The Criminal Liability of Artificial Intelligence Entities

Gabriel Hallevy

I. INTRODUCTION

In 1981, a 37-year-old Japanese employee of a motorcycle factory was killed by an artificial-intelligence robot working near him.[1] The robot erroneously identified the employee as a threat to its mission, and calculated that the most efficient way to eliminate this threat was by pushing him into an adjacent operating machine. Using its very powerful hydraulic arm, the robot smashed the surprised worker into the operating machine, killing him instantly, and then resumed its duties with no one to interfere with its mission. Unfortunately, this is not science fiction, and the legal question is: Who is to be held liable for this killing?

The technological world is changing rapidly. More and more simple human activities are being replaced by robots and computers. As long as humanity used computers as mere tools, there was no real difference between computers and screwdrivers, cars or telephones. When computers became sophisticated, we used to say that computers "think" for us. The problem began when computers evolved from "thinking" machines (machines that were programmed to perform defined thought processes/computing) into thinking machines (without quotation marks) – or Artificial

[1] The facts above are based on the overview in Yueh-Hsuan Weng, Chien-Hsun Chen and Chuen-Tsai Sun, *Towards the Human-Robot Co-Existence Society: On Safety Intelligence for Next Generation Robots*, 1 INT. J. SOC. ROBOT 267, 273 (2009).

Intelligence (AI). Artificial Intelligence research began in the early 1950s.[2] Since then, AI entities have become an integral part of modern human life, functioning much more sophisticatedly than other daily tools. Could they become dangerous?

In fact, they already are, as the above incident attests. In 1950, Isaac Asimov set down three fundamental laws of robotics in his science fiction masterpiece "I, Robot":1. A robot may not injure a human being or, through inaction, allow a human being to come to harm; 2. A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law. 3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Laws.[3]

These three fundamental laws are obviously contradictory.[4] What if a man orders a robot to hurt another person for the own good of the other person? What if the robot is in police service and the commander of the mission orders it to arrest a suspect and the suspect resists arrest? Or what if the robot is in medical service and is ordered to perform a surgical procedure on a patient, the patient objects, but the medical doctor insists that the procedure is for the patient's own good, and repeats the order to the robot? Besides, Asimov's fundamental laws of robotics relate only to

[2] N. P. PADHY, ARTIFICIAL INTELLIGENCE AND INTELLIGENT SYSTEMS 3-29 (2005, 2009).

[3] ISAAC ASIMOV, I, ROBOT (1950).

[4] Isaac Asimov himself wrote in his introduction to THE REST OF ROBOTS (1964) that "[t]here was just enough ambiguity in the Three Laws to provide the conflicts and uncertainties required for new stories, and, to my great relief, it seemed always to be possible to think up a new angle out of the 61 words of the Three Laws".

robots. AI software not installed in a robot would not be subject to Asimov's laws, even if these laws had any real legal significance.

The main question in that context is which kind of laws or ethics are correct and who is to decide. In order to cope with these same problems as they relate to humans, society devised criminal law. Criminal law embodies the most powerful legal social control in modern civilization. People's fear of AI entities, in most cases, is based on the fact that AI entities are not considered to be subject to the law, specifically to criminal law.[5] In the past, people were similarly fearful of corporations and their power to commit a spectrum of crimes, but since corporations are legal entities subject to criminal and corporate law, that kind of fear has been reduced significantly.[6]

---

[5] The apprehension that AI entities evoke may have arisen due to Hollywood's depiction of AI entities in numerous films, such as "2001: A Space Odyssey" (1968) [STANLEY KUBRICK, 2001: A SPACE ODYSSEY (1968)], and the modern trilogy "The Matrix" (1999, 2003, 2003) [JOEL SILVER, THE MATRIX (1999); JOEL SILVER, LAURENCE WACHOWSKI AND ANDREW PAUL WACHOWSKI, THE MATRIX RELOADED (2003); JOEL SILVER, LAURENCE WACHOWSKI AND ANDREW PAUL WACHOWSKI, THE MATRIX REVOLUTIONS (2003)], in which AI entities are not subject to the law. However, it should be noted that Hollywood did treat AI entities in an empathic way as well, by depicting them as human, as almost human, or as wishing to be human. See e.g. STEVEN SPIELBERG, STANLEY KUBRICK, JAN HARLAN, KATHLEEN KENNEDY, WALTER F. PARKES AND BONNIE CURTIS, A.I. ARTIFICIAL INTELLIGENCE (2001). This kind of treatment included, of course, clear subordination to human legal social control, and to criminal law.

[6] John C. Coffee, Jr., *"No Soul to Damn: No Body to Kick": An Unscandalised Inquiry Into the Problem of Corporate Punishment*, 79 MICH. L. REV. 386 (1981); STEVEN BOX, POWER,

Therefore, the modern question relating to AI entities becomes: Does the growing intelligence of AI entities subject them to legal social control, as any other legal entity?[7] This article attempts to work out a legal solution to the problem of the criminal liability of AI entities. At the outset, a definition of an AI entity will be presented. Based on that definition, this article will then propose and introduce three models of AI entity criminal liability:[8] the perpetration-by-another liability model, the natural-probable-consequence liability model and the direct liability model.

These three models might be applied separately, but in many situations, a coordinated combination of them (all or some of them) is required in order to

---

CRIME AND MYSTIFICATION 16-79 (1983);Brent Fisse and John Braithwaite, *The Allocation of Responsibility for Corporate Crime: Individualism, Collectivism and Accountability*, 11 SYDNEY L. REV. 468 (1988).

[7] See in general, but not in relation to criminal law, e.g. Thorne L. McCarty, *Reflections on Taxman: An Experiment in Artificial Intelligence and Legal Reasoning*, 90 HARV. L. REV. 837 (1977); Donald E. Elliott, *Holmes and Evolution: Legal Process as Artificial Intelligence* 13 J. LEGAL STUD. 113 (1984); Thomas E. Headrick and Bruce G. Buchanan, *Some Speculation about Artificial Intelligence and Legal Reasoning*, 23 STAN. L. REV. 40 (1971); Antonio A. Martino, *Artificial Intelligence and Law*, 2 INT'L J. L. & INFO. TECH. 154 (1994); Edwina L. Rissland, *Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning*, 99 YALE L. J. 1957 (1990).

[8] The Perpetration-by-Another Liability Model is discussed hereinafter at subparagraph III.B; The Natural Probable Consequence Liability Model is discussed hereinafter at subparagraph III.C; and the Direct Liability Model is discussed hereinafter at subparagraph III.D.

complete the legal structure of criminal liability.[9] Once we examine the possibility of legally imposing criminal liability on AI entities, then the question of punishment must be addressed. How can an AI entity serve a sentence of imprisonment? How can death penalty be imposed on an AI entity? How can probation, a pecuniary fine, etc. be imposed on an AI entity? Consequently, it is necessary to formulate viable forms of punishment in order to impose criminal liability practically on AI entities.[10]

II. WHAT IS AN ARTIFICIAL INTELLIGENCE ENTITY?

For some years, there has been significant controversy about the very essence of an AI entity.[11] Futurologists have proclaimed the birth of a new species, *machina sapiens*, which will share the human place as intelligent creatures on earth. Critics have argued that a "thinking machine" is an oxymoron. Machines, including computers, with their foundations of cold logic, can never be insightful or creative as humans are. This controversy raises the basic questions of the essence of humanity

[9] The coordination of the three liability models is discussed hereinafter at subparagraph III.E.

[10] The general punishment adjustment considerations are discussed hereinafter at paragraph IV.

[11] See e.g., Terry Winograd, *Thinking Machines: Can There Be? Are We?*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 167 (Derek Partridge and Yorick Wilks eds., 2006).

(Do human beings function as thinking machines?) and of AI (Can there be thinking machines?).[12]

There are five attributes that one would expect an intelligent entity to have.[13] The first is communication. One can communicate with an intelligent entity. The easier it is to communicate with an entity, the more intelligent the entity seems. One can communicate with a dog, but not about Einstein's theory of relativity. One can communicate with a little child about Einstein's theory, but it requires a discussion in terms that a child can comprehend. The second is internal knowledge. An intelligent entity is expected to have some knowledge about itself.

The third is external knowledge. An intelligent entity is expected to know about the outside world, to learn about it, and utilize that information. The fourth is goal-driven behavior.[14] An intelligent entity is expected to take action in order to achieve its goals. The fifth is creativity. An intelligent entity is expected to have some degree of creativity. In this context, creativity means the ability to take alternate action when the initial action fails. A fly that tries to exit a room and bumps into a window pane, tries to do that over and over again. When an AI robot bumps into a

---

[12] For the formal foundations of AI see e.g. Teodor C. Przymusinski, *Non-Monotonic Reasoning versus Logic Programming: A New Perspective*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 49 (Derek Partridge and Yorick Wilks eds., 2006); Richard W. Weyhrauch, *Prolegomena to a Theory of Mechanized formal Reasoning*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 72 (Derek Partridge and Yorick Wilks eds., 2006).

[13] Roger C. Schank, *Waht is AI, Anyway?*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 3 (Derek Partridge and Yorick Wilks eds., 2006).

[14] HANS WELZEL, DAS DEUTSCHE STRAFRECHT – EINE SYSTEMATISCHE DARSTELLUNG (11 Auf., 1969)

window, it tries to exit using the door. Most AI entities possess these five attributes by definition.[15] Some twenty first century types of AI entities possess even more attributes that enable them to act in far more sophisticated ways.[16]

An AI entity has a wide variety of applications, including in robots. A robot can be designed to imitate the physical capabilities of a human being, and these capabilities can be improved. A robot is capable of being physically faster and stronger than a human being. The AI software installed in it also enables the robot to calculate many complicated calculations faster and simultaneously, or to "think" faster. An AI entity is capable of learning and of gaining experience, and experience is a useful way of learning. All these attributes create the essence of an AI entity.[17]

---

[15] Schank, *supra* note 13, at pp. 4-6.

[16] In November 2009, during the Supercomputing Conference in Portland Oregon (SC 09), IBM scientists and others announced that they succeeded in creating a new algorithm named "Blue Matter," which possesses the thinking capabilities of a cat, Chris Capps, *"Thinking" Supercomputer Now Conscious as a Cat*, 11.19.2009, http://www.unexplainable.net/artman/publish/article_14423.shtml; last visited Jan 10, 2010; http://sc09.supercomputing.org/; last visited Jan 10, 2010. This algorithm collects information from very many units with parallel and distributed connections. The information is integrated and creates a full image of sensory information, perception, dynamic action and reaction, and cognition. This platform simulates brain capabilities, and eventually, it is supposed to simulate real thought processes. The final application of this algorithm contains not only analog and digital circuits, metal or plastics, but also protein-based biologic surfaces.

[17] See more e.g. Yorick Wilks, *One Small Head: Models and Theories*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 121 (Derek Partridge and Yorick Wilks eds., 2006); Alan Bundy and Stellan Ohlsson, *The Nature of AI Principles*, THE FOUNDATIONS OF ARTIFICIAL

AI robots and AI software are used in a wide range of applications in industry, military services, medical services, science, and even in games.[18]

III. MODELS OF THE CRIMINAL LIABILITY OF ARTIFICIAL INTELLIGENCE ENTITIES

A. General Requirements

The basic question of criminal law is the question of criminal liability; i.e., whether the specific entity (human or corporation) bears criminal liability for a specific offense committed at a specific point in time and space. In order to impose criminal liability upon a person, two main elements must exist. The first is the external or factual element; i.e., criminal conduct (*actus reus*), while the other is the internal or mental element; i.e., knowledge or general intent vis-à-vis the conduct element (*mens rea*). If one of them is missing, no criminal liability can be imposed.

INTELLIGENCE 135 (Derek Partridge and Yorick Wilks eds., 2006); Thomas W. Simon, *Artificial Methodology Meets Philosophy*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 155 (Derek Partridge and Yorick Wilks eds., 2006).

[18] See e.g. William B. Schwartz, Ramesh S. Patil and Peter Szolovits, *Artificial Intelligence in Medicine Where Do We Stand*, 27 JURIMETRICS J. 362 (1987); Richard E. Susskind, *Artificial Intelligence, Expert Systems and the Law*, 5 DENNING L. J. 105 (1990).

The *actus reus* requirement is expressed mainly by acts or omissions.[19] Sometimes, other external elements are required in addition to conduct, such as the specific results of that conduct and the specific circumstances underlying the conduct.[20] The *mens rea* requirement has various levels of mental elements. The highest level is expressed by knowledge, while sometimes it is accompanied by a requirement of intent or specific intention.[21] Lower levels are expressed by negligence[22] (a reasonable person should have known), or by strict liability offenses.[23]

---

[19] Walter Harrison Hitchler, *The Physical Element of Crime*, 39 DICK. L. REV. 95 (1934); MICHAEL MOORE, ACT AND CRIME: THE PHILOSOPHY OF ACTION AND ITS IMPLICATIONS FOR CRIMINAL LAW (1993).

[20] JOHN WILLIAM SALMOND, ON JURISPRUDENCE 505 (Glanville Williams ed., 11th ed., 1957); GLANVILLE WILLIAMS, CRIMINAL LAW: THE GENERAL PART §11 (2nd ed., 1961); OLIVER W. HOLMES, THE COMMON LAW 54 (1881, 1923); Walter Wheeler Cook, *Act, Intention, and Motive in Criminal Law*, 26 YALE L. J. 645 (1917).

[21] J. Ll. J. Edwards, *The Criminal Degrees of Knowledge*, 17 MOD. L. REV. 294 (1954); Rollin M. Perkins, *"Knowledge" as a Mens Rea Requirement*, 29 HASTINGS L. J. 953 (1978); United States v. Youts, 229 F.3d 1312 (10th Cir.2000); United States v. Spinney, 65 F.3d 231 (1st Cir.1995); State v. Sargent, 156 Vt. 463, 594 A.2d 401 (1991); State v. Wyatt, 198 W.Va. 530, 482 S.E.2d 147 (1996); People v. Steinberg, 79 N.Y.2d 673, 584 N.Y.S.2d 770, 595 N.E.2d 845 (1992).

[22] Jerome Hall, *Negligent Behaviour Should Be Excluded from Penal Liability*, 63 COLUM. L. REV. 632 (1963); Robert P. Fine and Gary M. Cohen, *Is Criminal Negligence a Defensible Basis for Criminal Liability?*, 16 BUFF. L. REV. 749 (1966).

[23] Jeremy Horder, *Strict Liability, Statutory Construction and the Spirit of Liberty*, 118 LAW Q. REV. 458 (2002); Francis Bowes Sayre, *Public Welfare Offenses*, 33 COLUM. L. REV. 55 (1933); Stuart P. Green, *Six Senses of Strict Liability: A Plea for Formalism*, APPRAISING

No other criteria or capabilities are required in order to impose criminal liability, not from humans, nor from any other kind of entity, including corporations and AI entities. An entity might possess further capabilities, however, in order to impose criminal liability, the existence of *actus reus* and *mens rea* in the specific offense is quite enough. A spider is capable of acting, but it is incapable of formulating the *mens rea* requirement; therefore, a spider bite bears no criminal liability. A parrot is capable of repeating words it hears, but it is incapable of formulating the *mens rea* requirement for libel.

In order to impose criminal liability on any kind of entity, it must be proven that the above two elements existed. When it has been proven that a person committed the criminal act knowingly or with criminal intent, that person is held criminally liable for that offense. The relevant question concerning the criminal liability of AI entities is: How can these entities fulfill the two requirements of criminal liability? This paper proposes the imposition of criminal liability on AI entities using three possible models of liability: the Perpetration-by-Another liability model; the Natural-Probable-Consequence liability model; and the Direct liability model. Following is an explanation of these three possible models.

B. The Perpetration-by-Another Liability Model

STRICT LIABILITY 1 (A. P. Simester ed., 2005); A. P. Simester, *Is Strict Liability Always Wrong?*, APPRAISING STRICT LIABILITY 21 (A. P. Simester ed., 2005).

This first model does not consider the AI entity as possessing any human attributes. The AI entity is considered an innocent agent. Accordingly, due to that legal viewpoint, a machine is a machine, and is never human. However, one cannot ignore an AI entity's capabilities, as mentioned above. Pursuant to this model, these capabilities are insufficient to deem the AI entity a perpetrator of an offense. These capabilities resemble the parallel capabilities of a mentally limited person, such as a child, or of a person who is mentally incompetent or who lacks a criminal state of mind.

Legally, when an offense is committed by an innocent agent, like when a person causes a child,[24] a person who is mentally incompetent[25] or who lacks a criminal state of mind, to commit an offense,[26] that person is criminally liable as a perpetrator-via-another.[27] In such cases, the intermediary is regarded as a mere instrument, albeit a sophisticated instrument, while the party orchestrating the offense (the perpetrator-via-another) is the real perpetrator as a principal in the first degree

---

[24] Maxey v. United States, 30 App. D.C. 63 (App.D.C.1907); Commonwealth v. Hill, 11 Mass. 136 (1814); Michael, (1840) 2 Mood. 120, 169 E.R. 48.

[25] Johnson v. State, 142 Ala. 70, 38 So. 182 (1904); People v. Monks, 133 Cal. App. 440, 24 P.2d 508 (Cal.App.4Dist.1933).

[26] United States v. Bryan, 483 F.2d 88 (3rd Cir.1973); Boushea v. United States, 173 F.2d 131 (8th Cir.1949); People v. Mutchler, 309 Ill. 207, 140 N.E. 820 (1923); State v. Runkles, 326 Md. 384, 605 A.2d 111 (1992); Parnell v. State, 323 Ark. 34, 912 S.W.2d 422 (Ark.1996); State v. Thomas, 619 S.W.2d 513 (Tenn.1981).

[27] Morrisey v. State, 620 A.2d 207 (Del.1993); Conyers v. State, 367 Md. 571, 790 A.2d 15 (2002); State v. Fuller, 346 S.C. 477, 552 S.E.2d 282 (2001); Gallimore v. Commonwealth, 246 Va. 441, 436 S.E.2d 421 (1993).

and is held accountable for the conduct of the innocent agent. The perpetrator's liability is determined on the basis of that conduct[28] and his own mental state.[29]

The derivative question relative to artificial intelligence entities is: Who is the perpetrator-via-another? There are two candidates: the first is the programmer of the AI software and the second is the user, or the end-user. A programmer of AI software might design a program in order to commit offenses via the AI entity. For example: a programmer designs software for an operating robot. The robot is intended to be 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 the perpetrator.

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 specific user is identified by the robot as that master, and the master orders the robot to assault any invader of the house. The robot executes the order exactly as ordered. This is not different than a person who orders his dog to attack any trespasser. The robot committed the assault, but the user is deemed the perpetrator.

---

[28] Dusenbery v. Commonwealth, 220 Va. 770, 263 S.E.2d 392 (1980).

[29] United States v. Tobon-Builes, 706 F.2d 1092 (11th Cir.1983); United States v. Ruffin, 613 F.2d 408 (2nd Cir.1979).

In both scenarios, the actual offense was committed by the AI entity. The programmer or the user did not perform any action conforming to the definition of a specific offense; therefore, they do not meet the *actus reus* requirement of the specific offense. The Perpetration-by-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 that is the instrumental usage of the AI entity as an innocent agent. No mental attribute required for the imposition of criminal liability is attributed to the AI entity.[30]

When programmers or users use an AI entity instrumentally, the commission of an offense by the AI entity is attributed to them. The internal element required in the specific offense already exists in their minds. 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 offenses 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.

This liability model does not attribute any mental capability, or any human 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

[30] The AI entity is used as an instrument and not as a participant, although it uses its features of processing information. See e.g. George R. Cross and Cary G. Debessonet, *An Artificial Intelligence Application in the Law: CCLIPS, A Computer Program that Processes Legal Information*, 1 HIGH TECH. L. J. 329 (1986).

burglar's. This is the same legal situation when using an animal instrumentally. An assault committed by a dog by 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 offense 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. Still, it is usage of an AI entity, due to its ability to execute an order to commit an offense. A screwdriver cannot execute such an order; a dog can. A dog cannot execute complicated orders; an AI entity can.[31]

The Perpetration-by-Another liability model is not suitable when an AI entity decides to commit an offense based on its own accumulated experience or knowledge. This model is not suitable when the software of the AI entity was not designed to commit the specific offense, but was committed by the AI entity nonetheless. This model is also not suitable when the specific AI entity functions not as an innocent agent, but as a semi-innocent agent.[32]

---

[31] Compare Andrew J. Wu, *From Video Games to Artificial Intelligence: Assigning Copyright Ownership to Works Generated by Increasingly Sophisticated Computer Programs*, 25 AIPLA Q. J. 131 (1997); Timothy L. Butler, *Can a Computer be an Author – Copyright Aspects of Artificial Intelligence*, 4 COMM. ENT. L. S. 707 (1982).

[32] NICOLA LACEY AND CELIA WELLS, RECONSTRUCTING CRIMINAL LAW – CRITICAL PERSPECTIVES ON CRIME AND THE CRIMINAL PROCESS 53 (2nd ed., 1998).

However, the Perpetration-by-Another liability model might be suitable when a programmer or user makes instrumental usage of an AI entity, but without using the AI entity's advanced capabilities. The legal result of applying this model is that the programmer and the user are fully criminally liable for the specific offense committed, while the AI entity has no criminal liability whatsoever.

C. The Natural-Probable-Consequence Liability Model

The second model of criminal liability assumes deep involvement of the programmers or users in the AI entity's daily activities, but without any intention of committing any offense via the AI entity. One scenario: during the execution of its daily tasks, an AI entity commits an offense. The programmers or users had no knowledge of the offense until it had already been committed; they did not plan to commit any offense, and they did not participate in any part of the commission of that specific offense.

One example of such a scenario: 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 (which is the AI entity), and the program is 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. 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.

Obviously, the programmer had not intended to kill anyone, especially not the human pilot, but nonetheless, the human pilot was killed as a result of the AI entity's actions, and these actions were done according to the program. Another example is AI software designed to detect threats from the internet and protect a computer system from these threats. A few days after the software is activated, it figures out that the best way to detect such threats is by entering web sites it defines as dangerous and destroying any software recognized as a threat. When the software does that, it is committing a computer offense, although the programmer did not intend for the AI entity to do so.

In these examples, the first model is not legally suitable. The first model assumes *mens rea*, the criminal intent of the programmers or users to commit an offense via the instrumental use of some of the AI entity's capabilities. This is not the legal situation in these cases. In these cases, the programmers or users had no knowledge of the committed offense; they had not planned it, and had not intended to commit the offense using the AI entity. For such cases, the second model might create a suitable legal response. This model is based upon the ability of the programmers or users to foresee the potential commission of offenses.

According to the second model, a person might be held accountable for an offense, if that offense is a natural and probable consequence of that person's conduct. Originally, the natural-probable-consequence liability was used to impose criminal

liability upon accomplices, when one committed an offense, which had not been planned by all of them and which was not part of a conspiracy. The established rule prescribed by courts and commentators is that accomplice liability extends to acts of a perpetrator that were a "natural and probable consequence"[33] of a criminal scheme that the accomplice encouraged or aided.[34] The natural-probable-consequence liability has been widely accepted in accomplice liability statutes and recodifications.[35]

Natural-probable-consequence liability seems to be legally suitable for situations in which an AI entity committed an offense, while the programmer or user had no knowledge of it, had not intended it and had not participated in it. The natural-probable-consequence liability model requires the programmer or user to be in a mental state of negligence, not more. Programmers or users are not required to know about any forthcoming commission of an offense as a result of their activity, but are required to know that such an offense is a natural, probable consequence of their actions.

---

[33] United States v. Powell, 929 F.2d 724 (D.C.Cir.1991).

[34] WILLIAM M. CLARK AND WILLIAM L. MARSHALL, LAW OF CRIMES 529 (7th ed., 1967); Francis Bowes Sayre, *Criminal Responsibility for the Acts of Another*, 43 HARV. L. REV. 689 (1930); People v. Prettyman, 14 Cal.4th 248, 58 Cal.Rptr.2d 827, 926 P.2d 1013 (1996); Chance v. State, 685 A.2d 351 (Del.1996); Ingram v. United States, 592 A.2d 992 (D.C.App.1991); Richardson v. State, 697 N.E.2d 462 (Ind.1998); Mitchell v. State, 114 Nev. 1417, 971 P.2d 813 (1998); State v. Carrasco, 122 N.M. 554, 928 P.2d 939 (1996); State v. Jackson, 137 Wash.2d 712, 976 P.2d 1229 (1999).

[35] State v. Kaiser, 260 Kan. 235, 918 P.2d 629 (1996); United States v. Andrews, 75 F.3d 552 (9th Cir.1996).

A negligent person, in a criminal context, is a person who has no knowledge of the offense, but a reasonable person should have known about it, since the specific offense is a natural probable consequence of that person's conduct.[36] The programmers or users of an AI entity, who should have known about the probability of the forthcoming commission of the specific offense, are criminally liable for the specific offense, even though they did not actually know about it. This is the fundamental legal basis for criminal liability in negligence cases. Negligence is, in fact, an omission of awareness or knowledge. The negligent person omitted knowledge, not acts.

The natural-probable-consequence liability model would permit liability to be predicated upon negligence, even when the specific offense requires a different state of mind.[37] This is not valid in relation to the person who personally committed the offense, but rather, is considered valid in relation to the person who was not the actual perpetrator of the offense, but was one of its intellectual perpetrators. Reasonable programmers or users should have foreseen the offense, and prevented it from being committed by the AI entity.

---

[36] Robert P. Fine and Gary M. Cohen, *Is Criminal Negligence a Defensible Basis for Criminal Liability?*, 16 BUFF. L. REV. 749 (1966); Herbert L.A. Hart, *Negligence, Mens Rea and Criminal Responsibility*, OXFORD ESSAYS IN JURISPRUDENCE 29 (1961); Donald Stuart, *Mens Rea, Negligence and Attempts*, [1968] CRIM. L.R. 647 (1968).

[37] THE AMERICAN LAW INSTITUTE, MODEL PENAL CODE – OFFICIAL DRAFT AND EXPLANATORY NOTES 312 (1962, 1985) (hereinafter "Model Penal Code"); State v. Linscott, 520 A.2d 1067 (Me.1987).

However, the legal results of applying the natural-probable-consequence liability model to the programmer or user differ in two different types of 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 offense. The second type of case is when the programmers or users programmed or used the AI entity knowingly and willfully in order to commit one offense via the AI entity, but the AI entity deviated from the plan and committed some other offense, in addition to or instead of the planned offense.

The first type of case is a pure case of negligence. The programmers or users acted or omitted negligently; therefore, there is no reason why they should not be held accountable for an offense of negligence, if there is such an offense in the specific legal system. Thus, 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. Consequently, if there is a specific offense of negligent homicide in that legal system, this is the most severe offense, for which the programmer might be held accountable, and not manslaughter or murder, which requires knowledge or intent.

The second type of case resembles the basic idea of the natural probable consequence liability in accomplice liability cases. The dangerousness of the very association or conspiracy whose aim is to commit an offense is the legal reason for more severe accountability to be imposed upon the cohorts. For example, a programmer programs an AI entity to commit a violent robbery in a bank, but the programmer did not program the AI entity to kill anyone. During the execution of the

violent robbery, the AI entity kills one of the people present at the bank who resisted the robbery. In such cases, the criminal negligence liability alone is insufficient. The danger posed by such a situation far exceeds negligence.

As a result, according to the natural-probable-consequence liability model, when the programmers or users programmed or used the AI entity knowingly and willfully in order to commit one offense via the AI entity, but the AI entity deviated from the plan and committed another offense, in addition to or instead of the planned offense, the programmers or users shall be held accountable for the offense itself, as if it had been committed knowingly and willfully. In the above example of the robbery, the programmer shall be held criminally accountable for the robbery (if committed), as well as for the killing, as an offense of manslaughter or murder, which requires knowledge and intent.[38]

The question still remains: What is the criminal liability of the AI entity itself when the natural-probable-consequence liability model is applied? In fact, there are two possible outcomes. If the AI entity acted as an innocent agent, without knowing anything about the criminal prohibition, it is not held criminally accountable for the offense it committed. Under such circumstances, the actions of the AI entity were not

[38] Cunningham, [1957] 2 Q.B. 396, [1957] 2 All E.R. 412, [1957] 3 W.L.R. 76, 41 Cr. App. Rep. 155; Faulkner, (1876) 13 Cox C.C. 550; United States v. Greer, 467 F.2d 1064 (7th Cir.1972); People v. Cooper, 194 Ill.2d 419, 252 Ill.Dec. 458, 743 N.E.2d 32 (2000); People v. Weiss, 256 App.Div. 162, 9 N.Y.S.2d 1 (1939); People v. Little, 41 Cal.App.2d 797, 107 P.2d 634 (1941); People v. Cabaltero, 31 Cal.App.2d 52, 87 P.2d 364 (1939); People v. Michalow, 229 N.Y. 325, 128 N.E. 228 (1920).

different from the actions of the AI entity under the first model (the Perpetration-by-Another liability model).[39] However, 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 offense directly. The direct liability model of AI entities is the third model, as described hereunder.

D. The Direct Liability Model

The third model does not assume any dependence of the AI entity on a specific programmer or user. The third model focuses on the AI entity itself.[40] As discussed above, criminal liability for a specific offense is mainly comprised of the external element (*actus reus*) and the internal element (*mens rea*) of that offense.[41] Any person attributed with both elements of the specific offense is held criminally accountable for that specific offense. No other criteria are required in order to impose criminal liability. A person might possess further capabilities, but, in order to impose criminal

---

[39] See above at subparagraph III.B.

[40] Compare e.g. Steven J. Frank, *Tort Adjudication and the Emergence of Artificial Intelligence Software*, 21 SUFFOLK U. L. REV. 623 (1987); S. N. Lehmanqzig, *Frankenstein Unbound – Towards a Legal Definition of Artificial Intelligence*, 1981 FUTURES 442 (1981); Maruerite E. Gerstner, *Liability Issues with Artificial Intelligence Software*, 33 SANTA CLARA L. REV. 239 (1993); Richard E. Susskind, *Expert Systems in Law: A Jurisprudential Approach to Artificial Intelligence and Legal Reasoning*, 49 MOD. L. REV. 168 (1986).

[41] See above at subparagraph III.A.

liability, the existence of the external element and the internal element required to impose liability for the specific offense is quite enough.

In order to impose criminal liability on any kind of entity, the existence of these elements in the specific entity must be proven. When it has been proven that a person committed the offense in question with knowledge or intent, that person is held criminally liable for that offense. The relevant questions regarding the criminal liability of AI entities is: How can these entities fulfill the requirements of criminal liability? Do AI entities differ from humans in this context?

An AI algorithm might have very many features and qualifications far exceeding those of an average human, but such features or qualifications are not required in order to impose criminal liability. When a human or corporation fulfills the requirements of both the external element and the internal element, criminal liability is imposed. If an AI entity is capable of fulfilling the requirements of both the external element and the internal element, and, in fact, it actually fulfills them, there is nothing to prevent criminal liability from being imposed on that AI entity.

Generally, the fulfillment of the external element requirement of an offense is easily attributed to AI entities. As long as an AI entity controls a mechanical or other mechanism to move its moving parts, any act might be considered as performed by the AI entity. Thus, when an AI robot activates its electric or hydraulic arm and moves it, this might be considered an act, if the specific offense involves such an act. For example, in the specific offense of assault, such an electric or hydraulic

movement of an AI robot that hits a person standing nearby is considered as fulfilling the *actus reus* requirement of the offense of assault.

When an offense might be committed due to an omission, it is even simpler. Under this scenario, the AI entity is not required to act at all. Its very inaction is the legal basis for criminal liability, as long as there had been a duty to act. If a duty to act is imposed upon the AI entity, and it fails to act, the *actus reus* requirement of the specific offense is fulfilled by way of an omission.

The attribution of the internal element of offenses to AI entities is the real legal challenge in most cases. The attribution of the mental element differs from one AI technology to other. Most cognitive capabilities developed in modern AI technology are immaterial to the question of the imposition of criminal liability. Creativity is a human feature that some animals also have, but creativity is a not a requirement for imposing criminal liability. Even the most uncreative persons are held criminally liable. The sole mental requirements needed in order to impose criminal liability are knowledge, intent, negligence, etc., as required in the specific offense and under the general theory of criminal law.

Knowledge is defined as sensory reception of factual data and the understanding of that data.[42] Most AI systems are well equipped for such reception.

---

[42] WILLIAM JAMES, THE PRINCIPLES OF PSYCHOLOGY (1890); HERMANN VON HELMHOLTZ, THE FACTS OF PERCEPTION (1878); In this context knowledge and awareness are identical. See e.g. United States v. Youts, 229 F.3d 1312 (10th Cir.2000); State v. Sargent, 156 Vt. 463, 594 A.2d 401 (1991); United States v. Spinney, 65 F.3d 231 (1st Cir.1995); State v. Wyatt,

Sensory receptors of sights, voices, physical contact, touch, etc., are not rare in most AI systems. These receptors transfer the factual data received to central processing units that analyze the data. The process of analysis in AI systems parallels that of human understanding.[43] The human brain understands the data received by eyes, ears, hands, etc., by analyzing that data. Advanced AI algorithms are trying to imitate human cognitive processes.[44] These processes are not so different.[45]

---

198 W.Va. 530, 482 S.E.2d 147 (1996); United States v. Wert-Ruiz, 228 F.3d 250 (3rd Cir.2000); United States v. Jewell, 532 F.2d 697 (9th Cir.1976); United States v. Ladish Malting Co., 135 F.3d 484 (7th Cir.1998); The Model Penal Code, supra note 37, at subsection 2.02(2)(b) (p. 21) even provides that "A person acts **knowingly** with a respect to a material element of an offense when: (i) if…, he is **aware** that his conduct is of that nature or that such circumstances exist; and (ii) if…, he is **aware** that it is practically certain that his conduct will cause such a result" (emphasis not in original).

[43] Margaret A. Boden, *Has AI Helped Psychology?*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 108 (Derek Partridge and Yorick Wilks eds., 2006); Derek Partridge, *What's in an AI Program?*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 112 (Derek Partridge and Yorick Wilks eds., 2006); David Marr, *AI: A Personal View*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 97 (Derek Partridge and Yorick Wilks eds., 2006).

[44] See above at paragraph II.

[45] Daniel C. Dennett, *Evolution, Error, and Intentionality*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 190 (Derek Partridge and Yorick Wilks eds., 2006); B. Chandraswkaran, *What Kind of Information Processing is Intelligence?*, THE FOUNDATIONS OF ARTIFICIAL INTELLIGENCE 14 (Derek Partridge and Yorick Wilks eds., 2006).

Specific intent is the strongest of the internal element requirements.[46] Specific intent is the existence of a purpose or an aim that a factual event will occur. The specific intent required to establish liability for murder is a purpose or an aim that a certain person will die.[47] As a result of the existence of such intent, the perpetrator of the offense commits the offense; i.e., he performs the external element of the specific offense. This situation is not unique to humans. An AI entity might be programmed to have a purpose or an aim and to take actions in order to achieve that purpose. This is specific intent.

One might assert that humans have feelings that cannot be imitated by AI software, not even by the most advanced software. Such feelings are love, affection, hatred, jealousy, etc. That might be correct in relation to the technology of the beginning of the 21st century. However, such feelings are rarely required in specific offenses. Most specific offenses are satisfied by knowledge of the existence of the external element. Few offenses require specific intent in addition to knowledge. Almost all other offenses are satisfied by much less than that (negligence, recklessness, strict liability). Perhaps in a very few specific offenses that do require

---

[46] Robert Batey, *Judicial Exploration of Mens Rea Confusion at Common Law and Under the Model Penal Code*, 18 GA. ST. U. L. REV. 341 (2001); State v. Daniels, 236 La. 998, 109 So.2d 896 (1958); Carter v. United States, 530 U.S. 255, 120 S.Ct. 2159, 147 L.Ed.2d 203 (2000); United States v. Randolph, 93 F.3d 656 (9th Cir.1996); United States v. Torres, 977 F.2d 321 (7th Cir.1992); Frey v. United States, 708 So.2d 918 (Fla.1998); State v. Neuzil, 589 N.W.2d 708 (Iowa 1999); People v. Disimone, 251 Mich.App. 605, 650 N.W.2d 436 (2002); People v. Henry, 239 Mich.App. 140, 607 N.W.2d 767 (1999).

[47] For the Intent-to-Kill murder see in WAYNE R. LAFAVE, CRIMINAL LAW 733-734 (4th ed., 2003).

certain feelings (e.g., crimes of racism, hate[48]), criminal liability cannot be imposed upon AI entities, which have no such feelings, but in any other specific offense, it is not a barrier.

If a person fulfills the requirements of both the external element and the internal element of a specific offense, then the person is held criminally liable. Why should an AI entity that fulfills all elements of an offense be exempt from criminal liability? One might argue that some segments of human society are exempt from criminal liability even if both the external and internal elements have been established. Such segments of society are infants and the mentally ill.

A specific order in criminal law exempts infants from criminal liability.[49] The social rationale behind the infancy defense is to protect infants from the harmful

---

[48] See e.g. Elizabeth A. Boyd, Richard A. Berk and Karl M. Hammer, *"Motivated by Hatred or Prejudice": Categorization of Hate-Motivated Crimes in Two Police Divisions*, 30 LAW & SOC'Y REV. 819 (1996); Projects, *Crimes Motivated by Hatred: The Constitutionality and Impact of Hate Crimes Legislation in the United States*, 1 SYRACUSE J. LEGIS. & POL'Y 29 (1995).

[49] See e.g. MINN. STAT. §9913 (1927); MONT. REV. CODE §10729 (1935); N.Y. PENAL CODE §816 (1935); OKLA. STAT. §152 (1937); UTAH REV. STAT. 103-I-40 (1933); State v. George, 20 Del. 57, 54 A. 745 (1902); Heilman v. Commonwealth, 84 Ky. 457, 1 S.W. 731 (1886); State v. Aaron, 4 N.J.L. 269 (1818); McCormack v. State, 102 Ala. 156, 15 So. 438 (1894); Little v. State, 261 Ark. 859, 554 S.W.2d 312 (1977); Clay v. State, 143 Fla. 204, 196 So. 462 (1940); *In re* Devon T., 85 Md.App. 674, 584 A.2d 1287 (1991); State v. Dillon, 93 Idaho 698, 471 P.2d 553 (1970); State v. Jackson, 346 Mo. 474, 142 S.W.2d 45 (1940).

consequences of the criminal process and to handle them in other social frameworks.[50] Do such frameworks exist for AI entities? The original legal rationale behind the infancy defense was the fact that infants are as yet incapable of comprehending what was wrong in their conduct (*doli incapax*).[51] Later, children can be held criminally liable if the presumption of mental incapacity was refuted by proof that the child was able to distinguish between right and wrong.[52] Could that be

---

[50] Frederick J. Ludwig, *Rationale of Responsibility for Young Offenders*, 29 NEB. L. REV. 521 (1950); *In re* Tyvonne, 211 Conn. 151, 558 A.2d 661 (1989); Andrew Walkover, *The Infancy Defense in the New Juvenile Court*, 31 U.C.L.A. L. REV. 503 (1984); Keith Foren, *Casenote: In Re Tyvonne M. Revisited: The Criminal Infancy Defense in Connecticut*, 18 Q. L. REV. 733 (1999); Michael Tonry, *Rethinking Unthinkable Punishment Policies in America*, 46 U.C.L.A. L. REV. 1751 (1999); Andrew Ashworth, *Sentencing Young Offenders*, PRINCIPLED SENTENCING: READINGS ON THEORY AND POLICY 294 (Andrew von Hirsch, Andrew Ashworth and Julian Roberts eds., 3rd ed., 2009); Franklin E. Zimring, *Rationales for Distinctive Penal Policies for Youth Offenders*, PRINCIPLED SENTENCING: READINGS ON THEORY AND POLICY 316 (Andrew von Hirsch, Andrew Ashworth and Julian Roberts eds., 3rd ed., 2009); Andrew von Hirsch, *Reduced Penalties for Juveniles: The Normative Dimension*, PRINCIPLED SENTENCING: READINGS ON THEORY AND POLICY 323 (Andrew von Hirsch, Andrew Ashworth and Julian Roberts eds., 3rd ed., 2009).

[51] SIR EDWARD COKE, INSTITUTIONS OF THE LAWS OF ENGLAND – THIRD PART 4 (6th ed., 1681, 1817, 2001).

[52] MATTHEW HALE, HISTORIA PLACITORUM CORONAE 23, 26 (1736) [MATTHEW HALE, HISTORY OF THE PLEAS OF THE CROWN (1736)]; McCormack v. State, 102 Ala. 156, 15 So. 438 (1894); Little v. State, 261 Ark. 859, 554 S.W.2d 312 (1977); *In re* Devon T., 85 Md.App. 674, 584 A.2d 1287 (1991).

similarly applied to AI entities? Most AI algorithms are capable of analyzing permitted and forbidden.

The mentally ill are presumed to lack the fault element of the specific offense, due to their mental illness (*doli incapax*).[53] The mentally ill are unable to distinguish between right and wrong (cognitive capabilities)[54] and to control impulsive behavior.[55] When an AI algorithm functions properly, there is no reason for it not to use all of its capabilities to analyze the factual data received through its receptors. However, an interesting legal question would be whether a defense of insanity might be raised in relation to a malfunctioning AI algorithm, when its analytical capabilities become corrupted as a result of that malfunction.

---

[53] See e.g. Benjamin B. Sendor, *Crime as Communication: An Interpretive Theory of the Insanity Defense and the Mental Elements of Crime*, 74 GEO. L. J. 1371, 1380 (1986); Joseph H. Rodriguez, Laura M. LeWinn and Michael L. Perlin, *The Insanity Defense Under Siege: Legislative Assaults and Legal Rejoinders*, 14 RUTGERS L. J. 397, 406-407 (1983); Homer D. Crotty, *The History of Insanity as a Defence to Crime in English Common Law*, 12 CAL. L. REV. 105 (1924).

[54] See e.g. Edward de Grazia, *The Distinction of Being Mad*, 22 U. CHI. L. REV. 339 (1955); Warren P. Hill, *The Psychological Realism of Thurman Arnold*, 22 U. CHI. L. REV. 377 (1955); Manfred S. Guttmacher, *The Psychiatrist as an Expert Witness*, 22 U. CHI. L. REV. 325 (1955); Wilber G. Katz, *Law, Psychiatry, and Free Will*, 22 U. CHI. L. REV. 397 (1955); Jerome Hall, *Psychiatry and Criminal Responsibility*, 65 YALE L. J. 761 (1956).

[55] See e.g. John Barker Waite, *Irresistible Impulse and Criminal Liability*, 23 MICH. L. REV. 443, 454 (1925); Edward D. Hoedemaker, *"Irresistible Impulse" as a Defense in Criminal Law*, 23 WASH. L. REV. 1, 7 (1948).

When an AI entity establishes all elements of a specific offense, both external and internal, there is no reason to prevent imposition of criminal liability upon it for that offense. The criminal liability of an AI entity 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. Criminal liability is not to be divided, but rather, added. The criminal liability of the AI entity is imposed in addition to the criminal liability of the human programmer or user.

However, the criminal liability of an AI entity is not dependent upon the criminal liability of the programmer or user of that AI entity. As a result, if the specific AI entity was programmed or used by another AI entity, the criminal liability of the programmed or used AI entity is not influenced by that fact. The programmed or used AI entity shall be held criminally accountable for the specific offense pursuant to the direct liability model, unless it was an innocent agent. In addition, the programmer or user of the AI entity shall be held criminally accountable for that very offense pursuant to one of the three liability models, according to its specific role in the offense. The chain of criminal liability might continue, if more parties are involved, whether human or AI entities.

There is no reason to eliminate the criminal liability of an AI entity or of a human, which is based on complicity between them. An AI entity and a human might cooperate as joint perpetrators, as accessories and abettors, etc., and the relevant criminal liability might be imposed on them accordingly. Since the factual and mental capabilities of an AI entity are sufficient to impose criminal liability on it, if these capabilities satisfy the legal requirements of joint perpetrators, accessories and

abettors, etc., then the relevant criminal liability as joint perpetrators, accessories and abettors, etc., should be imposed, regardless of whether the offender is an AI entity or a human.

Not only positive factual and mental elements might be attributed to AI entities. All relevant negative fault elements are attributable to AI entities. Most of these elements are expressed by the general defenses in criminal law; e.g., self-defense, necessity, duress, intoxication, etc. For some of these defenses (justifications),[56] there is no material difference between humans and AI entities, since they relate to a specific situation (*in rem*), regardless of the identity of the offender. For example, an AI entity serving under the local police force is given an order to arrest a person illegally. If the order is not manifestly illegal, the executer of the order is not criminally liable.[57] In that case, there is no difference whether the executer is human or an AI entity.

---

[56] JOHN C. SMITH, JUSTIFICATION AND EXCUSE IN THE CRIMINAL LAW (1989); Anthony M. Dillof, *Unraveling Unknowing Justification*, 77 NOTRE DAME L. REV. 1547 (2002); Kent Greenawalt, *Distinguishing Justifications from Excuses*, 49 LAW & CONTEMP. PROBS. 89 (Summer 1986); Kent Greenawalt, *The Perplexing Borders of Justification and Excuse*, 84 COLUM. L. REV. 949 (1984); Thomas Morawetz, *Reconstructing the Criminal Defenses: The Significance of Justification*, 77 J. CRIM. L. & CRIMINOLOGY 277 (1986); Paul H. Robinson, *A Theory of Justification: Societal Harm as a Prerequisite for Criminal Liability*, 23 U.C.L.A. L. REV. 266 (1975); Paul H. Robinson, *Testing Competing Theories of Justification*, 76 N.C. L. REV. 1095 (1998).

[57] Michael A. Musmanno, *Are Subordinate Officials Penally Responsible for Obeying Superior Orders which Direct Commission of Crime?*, 67 DICK. L. REV. 221 (1963).

For other defenses (excuses and exempts)[58] some applications should be adjusted. For example, the intoxication defense is applied when the offender is under the physical influence of an intoxicating substance (e.g., alcohol, drugs, etc.). The influence of alcohol on an AI entity is minor, at most, but the influence of an electronic virus that is infecting the operating system of the AI entity might be considered parallel to the influence of intoxicating substances on humans. Some other factors might be considered as being parallel to insanity, loss of control, etc.

It might be summed up that the criminal liability of an AI entity according to the direct liability model is not different from the relevant criminal liability of a human. In some cases, some adjustments are necessary, but substantively, it is the very same criminal liability, which is based upon the same elements and examined in the same ways.

E. Coordination

The three liability models described above are not alternative models. These models might be applied in combination in order to create a full image of criminal liability in the specific context of AI entity involvement. None of the three models is mutually exclusive. Thus, applying the second model is possible as a single model for

---

[58] Peter Arenella, *Convicting the Morally Blameless: Reassessing the Relationship Between Legal and Moral Accountability*, 39 U.C.L.A. L. REV. 1511 (1992); Sanford H. Kadish, *Excusing Crime*, 75 CAL. L. REV. 257 (1987); Andrew E. Lelling, *A Psychological Critique of Character-Based Theories of Criminal Excuse*, 49 SYRAC. L. REV. 35 (1998).

the specific offense, and it is possible as one part of a combination of two of the legal models or of all three of them.

When the AI entity plays the role of an innocent agent in the perpetration of a specific offense, and the programmer is the only person who directed that perpetration, the application of the Perpetration-by-Another model (the first liability model) is the most appropriate legal model for that situation. In that same situation, when the programmer is itself an AI entity (when an AI entity programs another AI entity to commit a specific offense), the direct liability model (the third liability model) is most appropriate to be applied to the criminal liability of the programmer of the AI entity. The third liability model in that situation is applied in addition to the first liability model, and not in lieu thereof. Thus, in such situations, the AI entity programmer shall be criminally liable, pursuant to a combination of the Perpetration-by-Another liability model and the direct liability model.

If the AI entity plays the role of the physical perpetrator of the specific offense, but that very offense was not planned to be perpetrated, then the application of the natural-probable-consequence liability model might be appropriate. The programmer might be deemed negligent if no offense had been deliberately planned to be perpetrated, or the programmer might be held fully accountable for that specific offense if another offense had indeed been deliberately planned, but the specific offense that was perpetrated had not been part of the original criminal scheme. Nevertheless, when 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 while the planner is an AI entity.

The coordination of all three liability models creates an opaque net of criminal liability. The combined and coordinated application of these three 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 offense, it will be far more difficult to evade criminal liability. The social benefit to be derived from such a legal policy is of substantial value. All entities – human, legal or AI – become subject to criminal law. 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 very context of AI entities.

IV. PUNISHMENT CONSIDERATIONS

Let us assume an AI entity is criminally liable. Let us assume it is indicted, tried and convicted. After the conviction, the court is supposed to sentence that AI entity. If the most appropriate punishment under the specific circumstances is one year of imprisonment, for example, how can an AI entity practically serve such a sentence? How can death penalty, probation or even a fine be imposed on an AI entity? In instances where there is no body to arrest (especially in cases of AI software that was not installed in a physical body, such as a robot), what is the

practical meaning of imprisonment? Where no bank account is available for the sentenced AI entity, what is the practical significance of fining it?

Similar legal problems have been raised when the criminal liability of corporations was recognized.[59] Some asked how any of the legitimate penalties imposed upon humans could be applicable to corporations. The answer was simple and legally applicable. When a punishment can be imposed on a corporation as it is on humans, it is imposed without change. When the court adjudicates a fine, the corporation pays the fine in the same way that a human pays the fine and in the same way that a corporation pays its bills in a civil context. However, when punishment of a corporation cannot be carried out in the same way as with humans, an adjustment is required. Such is the legal situation vis-à-vis AI entities.

The punishment adjustment considerations examine the theoretical foundations of any applied punishment. These considerations are applied in a similar manner and are comprised of three stages. Each stage may be explained by a question, as described

[59] Gerard E. Lynch, *The Role of Criminal Law in Policing Corporate Misconduct*, 60 LAW & CONTEMP. PROBS. 23 (1997); Richard Gruner, *To Let the Punishment Fit the Organization: Sanctioning Corporate Offenders Through Corporate Probation*, 16 AM. J. CRIM. L. 1 (1988); Steven Walt and William S. Laufer, *Why Personhood Doesn't Matter: Corporate Criminal Liability and Sanctions*, 18 AM. J. CRIM. L. 263 (1991); John C. Coffee, Jr., *"No Soul to Damn: No Body to Kick": An Unscandalised Inquiry Into the Problem of Corporate Punishment*, 79 MICH. L. REV. 386 (1981); STEVEN BOX, POWER, CRIME AND MYSTIFICATION 16-79 (1983);Brent Fisse and John Braithwaite, *The Allocation of Responsibility for Corporate Crime: Individualism, Collectivism and Accountability*, 11 SYDNEY L. REV. 468 (1988).

below: (1) What is the fundamental significance of the specific punishment for a human? (2) How does that punishment affect AI entities? and – (3) What practical punishments may achieve the same significance when imposed on AI entities?

The most significant advantage of these punishment adjustment considerations is that the significance of the specific punishment remains identical when imposed on humans and AI entities. This method of punishment adjustment considerations is referred to below in some of the punishments used in modern societies: death penalty, imprisonment, suspended sentencing, community service and fines.

Death penalty is considered the most severe punishment for humans, and there is no consensus regarding its constitutionality among the various jurisdictions.[60] Death penalty is the most effective method of incapacitating offenders as it relates to recidivism, since once the death sentence is carried out, the offender is obviously incapable of committing any further offense. The significance of death penalty for humans is the deprivation of life.[61] The "life" of an AI entity is its independent

[60] See e.g. the abolition of death penalty in Germany in 1949, Grundgesetz, Art. 102; in Britain for murder in 1965, Murder (Abolition of Death Penalty) Act, 1965, c.71; and the debate in the United States, *In re* Kemmler, 136 U.S. 436, 10 S.Ct. 930, 34 L.Ed. 519 (1890); Provenzano v. Moore, 744 So.2d 413 (Fla. 1999); Dutton v. State, 123 Md. 373, 91 A. 417 (1914); Campbell v. Wood, 18 F.3d 662 (9th Cir. 1994); Wilkerson v. Utah, 99 U.S. (9 Otto) 130, 25 L.Ed. 345 (1878); People v. Daugherty, 40 Cal.2d 876, 256 P.2d 911 (1953); Gray v. Lucas, 710 F.2d 1048 (5th Cir. 1983); Hunt v. Nuth, 57 F.3d 1327 (4th Cir. 1995); Gregg v. Georgia, 428 U.S. 153, S.Ct. 2909, 49 L.Ed.2d 859 (1979).

[61] ROBERT M. BOHM, DEATHQUEST: AN INTRODUCTION TO THE THEORY AND PRACTICE OF DEATH PENALTY IN THE UNITED STATES 74-78 (1999); Austin Sarat, *The Cultural Life of*

existence as an entity. Sometimes, it has a physical appearance (e.g., as a robot); sometimes it has only an abstract existence (e.g., as software installed on a computer system or on a network server). Considering death penalty's efficacy in incapacitating offenders, the practical action that may achieve the same results as death penalty when imposed on an AI entity is deletion of the AI software controlling the AI entity. Once the deletion sentence is carried out, the offending AI entity is incapable of committing any further offenses. The deletion eradicates the independent existence of the AI entity and is tantamount to the death penalty.

Imprisonment is one of the most popular sentences imposed in western legal systems for serious crimes. The significance of imprisonment for humans is the deprivation of human liberty and the imposition of severe limitations on human free behavior, freedom of movement and freedom to manage one's personal life.[62] The

---

Death penalty: Responsibility and Representation in Dead Man Walking *and* Last Dance, THE KILLING STATE – DEATH PENALTY IN LAW, POLITICS, AND CULTURE 226 (Austin Sarat ed., 1999); Peter Fitzpatrick, *"Always More to Do": Death penalty and the (De)Composition of Law*, THE KILLING STATE – DEATH PENALTY IN LAW, POLITICS, AND CULTURE 117 (Austin Sarat ed., 1999); Franklin E. Zimring, *The Executioner's Dissonant Song: On Death penalty and American Legal Values*, THE KILLING STATE – DEATH PENALTY IN LAW, POLITICS, AND CULTURE 137 (Austin Sarat ed., 1999); Anthony G. Amsterdam, *Selling a Quick Fix for Boot Hill: The Myth of Justice Delayed in Death Cases*, THE KILLING STATE – DEATH PENALTY IN LAW, POLITICS, AND CULTURE 148 (Austin Sarat ed., 1999).

[62] David J. Rothman, *For the Good of All: The Progressive Tradition in Prison Reform*, HISTORY AND CRIME 271 (James A. Inciardi and Charles E. Faupel eds., 1980); MICHAEL WELCH, IRONIES OF IMPRISONMENT (2004); Roy D. King, *The Rise and Rise of Supermax: An American Solution in Search of a Problem?*, 1 PUNISHMENT AND SOCIETY 163 (1999);

"liberty" or "freedom" of an AI entity includes the freedom to act as an AI entity in the relevant area. For example, an AI entity in medical service has the freedom to participate in surgeries; an AI entity in a factory has the freedom to manufacture, etc. Considering the nature of a sentence of imprisonment, the practical action that may achieve the same effects as imprisonment when imposed on an AI entity is to put the AI entity out of use for a determinate period. During that period, no action relating to the AI entity's freedom is allowed, and thus its freedom or liberty is restricted.

Suspended sentencing is a very popular intermediate sanction in western legal systems for increasing the deterrent effect on offenders in lieu of actual imprisonment. The significance of a suspended sentence for humans is the very threat of imprisonment if the human commits a specific offense or a type of specific offense.[63]

CHASE RIVELAND, SUPERMAX PRISONS: OVERVIEW AND GENERAL CONSIDERATIONS (1999); JAMIE FELLNER AND JOANNE MARINER, COLD STORAGE: SUPER-MAXIMUM SECURITY CONFINEMENT IN INDIANA (1997); Richard Korn, *The Effects of Confinement in the High Security Unit in Lexington*, 15 SOCIAL JUSTICE 8 (1988); Holly A. Miller, *Reexamining Psychological Distress in the Current Conditions of Segregation*, 1 JOURNAL OF CORRECTIONAL HEALTH CARE 39 (1994); FRIEDA BERNSTEIN, THE PERCEPTION OF CHARACTERISTICS OF TOTAL INSTITUTIONS AND THEIR EFFECT ON SOCIALIZATION (1979); BRUNO BETTELHEIM, THE INFORMED HEART: AUTONOMY IN A MASS AGE (1960); Marek M. Kaminski, *Games Prisoners Play: Allocation of Social Roles in a Total Institution*, 15 Rationality and Society 188 (2003); JOHN IRWIN, PRISONS IN TURMOIL (1980); ANTHONY J. MANOCCHIO AND JIMMY DUNN, THE TIME GAME: TWO VIEWS OF A PRISON (1982).

[63] MARC ANCEL, SUSPENDED SENTENCE (1971); Marc Ancel, *The System of Conditional Sentence or Sursis*, 80 L. Q. REV. 334 (1964); Anthony E. Bottoms, *The Suspended Sentence in England 1967-1978*, 21 BRITISH JOURNAL OF CRIMINOLOGY 1 (1981).

If the human commits such an offense, a sentence of imprisonment will be imposed for the first offense in addition to the sentencing for the second offense. As a result, humans are deterred from committing another offense and from becoming a recidivist offender. Practically, a suspended sentence is imposed only in the legal records. No physical action is taken when a suspended sentence is imposed. As a result, when imposing a suspended sentence, there is no difference in effect between humans and AI entities. The statutory criminal records of the state do not differentiate between a suspended sentence imposed on humans, and those imposed on corporations or AI entities, as long as the relevant entity may be identified specifically and accurately.

Community service is also a very popular intermediate sanction in western legal systems in lieu of actual imprisonment. In most legal systems, community service is a substitute for short sentences of actual imprisonment. In some legal systems, community service is imposed coupled with probation so that the offender "pays a price" for the damages he caused by committing the specific offense.[64] The significance of community service for humans is compulsory contribution of labor to

---

[64] John Harding, *The Development of the Community Service*, ALTERNATIVE STRATEGIES FOR COPING WITH CRIME 164 (Norman Tutt ed., 1978); HOME OFFICE, REVIEW OF CRIMINAL JUSTICE POLICY (1977); Ashlee Willis, *Community Service as an Alternative to Imprisonment: A Cautionary View*, 24 PROBATION JOURNAL 120 (1977); Julie Leibrich, Burt Galaway and Yvonne Underhill, *Community Sentencing in New Zealand: A Survey of Users*, 50 FEDERAL PROBATION 55 (1986); James Austin and Barry Krisberg, *The Unmet Promise of Alternatives*, 28 JOURNAL OF RESEARCH IN CRIME AND DELINQUENCY 374 (1982); Mark S. Umbreit, *Community Service Sentencing: Jail Alternatives or Added Sanction?*, 45 FEDERAL PROBATION 3 (1981).

the community. As discussed above,[65] an AI entity can be engaged as a worker in very many areas. When an AI entity works in a factory, its work is done for the benefit of the factory owners or for the benefit of the other workers in order to ease and facilitate their professional tasks. In the same way that an AI entity works for the benefit of private individuals, it may work for the benefit of the community. When work for the benefit of the community is imposed on an AI entity as a compulsory contribution of labor to the community, it may be considered community service. Thus, the significance of community service is identical, whether imposed on humans or AI entities.

The adjudication of a fine is the most popular intermediate sanction in western legal systems in lieu of actual imprisonment. The significance of paying a fine for humans is deprivation of some of their property, whether the property is money (a fine) or other property (forfeiture).[66] When a person fails to pay a fine, or has insufficient property to pay the fine, substitute penalties are imposed on the offender,

---

[65] See above at paragraphs I, II.

[66] GERHARDT GREBING, THE FINE IN COMPARATIVE LAW: A SURVEY OF 21 COUNTRIES (1982); Judith A. Greene, *Structuring Criminal Fines: Making an 'Intermediate Penalty' More Useful and Equitable*, 13 JUSTICE SYSTEM JOURNAL 37 (1988); NIGEL WALKER AND NICOLA PADFIELD, SENTENCING: THEORY, LAW AND PRACTICE (1996); Manfred Zuleeg, *Criminal Sanctions to be Imposed on Individuals as Enforcement Instruments in European Competition Law*, EUROPEAN COMPETITION LAW ANNUAL 2001: EFFECTIVE PRIVATE ENFORCEMENT OF EC ANTITRUST LAW 451 (Claus-Dieter Ehlermann and Isabela Atanasiu eds., 2001); STEVE UGLOW, CRIMINAL JUSTICE (1995); DOUGLAS C. MCDONALD, JUDITH A. GREENE AND CHARLES WORZELLA, DAY-FINES IN AMERICAN COURTS: THE STATEN-ISLAND AND MILWAUKEE EXPERIMENTS (1992).

particularly imprisonment.[67] The imposition of a fine on a corporation is identical to the imposition of a fine on a person, since both people and corporations have property and bank accounts, so the payment of a fine is identical, whether the paying entity is human or a corporate entity.

However, most AI entities have no money or property of their own, nor have they any bank accounts. If an AI entity does have its own property or money, the imposition of a fine on it would be identical to the imposition of a fine on humans or corporations. For most humans and corporations, property is gained through labor.[68] When paying a fine, the property, which is a result of labor, is transferred to the state. That labor might be transferred to the state in the form of property or directly as labor. As a result, a fine imposed on an AI entity might be collected as money or property and as labor for the benefit of the community. When the fine is collected in the form of labor for the benefit of the community, it is not different from community service as described above. Thus, most common punishments are applicable to AI entities. The imposition of specific penalties on AI entities does not negate the nature of these penalties in comparison with their imposition on humans. Of course, some general punishment adjustment considerations are necessary in order to apply these penalties, but still, the nature of these penalties remains the same relative to humans and to AI entities.

---

[67] FIORI RINALDI, IMPRISONMENT FOR NON-PAYMENT OF FINES (1976); *Use of Short Sentences of Imprisonment by the Court*, REPORT OF THE SCOTTISH ADVISORY COUNCIL ON THE TREATMENT OF OFFENDERS (1960).

[68] JOHN LOCKE, TWO TREATISES OF GOVERNMENT (1689).

V. CONCLUSION

If all of its specific requirements are met, criminal liability may be imposed upon any entity – human, corporate or AI entity. Modern times warrant modern legal measures in order to resolve today's legal problems. The rapid development of Artificial Intelligence technology requires current legal solutions in order to protect society from possible dangers inherent in technologies not subject to the law, especially criminal law. Criminal law has a very important social function – that of preserving social order for the benefit and welfare of society. The threats upon that social order may be posed by humans, corporations or AI entities.

Traditionally, humans have been subject to criminal law, except when otherwise decided by international consensus. Thus, minors and mentally ill persons are not subject to criminal law in most legal systems around the world. Although corporations in their modern form have existed since the fourteenth century,[69] it took hundreds of years to subordinate corporations to the law, and especially, to criminal law. For hundreds of years, the law stated that corporations are not subject to criminal law, as inspired by Roman law (*societas delinquere non potest*).[70]

---

[69] WILLIAM SEARLE HOLDSWORTH, A HISTORY OF ENGLISH LAW 471-476 (1923).

[70] William Searle Holdsworth, *English Corporation Law in the 16th and 17th Centuries*, 31 YALE L. J. 382 (1922); WILLIAM ROBERT SCOTT, THE CONSTITUTION AND FINANCE OF ENGLISH, SCOTISH AND IRISH JOINT-STOCK COMPANIES TO 1720 462 (1912); BISHOP CARLETON HUNT, THE DEVELOPMENT OF THE BUSINESS CORPORATION IN ENGLAND 1800-1867 6 (1963).

It was only in the seventeenth century that an English court dared to impose criminal liability on a corporation.[71] It was inevitable. Corporations participate fully in human life, and it was outrageous not to subject them to human laws, since offenses are committed by corporations or through them. But corporations have neither body nor soul. Legal solutions were developed so that in relation to criminal liability, they would be deemed capable of fulfilling all requirements of criminal liability, including external elements and internal elements.[72] These solutions were embodied in models of criminal liability and general punishment adjustment considerations. It worked. In fact, it is still working, and very successfully.

Why should AI entities be different from corporations? AI entities are taking larger and larger parts in human activities, as do corporations. Offenses have already been committed by AI entities or through them. AI entities have no soul, and some AI entities have neither body nor soul. Thus, there is no substantive legal difference between the idea of criminal liability imposed on corporations and on AI entities. It would be outrageous not to subordinate them to human laws, as corporations have been. Models of criminal liability do exist and general paths to impose punishment. What else is needed?

---

[71] Langforth Bridge, (1635) Cro. Car. 365, 79 E.R. 919; See in addition Clifton (Inhabitants), (1794) 5 T.R. 498, 101 E.R. 280; Great Broughton (Inhabitants), (1771) 5 Burr. 2700, 98 E.R. 418; Stratford-upon-Avon Corporation, (1811) 14 East 348, 104 E.R. 636; Liverpool (Mayor), (1802) 3 East 82, 102 E.R. 529; Saintiff, (1705) 6 Mod. 255, 87 E.R. 1002.

[72] Frederick Pollock, *Has the Common Law Received the Fiction Theory of Corporations?*, 27 L. Q. REV. 219 (1911).