI Robots in Committing Criminal Offences" 22 *Syracuse Science & Technology Law Reporter* 10-11 (2010).

46 CS Chaitali Jani & Prof. Dr. S.P. Rathor, "A Legal Framework for Determining The Criminal Liability and Punishment for Artificial Intelligence" 45 *Tuijin Jishu/Journal of Propulsion Technology* 815 (2024).

47 Hock Lai Ho, "Justification, Excuse, and Proof Beyond Reasonable Doubt" 31 *Philosophical Issues* 146 (2021).

precarious proposition due to the opacity that AI possesses on account of the Black Box Problem. .[48]

This necessitates an approach where the AI itself may be held liable. This is given by Hallevy within his third model, the "Direct Liability Model" where AI entities are directly held liable for offences. This is only possible through the bestowal of juristic personhood on AI, where the personhood of AI allows it to be sued, thereby making it possible for liability to be imposed directly on the AI. Notably, this may not be applicable to the current "Weak AI" models, considering their limited capabilities, yet, the creation of "Strong AI" in any capacity may prove this model of liability to be applicable. Hallevy argues that in the event where an AI is given the status of a 'juristic person', the bestowal of criminal liability upon the AI directly does not have any legal impediments, with the AI being capable of fulfilling both the mens rea and actus reus requirements.[49] He further posits that the structure of criminal liability is built on a matrix of essential requirements, ensuring that each offence embodies only the minimum criteria necessary for imposing liability. Meeting these requirements alone is sufficient to establish criminal responsibility and the offender doesn't have to exhibit additional psychological traits such as being 'evil' or 'wicked'.[50] This approach ensures that criminal liability is based solely on rational principles rather than subjective moral judgements and through a focus on objective legal standards, the law maintains consistency and fairness in determining culpability.

A combination of these approaches given by Hallevy is also possible where both humans and AI are used in conjunction to determine liability. In this variation, a model can arise where a human and an AI entity collaborate in committing a crime, with the human agent responsible for the mens rea, while the AI carries out the actus reus. Unlike the other models, primarily the "Perpetrator-via-Another-Model" where the AI merely acts as an instrument, here, this model assumes that the AI possesses awareness of the act's legal implications, albeit it may or may not possess the mens rea, and despite the knowledge about the illegality of the act, still proceeds to partake in it. As observed by one author, "autonomous robots and artificial agents have a unique capacity to splinter a criminal act, where a human manifests the mens rea and the robot artificial agent

---

48 *Supra Note* 26, at 22.

49 *Supra Note* 43, at 187.

50 Gabriel Hallevy, *When Robots Kill: Artificial Intelligence Under Criminal Law* 32-33 (UPNE: Lebanon, 2013).

commits the actus reus."[51] This distinction raises complex legal questions about attributing liability, as the AI's awareness of illegality blurs the boundary between being a mere tool and an active participant in the criminal act.

5.3 REMOVING MENS REA FROM THE EQUATION: A STRICT LIABILITY APPROACH?

Determining the mens rea of an offence where AI is to be judged through the traditional jurisprudential lens may be a redundant effort, at least at present, due to the complexities involved in ascertaining fault and culpability. Instead, a legal system that attributes liability without focusing specifically on mens rea elements to determine 'fault' could be more effective.

The English case of "Rylands v. Fletcher"[52] is considered the fountain from which the doctrine of strict liability emerged, and it recognized "tort liability without any wrongdoing."[53] Over time though, particularly since the 19th century onwards, for "public welfare offences" and "public torts", the requirement of mens rea began to be eroded, and the proof of criminal intent dispensed through a strict liability regime.[54] The reasoning for relying on strict liability, which negates one of the two pillars of traditional criminal culpability is that such offences are established to prevent significant social harm, and imposing strict liability ensures effective deterrence and prevention of that harm.[55] Since the 19th century, cases such as "State v. Lingberg"[56], where the Court held that the reasonableness of a defendant's mistake regarding the source of the borrowed funds was not a valid defence to a felony charge that implicated directors who borrow excessive funds from the bank; and, "Regina v. Prince"[57] where the Court held that the reasonableness of the defendant's belief regarding the girl's age was irrelevant when the State barred any person from unlawfully taking an unmarried girl below the age of sixteen from the custody of her parents, firmly established the use of

---

51 Amanda McAllister, "Stranger Than Science Fiction: The Rise of A.I. Interrogation in the Dawn of Autonomous Robots and the Need for an Additional Protocol to the U.N. Convention Against Torture" 101 *Minnesota Law Review* 2572 (2017).

52 (1868) LR 3 HL 330.

53 John C.P. Goldberg & Benjamin C. Zipursky, "The Strict Liability in Fault and the Fault in Strict Liability" 85 *Fordham Law Review* 745 (2016).

54 Claire D. Johnson, "Strict Liability Crimes" 33 *Nebraska Law Review* 462 (1954).

55 Christine T. Sistare, "On the Use of Strict Liability in the Criminal Law" 17 *Canadian Journal of Philosophy* 398 (1987).

56 125 Wash. 51, 215 Pac. 4.

57 13 Cox Crim. Cas. 138 (1875).

strict liability in criminal law jurisprudence in America and Britain.[58] Today, strict liability in criminal law has received acquiescence from premier human rights institutions, including the European Court of Human Rights,[59] thus making it a mainstay of criminal law jurisprudence.

A strict liability model for AI criminal liability, where AI is granted personhood and mens rea is negated, can function by attributing responsibility solely based on the commission of an unlawful act, regardless of intent or knowledge. This approach aligns with legal precedents where liability is imposed for public welfare offences to prevent significant social harm. Scholars dealing with the intersection of AI and criminal liability have also accepted that using strict liability, especially when the AI is capable of autonomy through mechanisms such as ML, a regime where the AI is extra-careful in fulfilling objectives can be formulated.[60] By treating AI as an entity subject to strict liability, the law can hold AI itself accountable without requiring proof of criminal intent.[61] Under this framework, AI's liability would arise when its actions cause harm, regardless of whether it was pre-programmed to act maliciously or if the harm was an unintended consequence of its autonomous decision-making, thus allowing for a more streamlined approach to liability and making all parties involved, including humans during the present regime of "Weak AI" and AI, if granted personhood after the creation of "Strong AI" more accountable. Thus, a doctrine of strict liability, wherein criminal liability is imposed for any act or omission regardless of mens rea could serve as a potential solution for challenges surrounding liability of AI entities.[62]

5.4 AI GENERAL DEFENCE AS A NATURAL COROLLARY TO AI MENS REA? AN ANALYSIS

As AI continues to evolve, its potential recognition as a legal person through the bestowal of juristic personhood, complex questions about corresponding rights arise. This is because rights are a natural corollary to liabilities, and if AI is held directly accountable and liable for crimes, with the mens rea and actus reus elements directly attributed to the AI, it should also be privy to the same general defences that negate the mens rea in its human counterparts to balance the

---

58 Richard A. Wasserstrom, "Strict Liability in the Criminal Law" 12 *Stanford Law Review* 733 (1960).

59 *Salibaku v. France*, (1998) EHRR 379.

60 Gabriel Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems* 143 (Springer International: Switzerland, 2015.

61 Gyandeep Chaudhary, "Artificial Intelligence: The Liability Paradox" *ILI Law Review* 155-156 (2020).

62 Ryan Abbott & Alex Sarch, "Punishing Artificial Intelligence: Legal Fiction or Science Fiction" 52 *UC Davis Law Review* 351 (2019).

imposition of liability. As one author has argued in the context of corporations, "if corporations have no guaranteed due process rights, then they can be fined for having committed crimes without the benefit of trial"[63], extrapolating this argument to the present issue, it is also similarly important to recognize certain protectionary criminal law rights, primarily the right to due process in some capacity, to ensure that the attribution of criminal liability or the determination of mens rea for AI fulfills the norms of criminal law jurisprudence.

Furthermore, an argument that the AI does not possess consciousness and therefore cannot be given any rights in accordance with those bestowed upon human beings can also be a controversial opinion if juristic personhood of the AI is accepted. Conscious AI entities cannot be dismissed merely due to present technological limitations as these may not persist indefinitely. As one author has suggested, "the empirical finding that novel types of entities develop some kind of self-consciousness and become capable of intentional actions seems reasonable, as long as we keep in mind that the emergence of such entities will probably require us to rethink notions of consciousness, self-consciousness and moral agency."[64] This suggests that non-biological entities governed by AI could eventually challenge the existing notions of personhood, requiring legal systems to adapt accordingly. Therefore, if AI is to be held liable for criminal acts, where the mens rea is directly attributed to the AI, it would align with justice and rule of law to extend the general defences to the AI as well. In such cases, the AI's actions could be perceived as self-defence, where applicable, negating mens rea elements like intention and knowledge in the same way it would apply for a human defendant.

Finally, imposing criminal liability on AI necessitates rethinking traditional culpability frameworks, particularly in the case of "Weak AI" systems. The doctrine of "dolus incapax", meaning "incapable of criminal intention or malice; not of the age of discretion; not possessed of sufficient discretion and intelligence to distinguish between right and wrong to the extent of being criminally responsible for his actions"[65] is undoubtedly something that can be made applicable to "Weak AI" systems. This is because even with knowledge and 'intent' of some sort, which fulfills the mens rea requirements, AI may lack the ability to distinguish between ethical or moral right and wrong, functioning solely

---

63 Adam Winkler, "The Long History of Corporate Rights" 98 *Boston University Law Review Online* 68 (2018).

64 Bert-Japp Koops et.al., "Bridging the Accountability Gap: Rights for New Entities in the Information Society" 11 *Minnesota Journal of Law, Science and Technology* 558-559 (2010).

65 Henry Campbell Black, *Black's Law Dictionary* 570 (West Publishing Co.: Minnesota, 1968).

on logic. While Hallevy has argued that notions of 'good' or 'evil' are immaterial for applying criminal liability as far as the constituent elements are fulfilled[66] applying the same standards of liability without giving the same defences would be a misnomer as far as the application of the law is concerned.

Therefore, this argument asserts that if AI is deemed capable of possessing mens rea through the grant of juristic personhood, it must also be entitled to the general defences available under criminal law. Recognizing only liability without corresponding rights would create an imbalance in legal reasoning, especially since the liabilities have been recognised through the bestowal of personhood. To ensure fairness and consistency within jurisprudence, any application of mens rea standards to AI must necessarily include access to appropriate legal defences, maintaining an equitable framework.

6. CONCLUSION

The rapid advancement of AI presents unprecedented challenges to traditional legal frameworks, particularly in determining criminal liability. As AI continues to evolve, questions surrounding mens rea in AI-driven offences become more complex, necessitating a re-evaluation of existing principles of criminal responsibility. While actus reus can be clearly identified in cases involving AI, the absence of a conscious mind, which is also autonomous and can be beyond the understanding of its human overseers, such as programmers and users, as characterized by the Black Box Problem, it complicates the attribution of mens rea to human counterparts. While scholars such as Gabriel Hallevy have tried to accommodate existing principles of criminal liability with the nuances brought forth by AI to attribute liability to human counterparts, this paper has attempted to examine the possibility of a legal regime of direct liability, where the AI is granted juristic personhood, and both actus reus and mens rea are derived from and attributable to the AI itself.

The current jurisprudence does not prove to be an impediment in recognizing the juristic personhood of AI. Furthermore, considering that the matrix of mens rea necessitates the presence of certain elements, without additional elements of good or evil, its fulfilment, primarily by "Strong AI" can fulfill the requisite characteristics of attributing mens rea to AI. It is also proposed that one way of bypassing the requisite of mens rea, for AI-driven crimes, would be through a regime of strict liability, wherein the mental element of the offence is negated, and liability can be imposed directly if the factual elements fulfill the offence's requirements. This will streamline the process of bestowing liability. However, as

66 Supra Note 60, at 32-33.

a natural corollary to recognizing the liability of AI, it is also important to ensure that certain safeguards, in line with due process, are given to AI entities, thereby maintaining the balance.

Finally, to form a comprehensive legal regime, two suggestions can be given. First, a regime of "Explainable AI", which can be defined as "to explain the logic behind black-box models"[67] can be legally mandated for both "Weak AI" and "Strong AI" models where the inner workings of the AI that are precluded from human understanding through the Black Box Model become clear, thus allowing for easier attribution of mens rea, to either the humans involved or to the AI generally. Second, in applying the "Direct-Liability-Model", it is necessary to exercise caution to prevent its misuse by human counterparts[68] and therefore, any model of AI liability needs to be formulated by keeping in mind the possibility of its abuse by humans.

---

67 Chinu Dhir & Urvashi Bansal, "Explainable AI: To Reveal the Logic of Black-Box Models" 42 *New Generation Computing* 54 (2024).

68 Pompeu Casanovas, Ugo Pagallo et.al. (eds.), *AI Approaches to the Complexity of Legal Systems* 150-151 (Springer: Berlin Heidelberg, 2014).