A CRITIQUE OF GABRIEL HALLEVY'S MODELS OF CRIMINAL LIABILITY OF ARTIFICIAL INTELLIGENCE ENTITIES*

Abstract

The rapid technological change in the society has tremendously influenced our lives as human beings. Machines are increasingly becoming more sophisticated in their functions. The positive impacts of these artificial entities are undisputable in the society. Robots and computers are gradually replacing human activities, ranging from autonomous cars to machine translation software, robots and medical diagnosis software. From the homes to hospitals, and other public spaces, there is no denying the rapid growth and impact of Artificial Intelligence (AI) entities. However, the rise of Artificial Intelligence entities also raises questions about liability for crimes associated with them. Though Criminal law embodies the most powerful legal social control to regulate crimes, yet the concern of people in most cases is based on the fact that Artificial intelligence entities are not ordinarily considered to be subjects of law. Unfortunately, there are no enough pieces of legislation and regulations addressing the question of liability in relation to Artificial Intelligence entities. Be that as it may, scholars have made attempts to address the thorny issue. One such scholar is the Israeli Professor of Criminal Law, Gabriel Hallevy, whose basic question for consideration is: Does the growing intelligence of AI entities subject them to legal social control as any other legal entity? Hallevy developed three models of response to the subject, namely, the perpetration-via-anotheer liability model, The Natural-Probable-Consequence Liability Model' and The Direct Liability Model. This study simply examines these models and draws a response. The conclusion is that Hallevy's models are novel but are not sufficiently capable of addressing existing questions on the culpability of the AI entities.

Keywords: Artificial Intelligence Entities, Models of Criminal Liabilty, Gabriel Halevy, Critique

1. Introduction

The world is evolving towards Artificial Intelligence (AI). The future is gradually tilting towards internet of things (IOT) which allows information to be sent to and received from objects and devices, using the internet.[1] Some researchers even postulated that these machines are destined to take over the world.[2] However, this evolving technology raises questions about liabilities for crimes an AI 'commits', mainly because the AI acts autonomously and with limited control from humans.[3] Research recorded numerous deaths around the world caused by the operations of AI entities.[4] For instance, a total of 26 deaths caused by robots' malfunctions was recorded over the past 30 years in the United States, while the United Kingdom recorded 77 robot related fatal accidents in 2005.[5] There was also a record of 37-year old Japanese employee of a motorcycle factory who was killed in 1981 by an AI robot working near him.[6] In error, the robot identified the employee as a threat to its mission, and calculated that the most efficient way to eliminate the threat was to push him into an adjacent operating machine. With its powerful hydraulic arm, the robot smashed the worker into the operating machine, killing him instantly, and then resumed its duties having removed all interference to its mission.[7] The legal questions then are: who is to be held liable for this cold blooded, premeditated murder?[8] Is it possible to apply criminal liability on AI entities, not ordinarily seen as subjects of law? Are AI entities fit to be accorded the legal personhood in order to make them culpable of associated crimes? What type of punishment is appropriate on an AI entity for its crime?

Bringing the matter homewards, will lawmakers in Nigeria, for instance, be required to make new laws or are the present ones sufficient? Will the courts be required to formulate or adopt new rules of adjudication? Will this be a desirable approach or an existential necessity? Conventional wisdom and the *status quo* hold that punishing AI

*** By K. OGUNNOIKI, BA (Hons.), BL, LLM, PhD Candidate**, Principal Partner: Sammy & Kemmy Law Partners, Lagos, Nigeria. E-mail: keminoiki@yahoo.com; and

*** Ikenga K.E. ORAEGBUNAM, PhD (Law), PhD (Phil.), PhD (Rel. & Soc.), PhD (Edu. Mgt.) (In view), MEd, MA (Rel & Soc). MA (Phil), BTh, BA, BPhil, BL**, Professor and Formerly Head, Department of International Law and Jurisprudence, Faculty of Law, Nnamdi Azikiwe University, P.M.B. 5025, Awka, Anambra State, Nigeria. Email: ikengaken@gmail.com; ik.oraegbunam@unizik.edu.ng. Phone Number: +2348034711211.

[1] Marriam-Webster Dictionary https://www.merriam-webster 3rd May 2022

[2] R Leenes and F. Lucivero 'Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design.'(2014) 6(2) *Law, Innovation and Technology* 193-220

[3] R Abbott, A Sarch 'Punishing Artificial Intelligence: Legal Fiction or Science Fiction (2019)53 (1) *UC Davis law Review* https://lawreview law.ucdavies.edu/issues/53/1/articles

[4] S M Solaiman, 'Legal Personality of Robots, Corporate, Idols and Chimpanzees: A Quest for Legitimacy' (2017) 25(2) *Artificial Intelligence and Law*158

[5] Ibid

[6] G Hallevy, 'The Criminal Liability of Artificial Intelligence Entities: From Science Fiction to Legal Social Control' (2010) 4 (2) *Akron Intellectual Property Journal*, 172

[7] Ibid

[8] Ibid

is incongruous with basic criminal law principles such as the capacity for culpability and the requirement of a guilty mind[9], that is, the presence of the *actus reus and mens rea*. Given this traditional principle of criminal law, can a machine commit or be culpable of a crime? In a seminal study, entitled 'The Criminal Liability of Artificial Intelligence Entities: From Science Fiction to Legal Social Control',[10] Gabriel Hallevy gave a reasoned response to the subject matter. This article examines Gabriel Hallevy's Models of culpability of Artificial Intelligence Entities.

2. Analyses of Key Terms

It may be *apropos* to clarify the meaning of the following key terms for a better understanding of the discourse.

Intelligence

Scherer notes that 'the difficulty in defining AI lies not in the concept of artificiality but rather in the conceptual ambiguity of intelligence'.[11] It is trite that there is no universally acceptable definition of intelligence. According to McCarthy, intelligence is defined as 'the computational part of the ability to achieve goals in the world.'[12] Gudwin[13] states that 'intelligent systems are expected to work, and work well in many different environments. Their property of intelligence allows them to maximize the probability of success even if full knowledge of the situation is not available.' Similarly, intelligence was defined as 'the ability for an information processing system to adapt to its environment with insufficient knowledge and resources'.[14] Legg and Hutter [15] succinctly hold that 'intelligence measures an agent's ability to achieve goals in a wide range of environments.' We can deduce three common features from these definitions: that intelligence is (i) a property that an individual agent has that interacts with its environments. ii) related to the agent's ability to succeed or profit with respect to some goals or objectives, and iii) depends on how able that agent is to adapt to different objectives and environments.[16] These definitions point to the fact that intelligence is not the ability to deal with a known environment but rather, it is the ability to deal with some range of possibilities which cannot be wholly anticipated.[17]

Artificial Intelligence

Artificial Intelligence as a term was introduced by John McCarthy in 1956. He defined it as 'the science and engineering of making intelligent machines, especially intelligent computer programs.'[18] By intelligent machines, John McCarthy was referring to the 'computational part of the ability to achieve goals in the world'.[19] Artificial Intelligence is further defined as the simulation of human behaviour and cognitive process on a computer and hence it is the study of the nature of the whole space of intelligent minds.[20] Authors such as Abbot and Sarch [21] refer to Artificial Intelligence as a machine that is capable of completing tasks otherwise typically requiring human cognition.[22] Gabriel Hallevy[23] defines AI as 'the simulation of human behaviour and cognitive processes on a computer and hence is the study of the nature of the whole space of intelligence minds.'[24]

From the definitions, this work highlights some key features common to all AI entities, which are:

[9] R Abbott, A Sarch 'Punishing Artificial Intelligence: Legal Fiction or Science Fiction' (2019) 53 (1) *UC Davis law Review* https://lawreview law.ucdavies.edu/issues/53/1/articles 101

[10] G Hallevy, 'The Criminal Liability of Artificial Intelligence Entities: From Science Fiction to Legal Social Control', Akron Intellectual Property Journal, deaexchange.uakron.edu. Accessed 5th September 2022.

[11] M Scherer 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies (2016)29 *Harvard Journal of Law &Tech* 353 at 359

[12] J McCarthy, *'What is Artificial Intelligence?'* (Stanford University, Computer Science Department 2004) p354 <http://www-formal.stanford.edu/jmc/whatisai.html> accessed May 4th 2020

[13] R. R Gudwin, 'Evaluating Intelligence: A Computational Semiotics Perspective' in IEEE International Conference on Systems, Man and Cybernetics (Nashville: Tenessee, USA,2000)2080-2085.

[14] P Wang 'On the Working Definition of Intelligence'. *Technical Report 94,* Centre for Research on Concepts and Cognition, Indiana University, 1995

[15] S Legg and M Hutter, 'A Formal Measure of Machine Intelligence' in 15th Annual Machine Learning Conference of Belgium and the Netherlands (Benelearn'06) Ghent, 2006

[16] S. Legg and M Hutter, A Collection of Definitions of Intelligence,(2007) 157 *Frontiers in Artificial Intelligence and Applications* 17-24 <http://www.idsia.ch/~shane/intelligence.html.> accessed 4th October 2020

[17] ibid

[18] J McCarthy, 'What is Artificial Intelligence? *Rev* (2007) Stanford University, Computer Science Department http://www-formal.stanford.edu/jmc/whatisai.html > accessed on 4 May 2020

[19] Ibid

[20] NP Padhy, 'Artificial Intelligence and Intelligent Systems' (Oxford: Oxford University Press, 2005)

[21] R Abbot and A Sarch, 'Punishing Artificial Intelligence: Legal Fiction or Science Fiction' (2019) 53 *UC Davies Law Review* 1, 323

[22] Ibid

[23] G Hallevy; 'The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control' (2010) 4 (2) *Akron Intellectual Property Journal* 171-201

[24] S Russell and P Norving, Artificial Intelligence: A Modern Approach (4th edn, NJ Prentice Hall 2009) 4-5

i) *Autonomy*:[25] An AI may cause harm without being directly controlled by an individual. Humans are only limitedly involved, or in the future not involved at all in the decision making of an AI.[26]Autonomy is one of the most relevant features of software agents such as the AI. The autonomy differs between different fields of AI. For instance, from the autopilot mode in autonomous cars where the driver is required to stay in charge of the car, to the high frequency trading algorithms that function without humans engaging in their activities.[27] As noted by Floridi and Sanders,[28] these artificial agents are sufficiently informed, 'smart', autonomous and able to perform morally relevant actions independently of the humans who created them. Yang et al[29] argued that this combination of autonomy and learning skills underpins, both beneficial and malicious uses of AI.

ii) *Unpredictability*:[30] Some leading AIs rely on machine learning or similar technologies which involve a computer program initially created by individuals, further developing in response to data without explicit programming. This is one means by which an AI can engage in activities its original programmers may not have intended or foreseen. [31] It may react totally differently than a human facing exactly same situation. The outcome of the AI could be unpredictable when the conduct is not a result of an instruction from the programmer, but a self-learned strategy.

iii) *Unaccountability:*[32] According to Mireille Hildebrandt, as long as AI lacks legal personality, they can behave in a way that if it were human, it would have legal consequence. It may be possible to determine what an AI has done, but not how or why it acted as it did.[33]This has led to some AIs been being described as 'black box' systems.[34]For instance, an algorithm may refuse a credit application but not be able to articulate why the application was rejected.[35]

However, Russel and Norving[36] opine that these traditional definitions of AI seem to be narrow in scope, wavering between computer as a machine and as a program, and ignore other platforms such as aircraft, drones and satellites.[37] The duo therefore define AI as 'the mechanical simulation system of collecting knowledge and information and processing intelligence of universe: (collating and interpreting) and disseminating it to the eligible in the form of actionable intelligence'.[38] This study uses the term Artificial Intelligence (AI) to refer to a machine that is capable of completing tasks otherwise typically requiring human cognition.[39]

Crime

Black's Law Dictionary[40] defines a crime as 'an act that the law makes punishable; the breach of a legal duty treated as the subject matter of a criminal proceeding' [41] At this stage, certain questions are apt. When is a crime committed? In Nigeria as with most parts of the world, a crime is committed when a person recognised by the law and who is not statutorily excluded from being criminally culpable, does an act or makes an omission defined by the statutes to be an offence, with the required criminal knowledge or intent. A crime consists of two elements: a voluntary criminal act or omission (*actus reus*) and an intention to commit a crime (*mens rea*). When then could an AI entity be said to have committed a crime? This paper attempts to answer this question by making copious

[25] R Calo, 'Robotics and the Lessons of Cyberlaw Review (2015)103 *Califonia Law Review* 513, 532

[26] N Johnson et al., 'Abrupt Rise of New Machine Ecology Beyond Human Response Time', *Science Reports* September 11, 2013

[27] M Scherer 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies (2016)29 *Harvard Journal of Law &Tech* 363

[28] L Floridi, L .and J.W Sanders (2004). 'On the morality of artificial agents.'14 (3) *Minds and Machines* 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

[29] G. Z Yang, et al. (2018) 'The Grand Challenges of Science Robotics' 3 (14) *Science Robotics,* 7650 https://doi.org/10.1126/scirobotics.aar7650.

[30] M Scherer, n28

[31] C DeBrusk, 'The Risk of Machine-Learning Bias (and How to Prevent it)' *MIT Sloan Management Rev.* March 26, 2018 https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/

[32]M Hildebrandt, 'Criminal Liability and Smart Environments' in RA Duff and Stuart P Green(eds) *Philosophical Foundations of Criminal Law* (Oxford: Oxford University Press 2011) p.201

[33] D Castelvecchi, 'Can We Open the Black Box of AI? *Nature,* October 5 2016https://www.nature.com/news/can can-we-open-the-black box-of -ai -1.20731

[34] Ibid

[35] Ibid

[36] Ibid

[37] Ibid

[38] Ibid

[39] R Abbot and A Sarch, n11

[40]Black's Law Dictionary (9th edn2009)

[41] Ibid 427

references to the three liability models suggested by Gabriel Hallevy[42] in determining when and who bears liability for acts or omissions of an AI entity.

3. Requirements for Criminal Liability: *Actus Reus And Mens Rea*

In order to impose criminal liability upon a person, two main elements must exist simultaneously.[43]The first is the external element also called the physical criminal conduct (*actus reus*) and the other is the internal or mental element (*mens rea*). If one element is missing, no criminal liability can be imposed. The *actus reus* requirement is expressed mainly by acts or omissions while the *mens rea* requirement is expressed by mental elements such as knowledge or intention[44] The simultaneous existence of *actus reus* and *mens rea* are sufficient in order to impose criminal liability. A spider is capable of acting, but it is incapable of formulating the *mens rea* requirement, hence it bears no criminal liability. Likewise, though a parrot is capable of repeating the words it hears, yet it is incapable of formulating the *mens rea* requirement for libel.[45] The relevant question then is: how can AI entities fulfil the two requirements of criminal liability?

4. Gabriel Hallevy's Models of Culpability of Artificial Intelligence Entities

Gabriel Hallevy[46] proposes the imposition of criminal liability on AI entities using three liability models:

Perpetration-Via Another Liability Model[47]

Hallevy states that when a crime involves an AI entity, the AI entity should not be considered as possessing any human attributes but as an innocent agent.[48]Due to that view point, a machine is considered as a machine and is never a human. That is, the AI entity is just a mere instrument in the commission of that crime and not an active (principal or secondary) participant. In this case, due to lack of *mens rea* of the actual perpetrator, the criminal charge will always pursue the producer, the programmer or the end-user of that particular AI entity. According to Halleyy, these capabilities resemble the parallel capabilities of a mentally limited person, such as a child,[49] a person who is mentally incompetent[50] or one who lacks a criminal state of mind.[51] Legally, when an offence is committed by an innocent agent (a child, a person who is mentally incompetent, or one who lacks a criminal state of mind to commit an offence) that person is not criminally liable as a perpetrator-via-another. In such cases, the intermediary is regarded as a mere instrument, though a sophisticated instrument, while the party orchestrating the offence (the perpetrator-via-another) is the real perpetrator as a principal in the first degree and is held accountable for the conduct of the innocent agent.[52]The perpetrator's liability is determined on the basis of the 'instrument's' conduct[53] and his mental state.[54]

The derivative question relating to AI is who is the perpetrator-via-another? There are two possible candidates here: The first candidate is the programmer of the AI software while the second candidate is the user or the end-user.[55] It is possible for a programmer of a software to design a program in order to commit offences through the AI entity. For instance, a programmer designs a software for an operating robot. The robot is intentionally placed in a factory, and its software is designed to torch the factory at night when no one is there. The robot committed the arson, but the programmer is deemed to be the perpetrator.[56] The second person who might be considered the perpetrator- via-another is the user of the AI entity. The user did not program the software, but he uses the AI entity including its software for his own benefit. For example, a user purchases a servant-robot, which is designed to execute any order given by its master. The robot identifies the specific user as its master, and the master orders the robot to assault any invader of the house. The robot executes the order exactly as instructed. This is not different than a person who orders his dog to attack a trespasser. The robot committed the assault, but the user is

[42] Ibid

[43] G Hallevy, 'I, Robot, I -Criminal -When Science Fiction Becomes Reality: Legal Liability of Robots Committing Criminal Offences' (2010) 22 *Syracuse Sci.&Tech. L. Rep*1-37

[44] R M. Perkins, 'Knowledge as a Mens Rea Requirement' (1978) *29 Hastings L.J. 953,* United States v. Youts, 229 F.3d 1312, 1317 (10th Cir.2000); State v. Sargent ,594 A.2d 401 (Vt.1991) State v. Wyatt,482 S.E.2d 147 (W.Va 1996); People v. Steinberg, 595(N.E.2d 845 (N.Y.1992)

[45] G Hallevy n13

[46] G Hallevy; n13, 172

[47] Ibid p179-183

[48] Ibid, p179

[49] Ibid, *Maxey v. United States,* 30 App.D.C 63,80 (App.D.C.1907)

[50] Ibid, *Johnson v. State,*142 Ala.70,71 (1904)

[51] Ibid, *United States v. Bryan*,483 F.2d 88,92 (3dCir.1973)

[52] Hallevy n13 p129

[53] Ibid, *Dusenbery v. Commonwealth,* 772 263 S.E.2d 392(Va.1980)

[54] Ibid, *United States v. Tobon-Builes,* 706 F.2d 1092,1101 (11th Cir.1983)

[55]Hallevy n13 p129

[56] Ibid p180

deemed the perpetrator.[57] In both cases, the actual offence was committed by the AI entity. The programmer or the user did not perform any action conforming to the definition of a specific offence; therefore, they do not meet the *actus reus* requirement of the specific offence.[58]The perpetrator-via-another liability model considers the action committed by the AI entity as if it had been the programmer's or the user's action. The legal basis for liability is the instrumental usage of the AI entity as an innocent agent.[59] When programmers or end-users use an AI entity instrumentally, the commission of an offence by the AI entity is attributed to them. The internal element required in the specific offence already exists in their minds.[60] The programmer had criminal intent when he ordered the commission of the arson, and the user had criminal intent when he ordered the commission of the assault, even though these offences were actually committed through a robot, an AI entity. When an end-user makes instrumental usage of an innocent agent to commit a crime, the end-user is deemed the perpetrator. [61]

This liability model does not attribute any mental capability to the AI entity. According to this model, there is no legal difference between an AI entity and a screwdriver or an animal. When a burglar uses a screwdriver in order to open up a window, he uses the screwdriver instrumentally, and the screwdriver is not criminally liable. The screwdriver's 'action' is in fact the bugler's. This is the same legal situation when using an animal instrumentally.[62] An assault committed by the order of its master is in fact, an assault committed by the master. This kind of legal model might be suitable for two types of scenarios. The first scenario is using an AI entity to commit an offence without using its advanced capabilities. The second scenario is using a very old version of an AI entity which lacks the modern advanced capabilities of the modern AI entities. In both scenarios, the use of the AI entity is instrumental usage. However, it is the usage of an AI entity, due to its ability to execute an order to commit an offence.[63] It should be noted that this model assumes that the AI is completely dependent on either the programmer or the user. it is not self-ruling or self -determining, but solely an instrument for which no specific mental state is required. [64] However, exceptions to this model can be seen where an AI entity decides to commit an offence based on its own accumulated experience or knowledge.[65] Perpetration-via-another liability model is not suitable when the software of the AI entity was not designed to commit the specific offence, but was committed by the AI entity nonetheless.[66] The model is also not suitable when the specific AI entity functions not as an innocent agent, but as a semi-innocent agent.[67] The legal result of applying this model is that the programmer and the user are criminally liable for the specific offence committed, while the AI entity has no criminal liability whatsoever.[68]

The Natural-Probable-Consequence Liability Model

The second model of criminal liability assumes deep involvement of the programmers or users in the AI's daily activities, but without any intention of committing any offence via the entity.[69] For example, during the execution of its daily tasks, an AI entity commits an offence. The programmers or users had no knowledge of the offence until it had already been committed; they did not plan to commit any offence, and they did not participate in any part of the commission of specific offence.[70] One example of such a scenario is an AI robot, or software, which is designed to function as an automatic pilot. The AI entity is programmed to protect the mission as part of the mission of flying the plane. During the flight, the human pilot activates the automatic pilot (the AI entity) and the program was initialized. At some point after activation of the automatic pilot, the human pilot sees an approaching storm and tries to abort the mission and return to base. Unfortunately, the AI entity deems the human pilot's action as a threat to the mission and takes action in order to eliminate that threat. It might cut off the air supply to the pilot or activate the ejection seat, etc as a result, the human pilot is killed by the AI entity's actions.[71] Obviously, the programmer had not intended to kill anyone, but nonetheless, the human pilot was killed as a result of the AI entity's actions, and these actions were done according to the programming. Another example is that of an AI software which was designed to detect threats from the internet and protect a computer from such threats. Few

[57] Ibid

[58]J Dressler and SE Garvey, *Cases and Materials on Criminal Law* (6TH edn, West Academic Publishing 2012)126

[59] L B. Solum, Legal Personhood for Artificial Intelligence, 70 *N.C.L.REV.*1231 (1992)

[60] G Hallevy; n13, 181

[61]Ibid

[62]Ibid p181

[63] T L. Butler, 'Can a Computer Be an Author-Copyrights Aspects of Artificial Intelligence, (1982)4 *Comm. Ent. L.S.* 707

[64] PM Freitas, FA and P N, 'Criminal Liability of Autonomous Agents: from the Unthinkable to the Plausible', Law School, Universidade do Minho, Braga, Portugal <pfreitas@direito.uminho.pt > accessed 3rd October 2021

[65] N Lacey and C Wells, Reconstructing Criminal Law-Critical Perspectives on Crime and the Criminal Process 53 (2nd ed. 1998)

[66] Ibid

[67] Ibid

[68] *People v. Monks*, 133 Cal. App.440,446 (Cal. Dist. Ct. App. (1933)

[69] Hallevy n13, p181

[70] Ibid p182

[71] Ibid

days after the software was activated, it figured out that the best way to detect such threats was by entering websites it defined as dangerous and destroyed any software it recognized as a threat. When the software did that, it committed an offence, although the programmer did not intend for the AI entity to do so.[72] In the examples above, the programmers or users had no knowledge of the offence; they had neither planned it or intended to commit the offence using the AI. In such cases, this second model might create a suitable legal response.[73] The second model is based upon the ability of the programmers or users to foresee the potential commission of offences. It postulates that a person be held accountable for an offence, if that offence is a natural and probable consequence of that person's conduct.[74]

Originally, the natural -probable -consequence liability model was used to imposed criminal liability upon accomplices, when one committed an offence, which had not been planned by all of them and which was not part of a conspiracy. Natural -probable -consequence liability seems legally suitable for situations in which an AI entity commits an offence, while the programmer or user had no knowledge of it, had not intended it, and had not participated in it.[75] The programmer or user is seen to be in a negligent mental state, not more.[76] Programmers or users are not required to know about any forthcoming commission of an offence as result of their activity, but are required to know that such an offence is a natural, probable consequence of their actions.[77] A negligent person, in a criminal context, is a person who has no knowledge of the offence, but a reasonable person should have known about it since the specific offence is a natural probable consequence of that person's conduct. [78] Negligence is, in fact, an omission of awareness or knowledge.[79]The programmers or users of an AI entity, who should have known about the probability of the forthcoming commission of the specific offence and prevent it from being committed by the AI entity, but did not do so, are criminally liable for the specific offence, even though they did not actually know about it.[80]

However, the legal implication of applying the natural-probable-consequence-liability model to the programmer or user differs in two different factual cases. The first type of case is when the programmers or users were negligent while programming or using the AI entity but had no criminal intent to commit any offence.[81] The second is when the programmers or users programmed or used the AI entity knowingly and wilfully in order to commit one offence via the Ai entity, but the AI deviated from the plan and committed some other offence, in addition to, or instead of the planned offence.[82] The first type of case is one of negligence.[83] As in the above example, where a programmer of an automatic pilot negligently programmed it to defend its mission with no restrictions on the taking of human life, the programmer is negligent and liable for the homicide of the human pilot.[84] In the second type case, the programmer shall be held criminally liable for both the offence he originally programmed it to commit and also for the offence it committed in addition to or instead of the original programming.[85] The dangerousness of the very association or conspiracy whose aim is to commit an offence is the legal reason for imposing more severe accountability upon the programmer.[86] For example, a programmer programs an AI entity to commit a robbery in a bank but did not program the AI entity to kill anyone who resisted the robbery. But unfortunately, the AI entity killed someone in the bank who resisted the robbery.[87] In such cases, the danger posed by such a situation exceeds the criminal negligence liability. Hence, the programmer shall be held criminally accountable for the robbery (if committed), as well as for the killing, just as an offence of manslaughter or murder, which requires knowledge and intent.[88] In this scenario, human activity is merely linked to the malfunction of the AI entity in the manner that the programmer or the user should have considered the possible consequence of a crime being committed in certain circumstances by the AI entity. This work therefore agrees with Hallevy that considers the criminal liability of the human factor as negligence, rather than intention, although there may be

[72] Ibid

[73] Ibid

[74] Ibid p 183

[75] *United States v. Andrews*, 75 F.3d 552 (9th Cir. 1996)

[76] Ibid

[77] Ibid

[78] D Stuart, 'Mens Rea, Negligence and Attempts, 1968 *Crim.L.Rev.*647 (1968)

[79] Dressler and Garvey, n56, Model Penal Code§2.02.

[80] Hallevy n13 p184

[81] ibid

[82] ibid

[83] Dressler, and Garvey n56

[84] Hallevy n13 p184

[85] Ibid p185

[86] Ibid

[87] Ibid

[88] *United States v. Greer*, 467 F.2d 106444,1069 (7th Cir.1972); *People v. Cooper*, 743 N. E2d 32,36 (III.2000)

situations when the human offender foresees the result of its actions (upon the AI entity), does not pursue it, while accepting this result to occur one day.[89]

However, the question still remains: what is the criminal liability of the AI entity itself when the natural-probable-consequence liability model is applied? [90] There are two possible answers: firstly, if the AI entity acted as an innocent agent, totally oblivious of the criminal prohibition, it is not held criminally accountable for the offence committed, as the action of the AI entity is not different from perpetration-via-another liability model. [91] But if the AI entity did not act merely as an innocent agent, then, in addition to the criminal liability of the programmer or user, pursuant to the natural-probable-consequence liability model, the AI entity itself shall be held criminally liable for the specific offence directly.[92]

The Direct Liability Model

In this third model, focus is on the AI entity itself, rather than assume any dependence of the AI entity on a specific programmer or user.[93] In order to impose criminal liability for a specific offence, both the external element (*actus reus*) and the internal element (*mens rea*) of that offence must be present simultaneously. Anyone found with both elements of the specific offence is held criminally culpable for that offence.[94] Hence, in order to impose criminal liability on any kind of entity (including AI), the existence of these elements must be proven.[95] The new technology developments prove that AI entities are able to interpret large amounts of data from its sensors, able to differentiate between right and wrong, and even to analyse what is permitted or forbidden.[96] Hence, as long as an AI entity is capable of the above, it should be culpable for its crimes. Thus, if an AI robot activates its electric or hydraulic arm and moves it, such movement of an AI robot that hits a person is considered as fulfilling the *actus reus* requirement of the offence of an assault. [97]Also, in the offence of an omission, the inaction of the AI entity is the legal basis for criminal liability, as long as there had been a duty to act and it fails to act.[98]

The question arises, if a person who fulfils the requirements of both the external and the internal elements of a specific offence is held criminally liable, why then should an AI entity that fulfils all elements of an offence be exempted from criminal liability?[99] The criminal liability of an AI does not replace the criminal liability of the programmers or the users, if criminal liability is imposed on the programmers and/or users by any other legal path.[100] One may opine that criminal liability should be imposed on the AI entity in addition to the criminal liability of the human programmer or user[101] as criminal liability is not to be divided, but are combined.

5. A Critique of Gabriel Hallevy's Liability Models

This study sees the three liability models as postulated by Hallevy as constituting a helpful starting point in determining the issue of culpability in relation to AI entities. However, the models attracted some criticisms from several writers. Scherer [102] opines that the models fail to recognise the complex processes of how AI entities are built while Beard[103]states that technological developments are collaborative and polymorphic. For instance, in an hypothetical situation, Chuks, a technology magnate, owns Creativity Technologies Ltd (CTL). CTL also employs 2,000 programmers across the globe as engineers. CTL engineers programmed an AI entity, Heaven's Grace (HG) which is a self-driving car. CTL outsources HG hardware to a manufacturer, Affordable Motors Ltd (AML), known for its cheap rates but has a history of violating safety features. HG accidentally kills Mrs Jude, an old widow. While it is possible to discern Chuks as the Head Programmer, his role was within the loop of the logic module project is unclear. While certain acts can be attributed to the heads of each company, attributing each line of code and task to individual programmers is a huge task.[104] Hallevy models fail to note that AI is not just software based; it fails to consider the criminal liability of each of the hardware manufacturers. In the hypothesis

[89] Maxim Dobrinoiu, The Influence of Artificial Intelligence on Criminal Liability, Faculty of Law, *Nicolae Titulescu,* University of Bucharest http://cks.univnt.ro accessed 8th Nov 2021

[90] Hallevy n13 p185

[91] Ibid

[92] *State v. Kaiser*, 918 P.2d 629,637 (Kan 1996); *United States v. Andrews,* 75 F.3d 552,556 (9th Cir.1996)

[93] Maruerite E. Gerstner, 'Liability Issues with Artificial Intelligence Software' (1993) 33 *Santa Clara L. Rev*.239

[94] Dressler and Garvey n56, p126

[95] Ibid

[96] M Dobrinoiu, 'The Influence of Artificial Intelligence on Criminal Liability', Faculty of Law, *Nicolae Titulescu,* University of Bucharest http://cks.univnt.ro accessed 8th Nov 2021

[97] Dressler and Garvey n56 p126

[98] Ibid

[99]Hallevy n13 p189

[100] Ibid

[101] Ibid p191

[102]M Scherer, n28 p 359

[103] J Beard, 'Autonomous Weapons and Human Responsibilities' (2014)45 *Georgetown J Int Law* 647 at 662

[104] Ibid

above, AML would not be charged under any of Hallevy's three models. However, from the hypothesis above, AML has an history for skimping on safety features, which may have contributed to the accident that resulted in the death of Mrs Jude.

Another criticism on Hallevy's liability model arises from the fact that AI code can be open source software.[105] Open source software is where the original creator 'surrenders all… rights granted by copyright', allowing anyone to study, change and redistribute it,[106] hence, making the AI entity vulnerable to modifications by third parties. Melz[107]is of the view that open-sourcing AI promotes effective peer review. This study holds that given the exponential pace of technological developments closing gap between humans and technology, and the idea that the law ought to evolve alongside economic norms,[108] Hallevy's models are a good starting point for the applicability of criminal law to AI entities, However, Hallevy's models cannot be applied in their current form, there is need for the review of the existing criminal law framework for the models to be applicable.

6. Conclusion

The three liability models as described above are not alternative models. None of the three models is mutually exclusive.[109] For instance, when the AI entity plays the role of an innocent agent in the perpetration of a specific offence, and the programmer, is the only person who directed that perpetration, the application of the perpetration-via- another model (the first liability model) is the most appropriate legal model for that situation.[110] In that same situation, when the programmer is itself an entity(when an AI entity programs another AI entity to commit a specific offence), the direct liability model (the third liability model) will be applied in addition to the first liability model, and not in lieu thereof.[111]Thus in such situations, the AI entity programmer shall be criminally liable, pursuant to a combination of the perpetration-via-another liability model and the direct liability model.[112] Likewise, if the AI entity plays the role of the physical perpetrator of the specific offence, but that very offence was not planned, then the application of the natural-probable-consequence liability model might be appropriate, the programmer might be deemed negligent if no offence had been perpetrated intentionally, or the programmer might be held accountable for that specific offence if another offence had been deliberately planned, but the specific offence that was perpetrated was not part of the original criminal scheme.[113] Nevertheless, if the programmer is not human, the direct liability model must be applied in addition to the simultaneous application of the natural-probable-consequence liability model; likewise when the physical perpetrator is human and the planner is an AI entity.[114]

The interaction of all the three liability models reveals a new legal situation in the specific context of AI entities and criminal law. As a result, when AI entities and humans are involved, directly or indirectly in the perpetration of a specific offence, it will be far more difficult to evade criminal liability. If the clearest purpose of the imposition of criminal liability is the application of legal social control in the specific society, then the coordinated application of all three models is necessary in the context of AI entities.[115] Provided that AI entities have self-awareness, self-consciousness, and free will, their criminal responsibility are present, and since the AI entities could embody social and ethical cores, as they are human creations, either directly or indirectly. Hence, this paper posits that there are sufficient dogmatic, juridical and technological apparatus to enable AI entities qualify as active legal actors in criminal justice.[116]

[105] A M Laurent, *Understanding Open Source &Free Software Licensing (Sebastopol:* O'Reilly Media, Inc (CA) 2004) http://www.oreilly.com accessed October 11, 2021

[106] Ibid

[107] C Metz, 'Inside Open AI, Elon Musk's Wild Plan to Set Artificial Intelligence Free+' *Wired,* April 27, 2016 http://www.wired.com/2016/04/openai-elon-musk-sam-altm-set-artificial-intelligence-free accessed October 11 2021

[108] R Charney, 'Can Androids Plead Automatism? A Review of *When Robots Kill: Artificial Intelligence Under the Criminal Law* by G.Hallevy' (2015)73 (69) *U T Fac L Rev* 70

[109] Ibid p193

[110] Ibid

[111] Ibid

[112] Ibid

[113] Hallevy n13 p194

[114] Ibid

[115] Ibid

[116] P M Freitas, F Andrade and P Novais, 'Criminal Liability of Autonomous Agents: from the Unthinkable to the Plausible' (2014) 8929 *Springer Science and Business Media LCC.* Law School, Universidade do Minho, Braga, Portugal. <pfreitas@direito.uminho.pt>accessed 10th October 2021.