South Carolina Law Review

Volume 69 | Issue 3 Article 8

Spring 2018

Could AI Agents Be Held Criminally Liable: Artificial Intelligence and the Challenges for Criminal Law

Dafni Lina

COULD AI AGENTS BE HELD CRIMINALLY LIABLE? ARTIFICIAL INTELLIGENCE AND THE CHALLENGES FOR CRIMINAL LAW

Dafni Lima*

I. INTRODUCTION

For the past few decades, artificial intelligence (AI) seemed like something out of a science fiction work; the concept of a human-made intellect that could gain sufficient autonomy in order to make its own, independent choices is still quite unfamiliar for most. In recent years, rapid technological development has led to products that have evolved to increasingly incorporate AI elements. From smart products to drones to the Internet of Things, social reality has advanced beyond what was technologically feasible when relevant laws were drawn up and enacted. Smart technical systems that can operate in the absence of constant human input pose a set of questions particularly challenging for concepts salient for criminal law and its application in practice.

While in the past AI applications have been used more and more broadly in fields ranging from computer science to finance to medicine, we now stand on the verge of the first major breakthrough in widespread application of AI

* Ph.D. Candidate in Criminal Law at the Aristotle University of Thessaloniki (Greece) and Fulbright Foundation Doctoral Dissertation Visiting Research Student at the Harvard University Initiative on Law and Philosophy (United States). The author can be reached at dafni.lima@cantab.net.

in a way that is recognizable by the mass public: autonomous vehicles.[1] Smart cars that can safely navigate traffic are hardly a fantasy anymore; they have been in development for some years now, and the first versions are already on the streets of major U.S. cities. In 2011, Nevada was the first state to allow and regulate the operation of autonomous vehicles, and as of 2017, thirty-three states have introduced legislation that is related to the issue; twenty of them have already passed relevant legislation, and a further five have seen relevant executive orders issued.[2]

Operation of autonomous vehicles comes with great advantages: it will arguably increase mobility for social groups like the elderly or people with disabilities, it will provide greater safety on the road by providing a more restful travel for professional drivers and arguably guarantee increased adherence to traffic laws, as well as allow drivers to be more productive when travelling, as the autonomous car could take over for the most part. The future of autonomous cars is still not entirely shaped as versions based on a varying degree of automation are developed, some requiring a standby human driver and others being fully autonomous, yet autonomous vehicles in general rely heavily on AI in order to operate.

The advent of what seems to be the first mass application of AI in everyday life and in particular one that tremendously affects transportation—an essential human activity that is intensely regulated by law and where ample opportunities can arise for criminal law to intervene—will undoubtedly have implications that will affect how criminal law is construed and how it is applied. More than that, it will provide an invaluable opportunity to revisit and reflect on traditional criminal law concepts such as personhood, harm,

---

1. See Sven A. Beiker, *Legal Aspects of Autonomous Driving*, 52 SANTA CLARA L. REV. 1145 (2012); Dorothy J. Glancy, *Autonomous and Automated and Connected Cars—Oh My: First Generation Autonomous Cars in the Legal Ecosystem*,16 MINN. J.L. SCI. & TECH. 619 (2015); Adeel Lari et al., *Self-Driving Vehicles and Policy Implications: Current Status of Autonomous Vehicle Development and Minnesota Policy Implications*, 16 MINN. J.L. SCI. & TECH. 735 (2015), for an introduction to the subject. See also AUTONOMOUS DRIVING: TECHNICAL, LEGAL AND SOCIAL ASPECTS (M. Maurer et al. eds., trans., Springer Nature 2016), for several different perspectives into autonomous vehicles, including a social and historical account. See Stephen P. Wood et al., *The Potential Regulatory Challenges of Increasingly Autonomous Motor Vehicles*, 52 SANTA CLARA L. REV. 1423 (2012), for regulatory issues in the U.S. context.

2. See *Self-Driving Vehicles*, NAT'L CONF. STATE LEGISLATURES, http://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx#Enacted%20Autonomous%20Vehicle%20Legislation (last visited Feb. 24, 2018), for these figures as well as further information on actions taken by the fifty states regarding autonomous vehicles.

and blame since it will introduce a new "agent" into the traditional agency spectrum that is defined by capable human actors.

II. WHAT'S IN AN ACT?

One of the first challenges that we encounter when we start to contemplate AI with regard to criminal justice is the implications for one of the fundamental concepts in criminal law: acting. Criminal law is defined by its function as a response to a *crime*, which is construed across western jurisdictions as an *act*.[3] It is a well-established principle of modern criminal law that only acts can incur criminal liability;[4] not thoughts, beliefs, or intentions alone. In both common law and civil law systems, the inquiry into criminal liability starts at the essential level of acting: the concept is reflected in *actus reus* in the first system and included in the German *Tatbestandsmäßigkeit* in the most prominent representative jurisdiction of the latter[5] (which translates to "fulfillment of the elements of the offense," while *Tat* itself means "act").

Although there exists no single, consistent definition that applies to all western jurisdictions about what constitutes an act—or "conduct" in the case of United States law—under criminal law, the same aspects of acting keep coming up in theory and in case law in different legal systems, which speaks to their importance, regardless of whether they are ultimately adopted or not. In the United States, for instance, the Model Penal Code defines criminal liability as such: "A person is not guilty of an offense unless his liability is based on *conduct* which includes a *voluntary act* or the *omission to perform an act* of which he is physically capable,"[6] while under "General Definitions" an act is defined as "bodily movement" (whether voluntary or not).[7] Furthermore, the act requirement is widely regarded as the most notable, or perhaps the only, exception to the rule that substantive criminal law in the United States is not regulated under constitutional law.[8]

---

3. *See* Markus D. Dubber, *Comparative Criminal Law*, *in* THE OXFORD HANDBOOK OF COMPARATIVE LAW 1288, 1320 (Mathias Reimann & Reinhard Zimmermann eds., Oxford Univ. Press 2008) (noting that western jurisdictions require an act to constitute criminal liability).

4. *Id.*

5. *Id.* at 1318.

6. MODEL PENAL CODE § 2.01 (AM. LAW INST. 1962).

7. *Id.* § 1.13.

8. MARKUS D. DUBBER & TATJANA HÖRNLE, CRIMINAL LAW: A COMPARATIVE APPROACH 197 (Oxford Univ. Press 2014).

In Germany, a leading jurisdiction in civil law, the prevailing opinion among criminal law scholars is that an act has to be controllable by the actor and "socially relevant"[9]—in other words, it needs to convey social meaning. An example of this would be, for instance, an act that refers, relates to, or is directed at another person, not just oneself, as liberal theorists would propose in line with John Stuart Mill's famous articulation of the Harm Principle—that power can only be exercised against someone's will in order to prevent harm to others.[10] Further to that, all western jurisdictions have incorporated omission or failure to act into the concepts of acting or conduct. Without going into too much detail, it seems that concepts like bodily movement (or failure thereof) that are voluntary, extroversive, and socially meaningful in a way that is relevant to criminal law are essential aspects of *acting*.

It is important to note here that when a perpetrator uses objects or tools or machines to bring about the desired result, the crime is still considered the perpetrator's acting. When the perpetrator takes advantage of sentient beings, like animals, that do not possess the capacity to reason or fully grasp a situation and the relevant legal implications, criminal law again regards the person manipulating the sentient being as the one "acting." Even in cases of human actors that do not possess full capacity, or alternatively, human actors with full capacity who are forced or tricked into acting for the benefit of another, criminal law often regards this as acting by the person "behind the scenes," while the person who physically committed the act is regarded as a mere instrument of the principal actor. For instance, the German Criminal Code explicitly states under section 25 that a principal is someone who "commits the offense himself or through another."[11]

Against this setting, artificial intelligence raises some extremely interesting questions. First of all, it invites us to consider whether AI agents are acting in the sense of criminal law. And secondly, it urges us to think about different modes of acting when it comes to human agents. These are the two sides of the same question, as an offense that is "committed" by an AI agent, for example an autonomous vehicle running over and thus killing a person, will have to be attributed to someone. Could it be attributed to the AI agent—in which case, we concede that the autonomous vehicle is acting? Should it be attributed to the person behind the scenes—the driver that failed to regain control or perhaps the designer that created an algorithm that allowed this

9. *Id.* at 194.

10. *See* JOHN STUART MILL, ON LIBERTY 17 (1859) ("That the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others.").

11. MICHAEL BOHLANDER, THE GERMAN CRIMINAL CODE: A MODERN ENGLISH TRANSLATION 43 (Hart Publ'g 1st ed. 2008).

development? Or should it be regarded as a product of luck, what would once be referred to as "an act of God"—something that is not attributable to anybody, that is caused by forces of nature or sheer bad luck? Or, finally, should the rise of more and more complex AI agents invite us to reconsider the mere notion of act as the bedrock of contemporary criminal law theory? Will we perhaps need to replace or expand or enrich the arguably obsolete notion of "voluntary bodily movement" against this new landscape?

These questions are far from easy and could hardly be fully answered in the scope of this Paper. Instead, what is attempted here is to illuminate the appropriate questions we need to ask in the face of these technological developments, and thus illustrate which elements will lead criminal law scholars and practitioners to different approaches according to how they evolve in the future, as well as draw an outline of these different approaches. It is also crucial to note that AI will pose challenges for criminal law theory and judicial practice not only because it might invite us to think of sophisticated AI agents as actors of crimes, but also because it introduces further human actors in the query to attribute criminal liability: an AI agent will be, both initially and in terms of how it learns from input and adapts, dependent upon its design and programming, which necessarily includes human agents such as its designers, programmers, and developers as relevant actors. AI agents will also sometimes—or rather, almost always, in the current stage of technological development—interact with an operator, as well as other human actors that they necessarily engage with—for example, with other drivers, in the case of smart cars. All these individuals are "brought" into the scene of the crime for questioning, forcing criminal law to make difficult yet interesting decisions when ascribing liability.

Of course, the answer to these questions cannot be given without knowledge of the answer to the most fundamental question of all: what is AI and what is it capable of doing?[12] Since AI is not one thing but is constantly evolving, the answer—and with it, criminal law's response—will hugely depend upon the individual facts of the case at hand. An autonomous vehicle that should at all times be supervised by a present, competent, and legally licensed driver,[13] for example, is a totally different scenario than a fully autonomous vehicle that drives a minor or a drunk person safely home. Yet

12. *See, e.g.*, Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 404–05 (2017).

13. *See* Frank Douma & Sarah Aue Palodichuk, *Criminal Liability Issues Created by Autonomous Vehicles*, 52 SANTA CLARA L. REV. 1157, 1163 (2012) ("If legislators determine they will require a licensed driver in the driver's seat, they must then decide whether to place criminal responsibility on that driver for failing to respond to technological malfunctions.").

criminal law needs to prepare for both these possibilities and provide tailored responses.

In general, artificial intelligence is associated with the ability to adapt according to the feedback received in order to solve problems and address situations that go beyond the predefined set of queries and instructions that the AI was programmed with. In essence, artificial intelligence mirrors the human ability to process information and learn. As such, it can "decide" how to respond to unprecedented scenarios and also "choose" how to navigate a novel situation towards successfully achieving some objective. As AI applications expand and humans become more comfortable with them, many envision AI that will become truly independent from its human counterparts and take on a life of its own.

Under the current state of development, it seems that AI actions could hardly fall under the definition of acting.[14] Even if we set aside as obsolete the "bodily" dimension of acting, which by definition could never apply to a machine, an intelligent agent's movements could neither be seen as "socially relevant" nor as "voluntary" in the sense that criminal law implies. Social relevance may be grounded in a specific historical context, but it is built over time through evolution of social dynamics and perceptions,[15] and AI agents are still too young to have gathered such a "critical mass" of social meaning and importance. This, however, may change in the future as individuals and societies become more and more familiar with AI agents, especially service robots that assimilate a human-like appearance.

As for voluntariness, this could be at first glance attributed to any agent that "chooses" based on a given set of facts, so that even a computer choosing one of two available options based on input and a set objective might be said to choose. But on a deeper level, voluntariness, even in bodily movements, is rooted in the ability for judgment and free will. That is why, for instance, a person's bodily movement while sleepwalking or as a reflex does not count as voluntary under criminal law,[16] and this emphasis on the ability for judgment is reflected even further in the context of blame and punishment.[17]

---

14. *See* Sabine Gless et al., *If Robots Cause Harm, Who is to Blame? Self-Driving Cars and Criminal Liability*, 19 NEW CRIM. L. REV. 412, 417 (2016).

15. For example, let us not forget that at one time even animals were put on criminal trial and sentenced—often to death. *See, e.g.*, E. P. EVANS, THE CRIMINAL PROSECUTION AND CAPITAL PUNISHMENT OF ANIMALS (The Lawbook Exchange, Ltd. 2009) (1906).

16. *See* Thomas Weigend, *Germany*, *in* THE HANDBOOK OF COMPARATIVE CRIMINAL LAW 260 (Stanford Law Books 1st ed. 2010).

17. *See generally* Anthony Honoré, *Causation in the Law*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY, https://plato.stanford.edu/archives/win2010/entries/causation-law (last updated Nov. 17, 2010).

Voluntariness in this sense implies the ability to act otherwise, and an agent that is programmed to choose *A* when it encounters *B* is not necessarily choosing. Thus, AI agents do not yet seem to possess the potential for fully independent, even self-destructive decisions. In other words, no one would regard a robot's choice to change its route when stumbling upon a table as voluntary, so long as the robot is simply following an algorithm, however intricate, that dictates it to change route when encountering a physical obstacle—or to put it more simply, so long as the robot does not have the choice to keep hitting at the obstacle if it so wishes. This holds true even when this choice is the only reasonable one and in the AI agent's "benefit" of achieving its objective.

The way that we approach the problem of acting in the context of AI agents will also determine the way we respond to the causality requirement that is inherent in criminal law. It is widely accepted that no criminal liability may arise for a result unless it is causally linked to an agent's conduct—in simpler terms, unless the result was brought about due to the agent's conduct, the agent is not considered the perpetrator of the act.

Causality or causation has been the subject of much debate in criminal law theory,[18] with approaches ranging from a focus on necessity (the "but-for" test and its variations) to sufficiency (such as the NESS test). A simple example of causation would be the following: intending to bring about *X*'s death, I drop poison in their tea, which *X* drinks. The poison needs a few hours to take effect and in the meantime, *X* is run over by a car while trying to cross a busy street. As much as I wanted *X* to be dead and as much as I tried to bring about this desirable result, their death is not my doing. The chain of causation has been broken by the reckless driver (or even the autonomous car) that claimed *X*'s life first.[19] But if the reckless driver merely injured *X* by running over them, and before dying from these injuries the poison took effect and brought about *X*'s death, then *X*'s death is still my doing even though they would have died anyway from the wounds sustained in the car accident.

Therefore, in order to break the causal link between a conduct and a harm, there needs to exist an intervention significant enough to warrant such a split, and this is usually the case in the event of forces of nature at play or

18. *See generally* MICHAEL BOHLANDER, PRINCIPLES OF GERMAN CRIMINAL LAW 45 (2009); H.L.A. HART & ANTHONY HONORÉ, CAUSATION IN THE LAW (2d ed. 2002); MICHAEL S. MOORE, CAUSATION AND RESPONSIBILITY: AN ESSAY IN LAW, MORALS, AND METAPHYSICS (2010); Jane Stapleton, *Law, Causation and Common Sense*, 8 OXFORD J. LEGAL STUD. 111 (1988); Honoré, *supra* note 17.

19. See R v. White [1910] 79 Crim. App. 854 at 856 (Eng.), for a variation of this example which established the "but-for" test in English case law.

involvement of third parties.[20] There are several more sophisticated versions of causation that are widely accepted. For example, if I plant a bomb in someone's car with the intent to kill that person, but in order for the bomb to go off the person needs to turn on the ignition and they do, then strictly speaking it was the person's action that ultimately brought about their death. Yet in such cases it is widely accepted that the result is still caused by the original agent's doing and that the "intervening" act is not enough to break the chain of causation[21]—much in the same vein as in "perpetration by another," the victim here is used instrumentally by the perpetrator. Similarly, setting up a machine or a mechanism that will ultimately bring about the harmful result even when the offender is not present is not sufficient to break the causal relationship; the end result is still the offender's doing.

In this context, whether or not one regards an AI agent's actions as acting in the criminal law sense is crucial for causation. If an AI agent is merely an instrument at the hands of the human agent, much like an inanimate tool such as a hammer or a knife, then the answer is simple. But matters become slightly more complex when we conceive of AI that is complex enough to perceive a situation and proceed with acting—or, fail to act where it could have acted and thus allow the harmful result to come about. Yet whether or not we understand the "choices" made by AI as acting is closely linked to how we perceive other issues, such as the pivotal question of personhood.

III. REVISITING PERSONHOOD AND BLAME

Artificial intelligence by definition mimics one of the essential traits of the human species, that of adapting to one's environment, and as such, it invites us to revisit our understanding of personhood.[22] Personhood is a concept that underlies not only criminal law, but every field of law, as it is closely linked to our capacity to perform legally meaningful acts and bring about legally relevant developments. Historically, our understanding of what it means to be a person has been connected to human ability for self-reflection

---

20. *See* Gless et al., *supra* note 14, at 429–30.

21. *See, e.g.*, R v. Pagett [1983] 76 Crim. App. 279 (Eng.); R v. Roberts [1971] 56 Crim. App. 95 (Eng.).

22. *See generally, e.g.*, Bert-Jaap Koops et al., *Bridging the Accountability Gap: Rights for New Entities in the Information Society?*, 11 MINN. J.L. SCI. & TECH. 497 (2010) (illustrating a very thorough account of the debate with several further references); Kamil Muzyka, *The Outline of Personhood Law Regarding Artificial Intelligences and Emulated Human Entities*, 4 J. ARTIFICIAL GEN. INTELLIGENCE 164 (2013); Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231 (1992) (discussing the broader issue of personhood with regard to AI).

and self-conscience,[23] that is, our ability to perceive our independent existence and its boundaries that stretch into the past and future. As things stand currently, AI units do not seem to possess that same degree of self-awareness (or any at all) that would allow us to consider their situation as equivalent to the human experience—although this might change in the future.[24]

On some level, personhood is also associated with our ability to set goals for ourselves and pursue them, which for now seems to be extremely restricted when it comes to AI agents. While they might possess the ability to scale and set independent, smaller objectives in order to reach their overall goal, this greater objective is still set by the human programmer or user (or even another AI programmer or user that has been in turn initially developed by a human). In the case of autonomous vehicles, for example, while the AI software might be in a position to make decisions on the spot regarding traffic, the overall goal of safely navigating to the occasional desired destination is predetermined.

It would be an oversight not to note that there is truth in the statement that our own humanly possible perception of our awareness and our degree of freedom in setting our own goals and in making choices is far from complete. Quite often there are factors at play that restrict our freedom and distort our awareness, while philosophers and scientists are still contemplating on how exactly we form our self-understanding and our conscience. Nonetheless, there is an obvious qualitative difference between our own, at times fuzzy or inexplicable, ability to self-reflect and an AI agent's shortcomings on the same matter.

If an AI agent cannot be considered a person, it could not prima facie enjoy rights and be bound by obligations as humans do. There is again a qualitative difference between a restriction and an obligation, and while an AI unit may be programmed to adhere to certain restrictions, so long as this adherence is not the product of its own volition, it cannot be considered an "obligation" as such. However, when we turn to the issue of rights,[25] things slightly change; it is widely accepted that rights function quite differently than obligations for subjects that are not considered capable of undertaking obligations under law. For example, a minor can often enter into contracts that

---

23. See Gless et al., *supra* note 14, at 415, for a fuller account.

24. *Id.* at 416.

25. "Rights" and "obligations" are used in a generalizing fashion in order to accommodate the scope of this Paper. For a more nuanced understanding of rights and obligations, as well as a starting point to consider more accurate descriptions of legal categories that might better fit AI agents, see Wesley N. Hohfeld, *Some Fundamental Legal Conceptions as Applied in Judicial Reasoning*, 23 YALE L.J. 16, 16–59 (1913).

convey upon them benefits but not obligations,[26] or which are valid with regard to rights conferred and void with regard to obligations. Lately, a lot has been said on the issue of recognizing animal rights,[27] not least because we have finally begun to understand that animals are sentient beings that experience a much wider range of feelings than previously acknowledged;[28] as both research and legal scholarship advances on this matter, it might be conceivable that certain developments might be suitable for transposing in the field of AI agents with regard to their "rights" or "freedoms."

In the context of criminal law, personhood is closely associated with blame, as only a person who can distinguish right from wrong and is in a position to choose can be blamed for choosing to do wrong.[29] Blame presupposes the ability to comprehend what each choice will entail and the ability to freely choose. Historically, this goes beyond simply associating one option with criminal law repercussions and the other with walking free—although in practice it might very well be reduced to that. In that respect, it must be noted that the focus of deterrence theories is precisely on simply discouraging people from committing crimes, regardless of their inner motives, while seminal legal positivist teachings[30] are in part dedicated to freeing adherence to legal rules from the burden of inextricable association with moral considerations.

Against this setting, it is important to note a sometimes overlooked aspect, namely that *mens rea* and blame requirements were originally devised as a safeguard against abuse of state power in the exercise of criminal law enforcement; they were meant to ensure that no one would be held accountable for a crime if the person was mentally unaware of what had happened or did not engage in it with some degree of volition or acquiescence. Anyone held criminally liable for a conduct should have had some level of

---

26. *See, e.g.*, BÜRGERLICHES GESETZBUCH [BGB] [CIVIL CODE], § 107, *translation at* https://www.gesetze-im-internet.de/englisch_bgb/englisch_bgb.html#p0323 (Ger.).

27. *See, e.g.*, ANIMAL RIGHTS, CURRENT DEBATES AND NEW DIRECTIONS (Cass R. Sunstein & Martha C. Nussbaum eds., 2004); MARK ROWLANDS, ANIMAL RIGHTS, MORAL THEORY AND PRACTICE (Palgrave Macmillan, 2d ed. 2009); Jessica Eisen & Kristen Stilt, *Protection and Status of Animals*, *in* MAX PLANCK ENCYCLOPEDIA OF COMP. CONST. LAW (Rainer Grote et al. eds., 2015).

28. For example, in 2009, the EU with the Lisbon Treaty recognized that animals are "sentient beings," building on its previous legacy of recognizing the Five Freedoms for animals kept for farming purposes: "Freedom from hunger and thirst, Freedom from discomfort, Freedom from pain, injury, and disease, Freedom to express normal behaviour, and Freedom from fear and distress." *Animal Welfare*, EUROPEAN COMMISSION, https://ec.europa.eu/food/animals/welfare_en (last updated Feb. 24, 2018).

29. *See* Gless et al., *supra* note 14, at 419.

30. *See* H.L.A. Hart, *The Concept of Law*, *in* CLARENDON L. SER., 167, 167–80 (Oxford Univ. Press 3d ed. 2012).

knowledge and intent (or the duty to have known and to take care to avoid) with regard to the results of their actions.

This once groundbreaking development tapped into our collective innate human ability to understand, pass moral judgment on, and manipulate our actions. It also reflected a deep respect for human beings, as it treated them on the basis of their informed choices; one would only suffer the consequences for their actions because they *chose* so. This approach rests, on a deeper level, on respect for the freedom to even act wrongly and inflict harm—it is only when one conscientiously makes that choice, that they will be punished. This is why children, for instance, who do not yet fully apprehend the consequences of their actions, or persons with mental health challenges that prevent them from reasoning properly, are treated differently under criminal law. Ultimately, criminal liability is a response reserved for those who *could* have risen to the occasion—but *chose* not to.

Again, this approach is arguably a shortcut; it casts aside any particularly sophisticated concerns about how human intent is formulated as well as any doubts about whether our free will is indeed free and our own after all. As law so often does in general, this is both a generalization and a simplification—and one might even sense a hint of declaration captured in it.

In any case, the move away from torture, forced labor as punishment, and capital punishment (for most of the Western world) equally mirrored respect for a perpetrator's innate humanity; in principle, the law is not allowed to touch a convict's body or take their life. Similarly, the general rule that a fair and just trial by a judicial body is required before any imprisonment can legitimately be imposed is again the result of respect for what it means to be human. In that sense, it seems that modern criminal law and all its progressive developments were designed by humans for humans and always revolved around the fact that we all share some innately human quality that needs to be respected even in our ugliest hour. Of course, this progressive undercurrent is not without exceptions or occasional regress, but it lies at the heart of modern criminal law theory and practice.

At the time of this development in criminal law theory, only human agents possessed this type of intellect that forms the basis of criminal liability. Animals, although they do have the ability to communicate and make qualified choices to some extent, do not possess the same level of ability to understand or choose between right and wrong—or, in any case, between what the law prohibits and what it allows or demands.

Legal persons, on the other hand, which are the sole prominent example of extending criminal liability beyond human actors, are still based on human agency. First, they are in essence legal fictions, a translation of our collective efforts into legally relevant terms, and as such are not endowed with any type of intellect—although there is something to be said about corporate culture

and the way a collective agent can over time establish mechanisms and processes that surpass its individual members. Yet in contrast to animals, which are clearly something radically different from humans but do not have the same legally relevant capabilities, corporations caught the eye of criminal law precisely because they are so closely entwined with human agents.

Corporations are made up by humans who sometimes deliberately use them to escape responsibility for criminal conduct, and this is part of the reason why criminal law in many jurisdictions has stepped in and introduced some form of "criminal liability" for legal persons. Yet there is something to be said for the fact that, in many jurisdictions, legal persons are not subject to criminal penalties, but only administrative sanctions,[31] precisely because criminal law cannot concern itself with agents that cannot make moral decisions and thus cannot be blamed.[32]

Artificial intelligence is completely different from both animals and legal persons. It is not alive, like animals, yet it is not simply a fiction, like corporations. Yet it could be conceived of existing (at least after its initial creation) independently and without the involvement of humans and it could reason, which sets it apart from both legal persons in the first respect and from animals in the latter. Ultimately, it is an open question whether AI might in the future develop a form of conscience and even the capacity for ethics and reasoning that might allow it to be subjected to blame on par with human agents[33]—which is not the case with legal persons or animals. But as long as both our understanding and the practicality of blame are associated with self-awareness and conscious decisions rooted in the human experience, AI agents cannot partake.

The same point could be made about punishment. Even though we could conceive of punishments for AI agents that are roughly "equivalent" to those for humans,[34] there is still an argument to be made that these equivalent

---

31. Weigend, *supra* note 16, at 266.

32. *Id.*; *see also* Gless et al., *supra* note 14, at 416–17; Thomas Weigend, *Societas Delinquere Non Potest?: A German Perspective*, 6 J. INT'L CRIM. JUST. 927, 936 (furthering the argument that corporations cannot act, be blamed, or be punished).

33. *See generally* WENDELL WALLACH & COLLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG 9–11 (Oxford Univ. Press 2008) (furthering the fascinating subject of robots, ethics, and morals); Amitai Etzioni & Oren Etzioni, *Keeping AI Legal*, 19 VAND. J. ENT. & TECH. L. 133, 133 (2016) (arguing about the dangers AI might pose for legal order).

34. *See* Gabriel Hallevy, *I, Robot- I, Criminal- When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses*, 22 SYRACUSE SCI. & TECH. L. REP. 1, 29–35 (2010) (theorizing punishment adjustments for AI robots).

sanctions are slightly beside the point. All major theories about punishment,[35] from retributivism to rehabilitation (save perhaps for specific deterrence),[36] presuppose a communicative aspect[37] among agents that in theory participate equally in a shared experience of the world and in awareness of their own and each other's existence. Punishment is a collective means of responding to crime directed at an agent that can understand its significance as well as its relevance to their criminal conduct[38]—which is why people with diminished capacity are, as a rule, not subject to criminal sanctions.

If an AI software were deleted as a form of capital punishment, would anyone say that "it got what it deserved" in the context of the "just deserts" approach? And if it was deactivated for a certain period of time, could we truly hope that other AI units would be deterred from engaging in similar conduct? Until a positive answer to at least one of these questions appears likely, debate about criminal punishment for AI agents seems somewhat misplaced.

IV. POTENTIAL OPTIONS FOR ASCRIBING CRIMINAL LIABILITY

In the case where a result is brought about by an "action" (or "omission") on part of an AI agent, then an inquiry about ascribing criminal liability arises. The answer to how—and if—criminal liability should be attributed will heavily depend on the circumstances of each case, as outlined below. In each of these cases, approaches and concepts already familiar to criminal law might offer the solution; however, the focus will shift to the way legal professionals, lawmakers, judges, and practitioners will adapt, enrich, or decide to firmly hold on to their current understandings of these concepts.

---

35. *See generally* OXFORD UNIV. PRESS, A READER ON PUNISHMENT (Antony Duff & David Garland eds., 1994); OXFORD UNIV. PRESS, RETRIBUTIVISM HAS A PAST: HAS IT A FUTURE? (Michael Tonry ed., 2011); R. A. DUFF, ANSWERING FOR CRIME: RESPONSIBILITY AND LIABILITY IN CRIMINAL LAW (Hart Publ'g ed. 2007) (resources discussing different theories of punishment).

36. *See generally* Joshua Dressler, *Deterrence*, *in* 2 THE ENCYCLOPEDIA OF CRIME & JUSTICE 507, 507–14 (Thomas J. Bernard et al. eds., 2002).

37. For a discussion on the communicative aspect of punishment, see generally R.A. DUFF, PUNISHMENT, COMMUNICATION, AND COMMUNITY (Oxford Univ. Press ed. 2001); JOEL FEINBERG, DOING AND DESERVING: ESSAYS IN THE THEORY OF RESPONSIBILITY 95–118 (Princeton Univ. Press ed. 1970); ANDREW VON HIRSCH, CENSURE AND SANCTIONS 6–19 (Oxford Univ. Press ed. 1996).

38. *See* Gless et al., *supra* note 14, at 421–22.

A. Instrumental Use of an AI Agent

The first and easiest scenario is quite straightforward: what if a human actor manipulates an AI agent into doing the human's bidding, with the intent to commit a particular crime? In such cases, the obvious solution is to hold the person manipulating the AI agent accountable. This could be a programmer that successfully inserts an algorithm designed to kill into AI software or an operator that instructs AI software so that it will inflict harm to others. In any case, the AI agent cannot be regarded as anything else but a tool in the hands of the human "behind the curtain."

However, the path by which to ascribe liability might differ according to the level of sophistication that the AI agent possesses. In the case of tools like a hammer, for example, we are never speaking of "ascribing" the action of the hammer to the human using it—the movement of the tool is immediately understood as the action of the human agent. In the case of animals, we often equate them in legal terms with things that can be manipulated by their master (although they could never be controlled in an absolute sense, like a tool). In both these cases, we regard the human actor as the perpetrator of the criminal act.

Things start to change when we encounter the possibility of a human using another human as a "means" to commit a crime. In these cases, for example when an individual is tricked in order to shoot at someone thinking that the individual was only shooting at an inanimate target or when a nurse is tricked into giving poison to a patient thinking they were only administering a medicine, we could talk of perpetration by another. Yet this approach commands the existence of an intermediary (the "another") who is, in theory, in a position to intervene as the events that constitute the criminal conduct unfold—a person who could understand what is going on or who, in any case, could choose to act otherwise. If this is not the case, we would not talk about perpetration by another but simply about "perpetration," as we do with animals. "Another" is a direct reference to "another human."

In order, then, for this theory to make sense in the context of AI agents, they should be sophisticated and intricate enough to have the capability to understand what was going on and to choose accordingly—even if in the end they were tricked into the desired conduct by the perpetrator behind the scenes. One could argue that an autonomous vehicle that was simply programmed to go on the street and run over people is quite a different scenario than a driver who manipulates an AI car into regarding a particular person as a mere object they can safely run over. One could even begin to feel the "pull" of an ethical condemnation against the human actor in the second case, as an (artificially) intelligent agent is manipulated into committing a harmful action it would otherwise never choose to do. Ultimately, it all

depends on whether technological progress will allow us to view AI agents as sufficiently human-like or not.

At this point, it is also interesting to note that there are cases that might occur where an AI agent goes beyond the originally intended criminal act. For example, an autonomous vehicle is programmed to go out and injure a human, but instead ends up killing the human. In those cases, the end result is something different than the human actor has intended, and the theory of ascribing liability based on the foreseeability and probability of the crime that was actually committed as a consequence of the intended criminal conduct might prove useful.[39]

This model is usually employed when ascribing liability to an accomplice or an instigator and is based on a type of negligence on part of the accomplice or instigator. Under this model, criminal liability is ascribed to an accomplice or an instigator when they could and should have foreseen the different result that came about as a probable consequence of the initially intended act. Thus, in our example, the human actor could be held liable if the killing was a probable and foreseeable consequence of the human's order to the autonomous vehicle to go out and injure a particular person.[40] If, however, the crime ultimately committed had nothing to do with the one intended (e.g., a robot is ordered to steal a letter and instead burns down a house), then the perpetrator behind the scenes cannot be held criminally liable.

B. Negligence and Recklessness

On a similar note, negligence is the model that most fittingly can be used to ascribe criminal liability for unintended conduct that occurs in the context of an AI agent's usual programming or use—that is, as it carries out its duties without malfunction. Here, the focus shifts on a benevolent designer or operator who neglected to take due care in order to prevent an undesirable outcome that could occur within the usual performance of the AI agent and which the programmer or user should have foreseen. In these cases, the AI agent functions appropriately and in the discharge of its duties commits a crime—a simple example would be a cleaning robot that destroys valuable property mistaking it for dirt.[41]

---

39. Hallevy, *supra* note 34, at 4.

40. *See generally* IRYNA MARCHUK, THE FUNDAMENTAL CONCEPT OF CRIME IN INTERNATIONAL CRIMINAL LAW: A COMPARATIVE LAW ANALYSIS 167–98 (2014) (discussing the concept of joint criminal enterprise under international criminal law and its origins).

41. Gless et al., *supra* note 14, at 423 (discussing negligence that results in bodily injury).

In such cases, the main question to be answered is whether the programmer or the user could have foreseen this development and whether they were in a position to act in order to prevent it. Negligence, in essence, revolves around the duty to take appropriate and reasonable care to prevent harm to others and focuses on the foreseeability of the undesirable outcome.[42] In cases where the human agent actually foresaw the outcome and decided to disregard it—and according to the jurisdiction—recklessness would be the appropriate model to ascribe liability.

C. Respondeat Superior?

Strict liability is not unheard of in criminal law, but it stands in stark tension with many of its underlying principles—some of which, regarding free will and the innately human capacity to make (even wrongful) decisions, were discussed above. Yet in many western jurisdictions, strict liability offenses exist, from drug possession to particularly minor offenses like driving infractions. The concept of vicarious liability (or, in very fitting to the theme at hand terms, of *respondeat superior*—"let the master answer")[43] derives mainly from tort law, where it is particularly applied to impose liability on a person in control of another (such as an employer with regard to an employee) for the wrongdoing of their agent.[44]

This relationship between an agent and a superior appears at first uniquely suitable to the situation at hand. Just like with AI agents, in the case of vicarious liability, the agent that committed the wrongdoing is an independently intelligent and capable one. However, the concept is radically transformed when transposed in criminal law—and for good reason. One cannot tolerate the same low threshold of intellectual and volitional involvement for the obligation to undertake responsibility for a tort and for a

---

42. *Id.* at 423–24 (discussing negligence more generally in the context of AI agents).

43. *Id.* at 416.

44. *See id.* at 414 (discussing damages caused by AI agents); *see generally* Sophia H. Duffy & Jamie Patrick Hopkins, *Sit, Stay, Drive: The Future of Autonomous Car Liability*, 16 SMU SCI. & TECH. L. REV. 453 (2013) (explaining how applying a strict liability regime for autonomous cars will equitably assess liability without unduly hindering innovation); Kyle Graham, *Of Frightened Horses and Autonomous Vehicles: Tort Law and Its Assimilation of Innovations*, 52 SANTA CLARA L. REV. 1241 (2012) (discussing the uncertainty in predicting the interplay of innovation and liability in the context of autonomous cars); Gary E. Marchant & Rachel A. Lindor, *The Coming Collision Between Autonomous Vehicles and the Liability System*, 52 SANTA CLARA L. REV. 1321 (2012) (discussing how autonomous cars will reduce the number of vehicular accidents yet still pose liability concerns for manufacturers); Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 354 (2016) (advocating for the application of a tort system as opposed to direct regulation of autonomous vehicles).

crime. Criminal law is typically associated with grave consequences for the one found to bear liability, so the threshold must be higher.

This point has also a more general insight to offer: any potential model of ascribing liability for the human agent who is somehow involved in a crime committed by an AI agent will have to vary not only depending on circumstances, such as the sophistication of the intelligence of the AI agent or the degree of control of the human agent, but also on the type of crime committed. In other words, the threshold should be higher for serious crimes, such as killing, and could be lower for relatively minor ones, such as the destruction of an inexpensive item that belongs to a third party.

In the case of strict liability, not only is our deeper understanding of what criminal law is and what it does at stake, but also different and competing policy concerns. Introducing strict liability might satisfy a social demand for accountability that could prove crucial in the acceptance and wider use of AI agents; on the other hand, it could undermine the potential to further develop AI applications because the designers or operators would be discouraged by the likelihood of being found criminally liable for acts they did not intend or concede to.[45] In this context, strict liability could either be reserved only for minor offenses when they fall within the margin of error on the part of the human agent, whether it is a programming or an operating error, or it could be discarded completely as a model for ascribing criminal liability. Perhaps the best way to consider strict liability is in a context where it is combined with negligence requirements, in an approach modeled after (criminal) liability for faulty products.[46]

D. Other Options: Direct Liability or Bad Luck

Even if everything is done properly on the part of human agents, an AI agent might still malfunction and thus cause harm. In these cases, no human is at fault, and the question of what to do with criminal liability remains open. Another important and extreme scenario to consider is when an AI agent "deliberately" inflicts harm.

The second scenario seems far-fetched for now. As AI is not yet at a stage where it could really choose to do wrong, as discussed above, imposing direct criminal liability should be ruled out.[47] If and when AI sufficiently develops

45. *See* Gless et al., *supra* note 14, at 432–33.

46. *See id.* at 425–31. *But cf.* Woodrow Hartzog, *Unfair and Deceptive Robots*, 74 MD. L. REV. 785, 811–32 (2015) (arguing the Federal Trade Commission is the appropriate authority for regulating robots).

47. *But see* Hallevy, *supra* note 34, at 19.

to conform to some of the criteria set out above, then this question might be reconsidered. Even in those cases, however, a malfunction cannot be blamed on an AI agent any more than acts performed while intoxicated can be blamed on a human agent.[48]

In such cases of malfunction, it is proposed that humans should learn to live with this unfortunate development, much in the same vein that they have learned to live with the results of a bridge collapsing due to a hurricane or a flat tire that leads to a car accident.[49] Not everything can be foreseen, prevented, or contained, and in everyday life there are several instances where no one is to blame—much more be held criminally liable—for an undesirable outcome. In other words, not everything can—or should—be regulated under criminal law. Depending on the familiarity that humans will develop with AI agents in the future, this option might prove to be a viable alternative to criminal liability, even though policy implications have to be considered as it is likely that AI acceptance rates might suffer at first.[50]

V. FINAL THOUGHTS: CAN AI AGENTS TRULY MURDER?

Artificial intelligence and its development in the next years will undoubtedly pose great challenges for criminal law, which go beyond the question of criminal liability. With new technology and a far more widespread use of AI agents than is currently conceivable, new opportunities for crime will arise. For instance, if autonomous vehicles become commonplace on our streets, we will sooner or later need to think about new types of crimes that could be committed by hackers and how to prevent the commission of terrorism offenses that could be perpetrated by using the extended capabilities of smart cars.[51] Furthermore, new legal rules will have to be devised to regulate safe driving and relevant crimes;[52] the relationship between an autonomous vehicle, its driver and passengers, and third parties (other drivers,

48. *See generally* Hallevy, *supra* note 34, at 23–27 (discussing the applicability of defenses for AI robots).

49. *See* Gless et al., *supra* note 14, at 19 (arguing that AI agents should be viewed as "an exceptional risk" as opposed to a "normal risk").

50. *See* Gless et al., *supra* note 14, at 430–31.

51. *See* Douma & Palodichuk, *supra* note 13, at 1164–68 (discussing the implications of autonomous vehicles being used in the perpetration of criminal activity or terrorism).

52. *See generally* Jeffrey K. Gurney, *Driving Into the Unknown: Examining the Crossroads of Criminal Law and Autonomous Vehicles*, 5 WAKE FOREST J.L. & POL'Y 393 (2015) (explaining the need for federal and state regulators to amend automobile criminal laws to aid the introduction of autonomous vehicles into society).

passengers, or pedestrians); insurance and tort claims;[53] and privacy with regard to autonomous vehicles.[54] Finally, law enforcement will have to be equipped with new powers and duties in order to address the new situation; for example, we will need to think about under which circumstances a law enforcement officer might be allowed to pull over an autonomous vehicle, and how.[55]

However, the very first wave of vibrations that will be felt in criminal law will undoubtedly include issues that revolve around criminal liability. In this context, legal professionals will be invited to revisit, enrich, and reshape fundamental concepts, as discussed above. Lawmakers and common law judges will have to come up with models that adequately address allocation and imposition of criminal liability, practitioners and adjudicators will have to understand how to best apply them in practice, and research by legal scholars will have to shift focus in order to inform this debate. The results might be as groundbreaking as AI technology itself; these reforms might even one day lead us to reconsider the very foundations of criminal liability, wrongful acts, and blame.

There exist among legal scholars opinions already in favor of imposition of criminal liability on AI agents.[56] Yet similar suggestions seem to rely, at least with regard to how things currently stand, on a circular argument that begs the question. They appear to take for granted the axiom that AI agents *can* fulfill the requirements for *mens rea*, even though *mens rea* as a concept was clearly conceived with human agents in mind—including criminal liability of legal persons, since these are no more than collective enterprises made up of human agents, in which case the criminal liability claim rests on the law's inability to "pierce the veil" and ascribe liability to the human behind the corporate fiction, as explained above.

Yet AI is something completely different. It is certainly no fiction anymore but independent and potentially able to become fully autonomous. If it is to be handled with legal tools that were devised for humans, we must

---

53. *See generally* Robert W. Peterson, *New Technology—Old Law: Autonomous Vehicles and California's Insurance Framework*, 52 SANTA CLARA L. REV. 1341 (2012).

54. *See generally* Dorothy J. Glancy, *Privacy in Autonomous Vehicles*, 52 SANTA CLARA L. REV. 1171 (2012) (discussing the challenges of reconciling privacy concerns with autonomous vehicle technology and suggesting some solutions); Sarah Aue Palodichuk, *Driving into the Digital Age: How SDVs Will Change the Law and Its Enforcement*, 16 MINN. J.L. SCI. & TECH. 827, 833–41 (2015) (discussing how autonomous vehicles will contain information that the Supreme Court has held in other contexts to be protected by the Fourth Amendment).

55. *See* Douma & Palodichuk, *supra* note 13, at 1167–68.

56. *See* Hallevy, *supra* note 34, at 35–37; *see generally* GABRIEL HALLEVY, WHEN ROBOTS KILL: ARTIFICIAL INTELLIGENCE UNDER CRIMINAL LAW (2013) (providing an elaborate account in favor of criminal liability of AI agents).

establish either that it is sufficiently human-like, which does not yet seem to be the case, or that the tools at hand are also suitable for non-humans, which especially in the case of *mens rea* and blame is, at the very least, a matter of dispute, as the whole concept reflects our collective experience of what it means to be human.

So, taking for granted that *mens rea* requirements could aptly be fulfilled by non-human (or, rather, non-human-like) intelligent agents necessarily presupposes the perception of historically and empirically informed concepts such as choice, voluntariness, knowledge, and intent as simply technical terms without any inextricable grounding in the human experience. This is a bold and perhaps forward-looking approach, but one that cannot be taken as self-evident without first examining those perspectives that would work against it—some of which this Paper has attempted to articulate.

If current criminal law concepts were devised for those sharing in the human experience of the world and its ethical dilemmas, and if the way AI agents experience the world is not (yet) at that point, then what is there left to do with criminal liability? It is important to note that even though artificial intelligence is still not at the same level of capacity for intellectual and emotional investment as humans, it may very well one day be—as countless works of science fiction have been trying to warn us. If and when that day comes, the situation might be very different with regard to criminal law and its application to AI agents. On that day, we may be prepared to directly ascribe criminal liability to AI actors and regard them as equally capable of making ethically informed choices and committing wrongdoing—we might even invite each other to share in the legislative and judicial process of responding to crime.

But until then, criminal law might not be the appropriate vessel for holding AI agents accountable. Although criminal law carries with it a connotation of moral condemnation that is very much socially desired in situations of harm to others, especially in serious crimes such as bodily injury or killing, a softer version of the State's powers to prohibit and punish behavior might be more appropriate—for example, administrative sanctions or a whole new field of law in-between. The desire to call a sanction "criminal" and as such satisfy the need to respond to an undesirable conduct by the gravity and resolution that criminal law means carry with them, bears a hidden yet crucial danger. Instead of strengthening our response to harmful and wrongful behavior, it might just weaken our perception of what criminal law is and what it has the power to do, and thus qualify it with a degree of levity that will in turn allow us to underestimate its potential to inflict harm on humans and sap our vigilance with regard to its advances.