The Basic Models of Criminal Liability of AI Systems and Outer Circles

Gabriel Hallevy*

A. Introduction

The way humans cope with breaches of legal order is through criminal law, operated by the criminal justice system. Accordingly, human societies define criminal offenses and operate social mechanisms to apply them. This is how criminal law works. Originally, this way has been designed by humans and for humans. However, as technology has developed, criminal offenses are committed not only by humans. The major development in this issue has occurred in the 17th century. In the 21st century criminal law is required to supply adequate solutions for commission of criminal offenses through artificial intelligent (AI) entities. Basically, there are three fundamental models to cope with this phenomenon within the current definitions of criminal law. These models are presented hereinafter.

B. The Perpetration-by-Another Liability Model

The first model does not consider any human feature of the AI system. The AI system is considered to be an innocent agent. Accordingly, due to that legal point of view, a machine is a machine, and never human. However, one cannot ignore the capabilities of an AI system. Due to this model, these capabilities are not enough in order to consider the AI system as a

* Full Professor, Faculty of Law, Ono Academic College. This manuscript has been presented at FOI, the Swedish Defense Research Agency, at Stockholm, Sweden on May 6, 2019. I thank Erik Zouave for the invitation, and to the FOI researchers for their fruitful comments. The Models are based on previous researches published around the world, including two of the author's books: GABRIEL HALLEVY, WHEN ROBOTS KILL – ARTIFICIAL INTELLIGENCE UNDER CRIMINAL LAW, Northeastern University Press, University Press of New England (2013) and GABRIEL HALLEVY, LIABILITY FOR CRIMES INVOLVING ARTIFICIAL INTELLIGENCE SYSTEMS, Springer-Verlag International Academic Press (2015).

perpetrator of an offense. These capabilities might resemble the parallel capabilities of a mentally limited person, such as a child, a mentally incompetent or a person who lacks a criminal state of mind to engage the conduct.

Legally, when an offense is committed by an innocent agent, as where a person causes a child,[1] a mentally incompetent,[2] or a person who lacks a criminal state of mind to engage the conduct,[3] that person is criminally liable as a perpetrator-by-another.[4] In such cases the intermediary is regarded as a mere instrument, even though it is a sophisticated instrument, and the originating actor (the perpetrator-by-another) is the real perpetrator as a principal of the first degree. That perpetrator-by-another is liable for the conduct of the innocent agent, and the perpetrator liability is determined on the basis of that conduct[5] and the perpetrator-by-another own mental state.[6]

The derivative question is who the perpetrator-by-another is. There are two possible persons who are entitled to become the perpetrators-by-another. The first is the programmer of the AI software and the second is the user, or the edge user. The programmer of the AI software might design the program in order to commit offenses by the AI system. For example, the programmer designs a software of a robot operating. The robot is intended to be placed in a factory, and its software is designed to burn the factory at night while no person is there. The robot has committed the arson, but the programmer is legally considered to be the perpetrator.

The second person who might be considered to be the perpetrator-by-another is the user of the AI system. The user did not program the software, but he uses the AI system, including its software, for its own benefit. For example, the user purchases a servant-robot, which is designed to execute any order given by its master. The specific user is identified by the robot

---

[1] 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.

[2] 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).

[3] United States v. Bryan, 483 F.2d 88 (3rd Cir.1973); Boushea v. United States, 173 F.2d 131 (8th Cir.1949).

[4] Morrisey v. State, 620 A.2d 207 (Del.1993); Conyers v. State, 367 Md. 571, 790 A.2d 15 (2002).

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

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

as that master, and the master orders the robot to assault an invader to the house. The robot executes the order exactly as it has been ordered. This is not different from a master who orders his dog to attack a passing-by person. The robot has committed an assault, but the user is legally considered to be the perpetrator.

In both possible scenarios, the physical commission of the specific offense has been done by the AI system. The programmer or the user did not commit any conduct which is defined in the specific offense definition; therefore no external element of the specific offense has been existed in them. The perpetration-by-another liability model considers the conduct committed by the AI system as if it is the programmer's or the user's. The legal basis for that is the instrumental usage of the AI system as an innocent agent. No mental feature required for the imposition of criminal liability is attributed to the AI system.[7]

When the programmers or the users use the AI system instrumentally, the commission of the conduct by the AI system is attributed to them. The internal element required in the specific offense already exists in their mind. The programmer has the intention towards the commission of the arson, and the user intends to the commission of the assault, even though the factual commission of these offenses is committed through a robot, which is an AI system. The instrumental usage of an innocent agent is considered committed the perpetration of the user himself.

This liability model does not attribute any mental capability, or any human mental capability, to the AI system. According to this model, there is no legal difference between an AI system and a screwdriver or an animal having no capacity of making decisions. 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 "action" is, in fact, the burglar's. This is the same legal situation when using an animal instrumentally. An assault committed by a dog due to its master's order is, in fact, an assault committed by the master.

[7] The AI system 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).

That kind of a legal model might be suitable for two types of situations. The first situation is a usage of an AI system to commit an offense with no usage of its advanced capabilities. The second situation is a usage of an older version of AI system, which lacks the modern advanced capabilities of the advanced AI systems. In both situations, the usage of the AI system is instrumental. Still it is a usage of an AI system, due to the capabilities 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 system can.[8]

The perpetration-by-another liability model is not suitable when the AI system has decided to commit an offense due to the experience or knowledge it has gained by itself. This model is neither suitable when the software of the AI system was not designed in order to commit the specific offense, but still it has been committed by the AI system. Even when the specific AI system functions not as an innocent agent, but as a semi-innocent agent, this model is not suitable.[9]

However, the perpetration-by-another liability model might be suitable when an instrumental usage of the AI system was done by the programmer or by the user, and there was no usage of the AI system advanced capabilities. The legal result of applying this model is that the programmers or the users are fully criminally liable for the specific offense committed, and the AI system has no criminal liability at all.

C. The Natural Probable Consequence Liability Model

The second model of criminal liability assumes a deep involvement of the programmers or the users in the AI system's daily activities, but they did not plan to commit any offense through the AI system. However, during the execution of its daily missions, the AI system commits an offense. The programmers or the users did not know about the commission of the offense until

[8] 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).

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

it has already been done, they did not plan any commission of any offense, and they did not participated in any part of the commission of that specific offense.

Such a situation might be demonstrated by a design of an AI robot or software which is developed to function as an automatic pilot. The AI system 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 system, and the program is initialized. When the automatic pilot is activated, the human pilot sees a storm coming closer to the plane and tries to abort the mission and return back. The AI system understands the human pilot acts as a threat upon the mission and acts in order to eliminate that threat. It might cut off the air supply to the pilot or activate the ejection seat etc. Consequently, the human pilot is killed by the AI system acts.

Of course, the programmer did not intend to kill anyone, especially not the human pilot, but still the human pilot is killed out of the AI system acts, and these acts were done according to the program. Another example is an AI software, which is 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 into websites which it define as dangerous and destroy any software which is recognized as a threat. When the software does that, it commits a computer offense, although the programmer did not intend the AI system would do so.

In these cases, the first model is not legally suitable. The first model assumes a plan of the programmers or the users to commit an offense through the AI system, using some of its capabilities instrumentally. This is not the legal situation in these cases. In these cases the programmers or the users did not know about the commission of the offense, they did not plan it, and they did not intend to do that using the AI system. For such cases the second model might form a suitable legal response. This model is based upon the ability of the programmers or the users to foresee the forthcoming commission of the offense.

According to the second model, a person might be held liable for an offense, since that offense is a natural and probable consequence of that person conduct. Originally, the natural probable consequence liability is used to impose criminal liability upon accomplices, that one of them has committed an offense which was not planned by them all and which was not part of the conspiracy. The established rule stated by courts and commentators is that accomplice liability

extends to acts of the perpetrator which were a "natural and probable consequence"[10] of a criminal scheme the accomplice encouraged or aided.[11] The natural probable consequence liability has been widely accepted in accomplice liability statutes and recodifications.[12]

The natural probable consequence liability seems to be legally suitable for the situations where an AI system committed an offense while the programmer or user did not know about it, intended it or participated in it. The natural probable consequence liability model requires the programmer or the user to be in a mental state of negligence, not more. The programmers or users are not required to know about any forthcoming commission of any offense as a result of their activity, but such a commission is a natural probable consequence of their activity.

A negligent person, in a criminal context, is a person who does not know about the offense, but a reasonable person could have known about it, since the specific offense is a natural probable consequence of that person conduct.[13] The programmers or users of an AI system, who could have known about the probable possibility 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, however, the very legal basis for the criminal liability in negligence cases. Negligence is in fact an awareness omission, or a knowledge omission. The negligent person omitted knowledge, not acts.

The natural probable consequence liability model would permit liability to be predicated upon negligence even when the crime involved requires a different state of mind.[14] Such is not legally legitimate as to one who has personally committed the offense. It is considered legally

---

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

[11] 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).

[12] GABRIEL HALLEVY, THE MATRIX OF DERIVATIVE CRIMINAL LIABILITY 241-247 (Springer Verlag, 2012). See more at State v. Kaiser, 260 Kan. 235, 918 P.2d 629 (1996); United States v. Andrews, 75 F.3d 552 (9th Cir.1996).

[13] 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).

[14] 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).

legitimate as to a person who has not been the physical perpetrator of the offense, but has been one of its intellectual reasons. Reasonable programmers or users could have foresee the offense, and therefore prevent it from being committed by the AI system.

However, the legal results of applying the natural probable consequence liability model upon the programmer or the user are different in two different types of factual cases. The first type of cases is when the programmers or users were negligent as to the commission of the offense while programming or using the AI system with no criminal intent to commit any offense. The second type of cases is when the programmers or users programmed or used the AI system knowingly and willfully in order to commit one offense through the AI system, but the AI system swerved the plan and committed any other offense in addition to the planed offense or in stead of it.

The first type of cases is a pure case of negligence. The programmers or users acted or omitted negligently, therefore there is no reason why not to hold them liable for a negligence offense, if there is such an offense in the specific legal system. Thus, as to the above example, where a programmer of an automatic pilot negligently programmed it to defend its mission with no restrictions on human life taking, the programmer is negligent as to the homicide of the human pilot. As a result, if there is a specific offense of a negligent homicide in that legal system, this is the most severe offense the programmer might be liable to, not manslaughter nor murder that require either knowledge or intent.

The second type of cases resembles the basic idea of the natural probable consequence liability in accomplice liability. The dangerousness of the very association or conspiracy in purpose to commit an offense is the legal reason for the more severe liability to be imposed upon the conspiracy members. For example, a programmer programs an AI system to commit a violent robbery in a bank, but the programmer did not program the AI system to kill anyone. During the execution of the violent robbery the AI system kills one of the people that were present at the bank and resisted the robbery. In such cases criminal negligence liability alone is not enough. The dangerousness of such a situation might not be expressed in negligence alone.

As a result, according to the natural probable consequence liability model, when the programmers or users programmed or used the AI system knowingly and willfully in order to commit one offense through the AI system, but the AI system swerved the plan and committed

any other offense in addition to the planed offense or in stead of it, the programmers or users shall be liable for the offense itself as if it had been committed knowingly and willfully. In the above example of the robbery, the programmer shall be criminally liable for the robbery (if committed) and for the killing as an offense of manslaughter or murder, which require either knowledge or intent.[15]

Still remains the question, what is the criminal liability of the AI system itself when the natural probable consequence liability model is applied. In fact, there are two possible results. If the AI system acted as an innocent agent for not knowing anything about the criminal prohibition, it is not criminally liable for the offense it has committed. In such situations the AI system functions not differently than the functions of the AI system under the first model (the perpetration-by-another liability model). But, if the AI system did not function as an innocent agent, in addition to the criminal liability of the programmer or the user due to the natural probable consequence liability model, the AI system shall be criminally liable for the specific offense directly. A direct liability model of the AI system is the core of the third model as described hereinafter.

D. The Direct Liability Model

The third model does not assume any dependence of the AI system in a specific programmer or user. The third model focuses on the AI system itself, and enables to derivate the outer circles' criminal liability more accurately.[16] Criminal liability for a specific offense is mainly combined out of the external element and the internal element of that offense. Any person, that both elements of the specific offense are attributed to, is held criminally liable for that specific offense. No other qualifications are required in order to impose criminal liability. A person

[15] 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).

[16] 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).

might possess further capabilities, but in order to impose criminal liability the existence of the external element and the internal element required in the specific offense is quite enough.

In order to impose criminal liability upon any kind of entity, these requirements should be proven as existed in the specific entity. When it is proven that a person committed the relevant conduct accompanied with relevant knowledge or intention, the person is criminally liable due to the specific offense. The relevant question towards the criminal liability of AI systems is how these entities might formulate these relevant requirements of criminal liability. Are AI systems different from human persons at this context?

An AI algorithm might have very many features and qualifications, which might be much higher than those of an average human. But, no such features or qualifications are required in order to impose criminal liability. When a human or corporation formulated both the external element and the internal element, a criminal liability is imposed. If an AI system has the capability to formulate both external element and internal element, and in fact it really formulates it, there is nothing to prevent the criminal liability from being imposed upon that AI system.

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

When the offense is committed by an omission, it is even simpler. Under an omission requirement, the AI system is not required to act at all. The very inaction is the legal basis for the omission, as long as there was a duty to act. If a duty to act is imposed upon the AI system, and it does not act, the external element requirement of the specific offense is fulfilled by way of omission.

The attribution of the internal element of the offense to the AI system is the real legal problem in most cases. The attribution of the mental element differs from one AI technology to other.

Most cognitive capabilities developed in the modern AI technology are immaterial for the question of the criminal liability imposition. Creativity is a human feature that humans do share with some animals, but creativity is not required in order to impose criminal liability. Even the most uncreative persons may be criminally liable. All mental requirements that are required 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 their understanding.[17] Most AI systems are well equipped for such a reception. Sensory receptors of sights, voices, physical contacts, touches, etc. are not rare in most AI systems. These receptors transfer the factual data received to the central processing units that analyze them. The analysis process is a process which is parallel to human understanding.[18] The human brain understands the data received by eyes, ears, hands etc. by analyzing that data. Advanced AI algorithms are trying to imitate the human brain understanding process. These processes are not so different.[19]

A requirement of a specific intent is the strongest requirement of the internal element.[20] A specific intent is the existence of a purpose or an aim that a factual event will occur. The specific intent which is required in murder is a purpose or an aim that a certain person will be

---

[17] 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). The Model Penal Code, supra note 14, at subsection 2.02(2)(b) (p. 21) even provides as follows:

> "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).

[18] 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).

[19] 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).

[20] 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).

dead.[21] As a result of the existence of such an intent, the perpetrator of the offense commits the offense, i.e. performs the external element of the specific offense. This situation is not unique to humans. An AI system might be programmed to have a purpose or an aim and to take actions in order to achieve that purpose. This is a specific intent.

One might assert that humans do have feelings which cannot be imitated by AI software, even 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. In spite of that, such feelings are most rarely required in specific offense. Most specific offenses are satisfied with knowledge about the existence of the external element. Few offenses require a specific intent requirement in addition to knowledge. Almost all other offenses are satisfied with much less than that (negligence, recklessness, strict liability). Perhaps in very few specific offenses that do require certain feelings (e.g., crimes of racism out of hatred),[22] it is not legally legitimate to impose criminal liability upon AI system which has no such feelings, but in any other specific offenses it is not a barrier.

If a person established both external and internal elements of a specific offense, the person is criminally liable. Therefore, why should an AI system that established all elements of an offense be exempt from criminal liability? One might argue that some parts of human population are exempt of criminal liability, although both external and internal elements are established. Such parts of the population are, for example, infants and mentally ills.

Infancy do exempt from criminal liability due to a specific legal provision in criminal law.[23] The social rationale of infancy defense is to protect the infants from the harmful consequences

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

[22] 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).

[23] 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).

of the criminal process and let them be socially handled in other social frames.[24] Are there such frames for AI systems? The original legal rationale of infancy defense was the incapability of an infant to understand the fault reflected out of the infant's conduct (*doli incapax*).[25] Later an infant could have been criminally liable if the presumption of mental incapability was rebutted by proof of an ability to distinguish between good and evil.[26] Could that be similarly asserted upon AI systems? Most of AI algorithms are capable of analyzing permitted and forbidden.

Mentally ills are presumed to lack the fault element of the specific offense due to their mental illness (*doli incapax*).[27] That illness defects the mental capabilities to distinguish good and evil

[24] 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).

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

[26] 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).

[27] GABRIEL HALLEVY, THE MATRIX OF INSANITY IN MODERN CRIMINAL LAW (Springer, 2015). See more at 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).

(cognitive capabilities),[28] and to control internal impulses.[29] When an AI algorithm functions properly there is no reason for not using all of its capabilities to process an analysis of the factual data received through its receptors. However, it is an interesting legal question whether a defense of insanity might be attributed to a malfunctioning AI algorithm, which its internal capabilities are defected as a result of that malfunction.

When an AI system establishes all elements of a specific offense, both external and internal, there is no reason to prevent an imposition of criminal liability upon it to that offense. The AI system criminal liability does not replace the criminal liability of the programmers or the users, if criminal liability is imposed on the programmers and users by any other legal path. Criminal liability is not to be divided, but to be added. The criminal liability of the AI system is imposed in addition to the criminal liability of the human programmer or user.

However, the criminal liability of the AI system is not depended on the criminal liability of the programmer or user of that AI system. As a result, if the specific AI system was programmed or used itself by another AI system, the criminal liability of the programmed or used AI system is not influenced by that fact. The programmed or used AI system shall be criminally liable for the specific offense due to the direct liability model, unless it is an innocent agent. In addition, the programmer or user AI system shall be criminally liable for that very offense due to one of the three liability models, according to the specific role of it. The chain of criminal liability might go on if there are more involved entities, human entities or AI systems.

There is no reason to eliminate any criminal liability of an AI system or a human entity, which is based on complicity between them. An AI system and human entity 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 system are sufficient to impose criminal liability upon it, if these capabilities are satisfying the legal

---

[28] 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).

[29] 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).

requirements of the joint perpetrator, the accessories and abettors etc., the relevant criminal liability as joint perpetrator, as an accessory, should be imposed whether it is an AI system or human.

Not only positive factual and mental elements might be attributed to the AI system. All relevant negative fault elements are attributable to AI systems. 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),[30] there is no material difference between humans and AI systems, since they relate to a specific situation (*in rem*) regardless what the offender specific identity is. For example, an AI system that serves under the local police forces is given an order to arrest a person illegally. If the order given is not manifestly illegal, the executer of the order is not criminally liable.[31] In that case, there is no difference whether the executer is human or AI system.

For other defenses (excuses and exempts),[32] some application adjustments should be applied. For example, intoxication defense is applied when the offender is under the physical influence of intoxicating matter (e.g. alcohol, drugs, etc.). The influence of alcohol on AI system is minor, at most, but the influence of an electronic virus that has been infecting the operating system of the AI system might be considered parallel to the influence of intoxicating matters on humans. Some other influences might be considered to be parallel to insanity, automatism etc.

---

[30] 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).

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

[32] 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).

It might be summed up that the criminal liability of an AI system according to the direct liability model is not different from the relevant criminal liability of humans. In some cases some adjustments are required, but substantively it is the very same criminal liability which is founded upon the same elements and examined by the same ways.

E. Coordination of the Three Liability Models

The three liability models described above are not alternative models. These models might be applied coordinately in order to create a full image of criminal liability in the specific context of AI system involvement. Each of the three models does not negate any of the rest of them. Thus, applying the second model is possible as a single model for the specific situation and is possible as one part of a combination of two of the legal models or of all three of them.

When the AI system plays a role of an innocent agent in the perpetration of the specific offense, and the programmer is the only person who directs that perpetration, the application of the perpetration-by-another model (first liability model) is the most appropriate legal model for that situation. In that very situation, when the programmer is itself an AI system (when an AI system programs another AI system to commit a specific offense), the direct liability model (third liability model) is most appropriate to be applied as to the criminal liability of the AI system programmer. The third liability model in that situation is applied in addition to the first liability model, and not instead. Thus, in such situations, the AI system programmer may be criminally liable due to a combination of the perpetration-by-another liability model and the direct liability model.

When the AI system plays a role of the physical perpetrator of the specific offense, but that very offense was not planned to be perpetrated, the natural probable consequence liability model might be appropriate to be applied. The programmer might be considered to be negligent if no offense was deliberately planned to be perpetrated, or the programmer might be considered to be fully liable for that specific offense if another offense was deliberately planned, and the specific offense was perpetrated not as part of the criminal scheme. Nevertheless, when the programmer is not human, the direct liability model is necessary to be applied in addition to the simultaneous application of the natural probable consequence liability

model. So is the situation when the physical perpetrator is human, and the planner is an AI system.

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 systems and criminal law. As a result, when AI systems and human entities are involved, directly or indirectly, in a perpetration of a specific offense, it would be much more difficult to evade criminal liability. The social benefit of such a legal policy is of a very high value. All entities, human, legal or AI, are subordinated to the criminal law. If the clearest purpose of the imposition of criminal liability is the application of the legal social control in the specific society, then the coordinated application of all three models is necessary in the very context of AI systems involvement within the commission of offenses.