ABSTRACT: Whether causing flash crashes in financial markets, purchasing illegal drugs, or running over pedestrians, AI is increasingly engaging in activity that would be criminal for a natural person, or even an artificial person like a corporation. We argue that criminal law falls short in cases where an AI causes certain types of harm and there are no practically or legally identifiable upstream criminal actors. This Article explores potential solutions to this problem, focusing on holding AI directly criminally liable where it is acting autonomously and irreducibly. Conventional wisdom holds that punishing AI is incongruous with basic criminal law principles such as the capacity for culpability and the requirement for a guilty mind.

Drawing on analogies to corporate and strict criminal liability, as well as familiar imputation principles, we show a coherent theoretical case can be constructed for AI punishment. AI punishment could result in general deterrence and expressive benefits, and it need not run afoul of negative limitations such as punishing in excess of culpability. Ultimately, however, punishing AI is not justified, because it might entail significant costs and it would certainly require radical legal changes. Modest changes to existing criminal laws that target persons, together with potentially expanded civil liability, are a better solution to AI crime.
I. INTRODUCTION ................................................................................................................................. 3

II. ARTIFICIAL INTELLIGENCE AND PUNISHMENT ................................................................................. 5
   A. INTRODUCTION TO ARTIFICIAL INTELLIGENCE ................................................................. 5
   B. A FRAMEWORK FOR UNDERSTANDING AI CRIME ............................................................. 8
   C. A MAINSTREAM THEORY OF PUNISHMENT .............................................................................. 11

III. THE AFFIRMATIVE CASE ................................................................................................................ 16
   A. CONSEQUENTIALIST BENEFITS ............................................................................................... 16
   B. EXPRESSIVE CONSIDERATIONS ............................................................................................... 17

IV. RETRIBUTIVE LIMITATIONS ......................................................................................................... 19
   A. THE ELIGIBILITY CHALLENGE ................................................................................................. 19
   B. OTHER RETRIBUTIVIST CHALLENGES: REDUCIBILITY AND SPILLOVER ......................... 28

V. FEASIBLE ALTERNATIVES .............................................................................................................. 30
   A. FIRST ALTERNATIVE: THE STATUS QUO ............................................................................... 31
   B. THE COSTS OF PUNISHING AI ............................................................................................... 34
   C. SECOND ALTERNATIVE: MINIMALLY EXTENDING THE CRIMINAL LAW ......................... 37
   D. THIRD ALTERNATIVE: MODERATE CHANGES TO CIVIL LIABILITY .............................. 40
   E. CONCLUDING THOUGHTS .......................................................................................................... 41
I. INTRODUCTION

In 2015, an artist going by the moniker “Random Darknet Shopper” (RDS) purchased Ecstasy and a Hungarian passport for display in an art exhibit. This was part of a performance project where RDS was given $100 in the cryptocurrency bitcoin each week to make a purchase from an online marketplace. The items were then shipped to a Swiss art gallery and put on exhibit. After learning about the exhibit from social media, Swiss police took RDS into custody along with the purchases.

What makes this story interesting for our purposes is that RDS was an artificial intelligence (AI), and hardly the first to have a run in with law enforcement. If RDS had been a person located within the U.S., it could be criminally prosecuted under U.S. law. For that matter, entities involved in this activity other than RDS might also be criminally prosecuted, such as those supplying the bitcoin and hosting the exhibition. Luckily for RDS and crew, the Swiss authorities were art fans.

Cases like this will pose new challenges. The RDS case may be relatively straightforward, but programs exist that are autonomous, decentralized, and “unoppable”. What if RDS had been open source software that individuals from around the world independently helped to program? What if RDS was instead “Random Shopper”, designed to purchase necessities for college dorms while relying on machine learning to improve? What if it had been initially programmed to only purchase items from Amazon, but learned from online user content that some necessities could be purchased at lower cost from other websites, and that a broader understanding of “necessities” exists? What if Random Shopper existed only on a distributed ledger, a decentralized technology that supports smart contracts and cryptocurrencies, so that it could not easily be deactivated? What if it earned its own funds from financial trading?

These scenarios suggest that criminal law may soon be confronted with some hard questions. For example, if Random Shopper autonomously bought Ecstasy, in a manner not foreseeable to its developers, should those individuals be criminally liable? For that matter, who


4 Id.

5 See Matt Novak, Was This The First Robot Ever Arrested?, GIZMODO (Feb. 18, 2014), https://paleofuture.gizmodo.com/was-this-the-first-robot-ever-arrested-1524686968 (describing police confiscation in 1982 of a robot: “The police considered citing [its owner] for failing to obtain a permit for advertising…but no charges were filed and the robot was ultimately returned.”). Robot encounters with law enforcement are becoming more common. See, e.g., Peter Dockrill, A Robot Was Just ‘Arrested’ by Russian Police, SCIENCE ALERT (Sept. 20, 2016), https://www.sciencealert.com/a-robot-was-just-arrested-by-russian-police.


7 18 U.S.C. § 2(a) (making it a crime to aid and abet offenses).

8 Random Darknet Shopper was eventually returned to its creators together with all of the purchases except the Ecstasy. Kharpal, supra note 1 (noting that prosecutor’s comment “that the possession of Ecstasy was indeed a reasonable means for the purpose of sparking public debate about questions related to the exhibition”). Apparently, the Hungarian passport was also returned. Id.

9 See Section II A, infra (discussing the DAO).

should count as its developers, and which ones would be liable? Should its owners be liable, and what if it has no owners? Should its users be liable, and what if it has no users? Perhaps Random Shopper itself should be held criminally liable.

The possibility of directly criminally punishing AI is receiving increased attention by the popular press and legal scholars alike. Perhaps the best-known defender of punishing AI is Gabriel Hallevy. He contends that “[w]hen an AI entity establishes all elements of a specific offense, both external and internal, there is no reason to prevent imposition of criminal liability upon it for that offense.” In his view, “[i]f all of its specific requirements are met, criminal liability may be imposed upon any entity—human, corporate or AI entity.” Drawing on the analogy to corporations, Hallevy asserts that “AI entities are taking larger and larger parts in human activities, as do corporations,” and he concludes that “there is no substantive legal difference between the idea of criminal liability imposed on corporations and on AI entities.”

In contrast to punishment expansionists like Hallevy, skeptics might be inclined to write off the idea of punishing AI from the start as conceptual confusion—akin to hitting one’s computer when it crashes. If AI is just a machine, then surely the fundamental categories of the criminal law like criminal culpability—a “guilty mind” that is characterized by insufficient regard for legally protected values—would be inconsistent with the idea of punishing AI. One might think the whole idea of punishing AI can be easily dispensed with as inconsistent with basic criminal law principles.

The idea of punishing AI is due for fresh consideration. This Article takes a measured look at the proposal, informed by theory and practice alike. We argue punishment of AI cannot be categorically ruled out. Harm caused by a sophisticated AI may be more than a mere accident where no wrongdoing is implicated. Some AI-generated harms may stem from difficult to reduce behaviors of an autonomous system, whose actions resemble those of other subjects of the criminal law, especially corporations. Corporations can directly face criminal charges when their defective procedures generate condemning harms—particularly in scenarios where structural problems in corporate systems and processes are difficult to reduce to the wrongful actions of

12 Id. at 199.
13 Id. at 200 (asking why AI entities should be treated “different from corporations”).
14 Id. at 201
15 Id. at 199.
16 Id. at 199.
18 Model Penal Code § 2.07 (outlining conditions under which a corporation could be convicted of an offense).
individuals. It is necessary to do the difficult pragmatic work of thinking through the theoretical costs and benefits of AI punishment, how it could be implemented in practice, and to consider the alternatives.

Our inquiry focuses on the strongest case for punishing AI: scenarios where crimes are functionally committed by machines and there is no identifiable person who has acted with criminal culpability. We call these Hard AI Crimes. This can occur when no person has acted with criminal culpability, or when it is not practically defensible to reduce an AI’s behavior to bad actors. There could be general deterrent and expressive benefits from imposing criminal liability on AI in such scenarios. Moreover, the most important negative, retributivist-style limitations that apply to persons need not prohibit AI punishment. On the other hand, there may be costs associated with AI punishment: conceptual confusion, expressive costs, spillover, and rights creep. In the end, our conclusion is this: While a coherent theoretical case can be constructed for AI punishment, it is not ultimately justified in light of the less disruptive alternatives that can provide substantially the same benefits.

This Article proceeds as follows. Part II provides a brief background to AI and “AI crime.” It then provides a framework for justifying punishment that considers affirmative benefits, negative limitations, and feasible alternatives. Part III considers potential benefits to AI punishment, and argues it could provide general deterrence and expressive benefits. Part IV examines whether punishment of AI would violate any of the negative limitations on punishment that relate to desert, fairness and the capacity for culpability. It finds that the most important constraints on punishment, such as requiring a capacity for culpability for it to be appropriately imposed, would not be violated by AI punishment.

Finally, Part V considers feasible alternatives to AI punishment. It argues the status quo is likely inadequate for properly addressing AI crime. While direct AI punishment is a solution, this would require problematic changes to the criminal law. Alternately, AI crime could be addressed through modest changes to criminal laws applied to individuals together with potentially expanded civil liability. We argue that civil liability is generally preferable to criminal liability for AI activity as it is proportionate to the scope of the current problem and a less significant departure from existing practice with fewer costs. In this way, the Article aims to map out the possible responses to the problem of harmful AI activity and makes the case for approaching AI punishment with extreme caution.

II. ARTIFICIAL INTELLIGENCE AND PUNISHMENT

A. Introduction to Artificial Intelligence

---

19 See infra notes 159–161 and accompanying text (discussing ways to defend the irreducibility of corporate culpability); William Laufer, Corporate Bodies and Guilty Minds, 43 EMORY L. J. 647, 664–68 (1994) (outlining prevalent models of “genuine corporate culpability” including proactive fault, reactive fault, corporate ethos and corporate policy).

20 See Part IV.
We use the term “AI” to refer to a machine that acts intelligently.\textsuperscript{21} AI only sometimes has the ability to directly act physically, as in the case of a “robot,” but it is not necessary for an AI to directly affect physical activity to cause harm (as the RDS case demonstrates).

AI is rapidly improving, driven by advances in software, computing power, and big data.\textsuperscript{22} Hardly a day goes by without a new report of some impressive feat achieved by AI. In 2017, Alphabet’s flagship DeepMind AI beat the world champion of the board game Go.\textsuperscript{23} This was considered an important feat in the AI community, because of the sheer complexity of the game.\textsuperscript{24} There are more possible Go board configurations than there are atoms in the universe.\textsuperscript{25} Thus, a machine designed to play Go cannot simply be preprogrammed with optimal predetermined moves, or solely rely on a brute force approach to considering a large number of future moves.\textsuperscript{26} Go was the last traditional board game at which people had been able to outperform machines.\textsuperscript{27}

In some areas, AI already makes significant practical contributions. For instance, Google translate supports more than 100 languages, including 37 by photo, 32 by voice, and 27 in “augmented reality mode.”\textsuperscript{28} The increasing prevalence and capability of AI will lead to widespread social benefit, but will also cause harm. Virtually all activity involves a risk of harm, and as AI comes to do more it will inevitably cause more harm.\textsuperscript{29}

A few of features of AI are important to highlight. First, AI has the potential to act unpredictably.\textsuperscript{30} Some leading AIs rely on machine learning or similar technologies which involve a computer program, initially created by individuals, further developing in response to data without explicit programming.\textsuperscript{31} This is one means by which AI can engage in unforeseeable activities its original programmers may not have intended.\textsuperscript{32}


\textsuperscript{23} Id at 25.
\textsuperscript{24} Id.
\textsuperscript{25} Id.
\textsuperscript{26} Id.
\textsuperscript{27} Id.
\textsuperscript{31} Abbott, supra note 22 at 25.
\textsuperscript{32} There has been a recent focus on biased decisions by machine learning algorithms—sometimes due to a programmer’s implicit bias, sometimes due to biased training data. See, e.g., Chris DeBrusk, \textit{The Risk of Machine-Learning Bias (and How to Prevent It)}, MIT SLOAN MANAGEMENT REVIEW (2018), https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/.
Second, AI has the potential to act unexplainably. It may be possible to determine what an AI has done, but not how or why it acted as it did. This has led to some AIs being described as “black box” systems. For instance, an algorithm may refuse a credit application but not be able to articulate why the application was rejected. That is particularly likely in the case of AIs that learn from data, and which may have been exposed to millions or billions of data points. Even if it is theoretically possible to explain an AI outcome, it may be impracticable given the potentially resource intensive nature of such inquiries, and the need to maintain earlier versions of AI and specific data.

Third, AI may act autonomously. For our purposes, that is to say an AI may cause harm without being directly controlled by an individual. Suppose an individual creates an AI to steal financial information by mimicking a bank’s website, stealing user information, and posting that information online. While the theft may be entirely reducible to an individual who is using the AI as a tool, the AI may continue to act without further human involvement. It may even be the case that the individual who sets an AI in motion is not able to regain control of the AI, which could be by design.

Fourth, while AI can already outperform people in spectacular fashion in some domains, like board games, in other domains AI is not competitive with toddlers. That is because all AI is designed to perform “narrow” or “specific” tasks. DeepMind can beat the world’s best human player at Go, but it could not translate English to French without being programmed to do so. By contrast, the holy grail of computer science research is developing “general” AI that could perform any task that a person could perform. Experts are divided on whether, and when, general AI will be developed. For now, the weight of expert opinion holds there will, probably, be no general AI for at least a couple of decades.

Of course, it is possible for a conventional machine to perform unpredictably, unexplainably, or autonomously. However, at a minimum, AI is far more likely to exhibit these characteristics, and to exhibit them to a greater extent. Even a sufficient difference in degree along several axes makes AI worth considering as a distinctive phenomenon, possibly meriting novel legal responses.

---

33 Abbott, supra note 21 (providing background on this definition).
34 Id.
35 Id.
36 Id.
37 “The DAO” was the most famous attempt to create a decentralized autonomous organization. See Samuel Falkon, The Story of the DAO—Its History and Consequences, THE STARTUP (Dec. 24, 2017), https://medium.com/swlh/the-story-of-the-dao-its-history-and-consequences-71e56a551ee. The concept was to deploy an entity could no longer be controlled by its creators and would act without further direction. The DAO would operate according to smart contract, or pre-programmed rules dictating future behavior. A DAO might be used to create an entity operating according to publicly available, unalterable code on a distributed ledger to prevent corporate mismanagement. Unfortunately, the DAO failed shortly after launch on Ethereum due to programming flaws and hacker interference. Id.
38 Abbott, supra note 22 at 23–28.
39 Id.
40 Id.
41 Id.
Finally, general AI, and even super- or ultra-intelligent AI, is different than the sort of self-aware, conscious, sentient AIs that are common in science fiction. The latter sorts of AIs, sometimes referred to as “strong AI,” are portrayed as having human-like abilities to cognitively reason and to be morally culpable for their actions. Today, even the prospect of such machines is safely within the realm of science fiction. We will not consider punishment of strong AI.

B. A framework for understanding AI crime

We use the term “AI crime” as a loose shorthand for cases in which an AI would be criminally liable if a natural person acted similarly. Machines have caused harm since ancient times, and robots have caused fatalities since at least the 1970s. However, besides machines being intentionally used to inflict harm (as when I run someone over with my car), most harms caused by machines are seen as mere accidents. The exception is when the culpable carelessness of people using a machine caused the harm (as when negligence in using drilling machinery caused the BP oil spill). Such harms are not mere accidents; rather, they are accidents that implicate the criminal law. Nonetheless, even in such cases, the criminal law is not usually deployed as against the harmful machines themselves (at least outside of some intriguing but archaic prosecutions of inanimate objects). It may be that AI crimes are no different than any other harm for which a machine is involved.

Yet AI can differ from conventional machines in a couple of essential respects that makes the direct application of the criminal law more worthy of consideration. Specifically, AI can behave in ways that display high degrees of autonomy and irreducibility. In terms of autonomy, AI is capable of behaving largely independently of human control. AI can receive sensory input, set

---

43 Abbott, supra note 22 at 23–28 (describing super- and ultra-intelligent AI).
44 Jesus Rodriguez, Consciousness and the Weak vs. Strong AI Debate, TOWARDS DATA SCIENCE (August 23, 2018), https://towardsdatascience.com/g%C3%B6del-consciousness-and-the-weak-vs-strong-ai-debate-51e71a9189ca
45 Id.
46 Id.
47 If and when such machines come into existence, we will certainly enjoy reading their works on AI criminal liability.
51 Oliver Wendell Holmes, The Common Law, 24 (during Edward I’s reign “[i]f a man fell from a tree, the tree was deodand”— forfeited as an “accursed thing” and given to God); Blackstone, Commentaries on the Law of England: In Four Books, ch. 8, The King’s Revenue, p. 302; Albert Alschuler, Ancient Law and the Punishment of Corporations: Of Frankpledge and Deodand, 71 Boston U. L. Rev. 307, 312 (1991) (“Just as primitive people hated and punished the wheel of a cart that had run someone over...some of us truly manage to hate the corporate entity.”).
52 We will not attempt to articulate the non-functional differences between human and algorithmic reasoning, a subject which has fascinated and confounded computer scientists since the 1950s. A.M. Turing, Computing Machinery and Intelligence, 59 MIND 433, 433–51 (1950). Functionally, AI and people can exhibit similar patterns of behavior and information processing, regardless of whether machines “think” or understand what they do. See Chalmers, infra note 118 (distinguishing intellectual capacities from phenomenal consciousness).
targets, assess outcomes against criteria, make decisions and adjust behavior to increase its likelihood of success—all without being directed by human orders.\textsuperscript{52} Reducibility is also critical, because if an AI engages in an act that would be criminal for a person and the act is reducible, then there typically will be a person that could be criminally liable.\textsuperscript{53} If an AI act is not effectively reducible, there may be no other party that is aptly punished, in which case intuitively criminal activity could occur without the possibility of punishment.

Many, likely most, AI crimes are likely to be reducible. For instance, if an individual develops an AI to hack into a self-driving car to disable vital safety features, that individual has directly committed a crime.\textsuperscript{54} If someone strikes another person with a rock, the rock has not committed battery—the individual throwing the rock has. Even where AI behaves autonomously, to the extent that a person uses AI as a tool to commit a crime, and the AI functions foreseeably, the crime involves an identifiable defendant causing harm. Even when AI causes unforeseeable harm this may still be reducible, for example, if an individual creates an AI to steal financial information, but a programming error results in the AI shutting down an electrical grid thus disrupting hospital care. This is a familiar problem for criminal law.\textsuperscript{55} If someone commits a robbery and in so doing causes injuries to bystanders in unforeseeable ways (imagine a tripped bank alarm startles the animals in a neighboring zoo and they break loose and trample pedestrians), the criminal law has doctrinal tools by which liability could still be imposed.\textsuperscript{56}

Some of the time, however, it may be difficult to reduce AI crime to an individual due to AI autonomy, complexity, or unexplainability. A large number of individuals may contribute to development of an AI over a long period of time. For instance, with some open source software, thousands of people can collaborate informally to create an AI.\textsuperscript{57} In the case of AI that develops in response to training with data, it may be difficult to attribute responsibility for an AI output where the machine has learned how to behave based on accessing millions or billions of data points from heterogenous sources.\textsuperscript{58} Thus, it may be more difficult to assign fault to individuals where AI is concerned versus a conventional product such as a car where one component is faulty. In fact, it may be \textit{practically} impossible to reduce an AI generated harm to the actions of individuals.

Even where AI developers are known, an AI might end up causing harm without any unreasonable human behavior. Suppose two experienced and expert programmers separately contribute code for the software of an autonomous vehicle, but the two contributions unforeseeably interact in ways that cause the vehicle to deliberately collide with individuals

\footnotesize{\textsuperscript{52} See supra notes 22-37.}  
\footnotesize{\textsuperscript{53} See Part V.A.}  
\footnotesize{\textsuperscript{54} Jeff Gurney, \textit{Driving into the Unknown: Examining the Crossroads of Criminal Law and Autonomous Vehicles}, \textit{5 Wake Forrest J. L. \\ \\ Policy} 393, 433 (2015) (discussing crimes applicable to this scenario).}  
\footnotesize{\textsuperscript{55} See infra Part V.}  
\footnotesize{\textsuperscript{56} See discussion below on constructive liability offenses (Part V.A).}  
\footnotesize{\textsuperscript{57} In 2017, for instance, more than 4,500 Microsoft employees contributed to open source software hosted on GitHub. Matt Asay, \textit{Who really contributes to open source}, \textit{Infoworld} (Feb. 7, 2018), https://www.infoworld.com/article/3253948/open-source-tools/who-really-contributes-to-open-source.html#tk.twt_ifw. GitHub is a development platform that hosts open source code. Frederic Lardinois, et al., \textit{Microsoft has acquired GitHub for $7.5B in stock}, \textit{TechCrunch} (June 2018), https://techcrunch.com/2018/06/04/microsoft-has-acquired-ghibli-for-7-5b-in-microsoft-stock/.}  
\footnotesize{\textsuperscript{58} See Lothar Determann \\ \\ & Bruce Perens, \textit{Open Cars}, 32 \textit{Berkeley Tech. L.J.} 915 (2017).}
wearing striped shirts. If this was the result of some not-reasonably-foreseeable interactions between the two programmers’ contributions, then presumably neither programmer would have criminal liability. Generally, to be criminally liable, an individual has to intend a certain prohibited socially undesirable outcome—or at least act recklessly, which is acting despite being aware of a substantial and unjustified risk that one’s conduct may produce a prohibited outcome.\(^\text{59}\) Sometimes, although more controversially,\(^\text{60}\) criminal liability can be imposed on a negligence basis when one causes harm that a reasonable person would have foreseen and taken precautions to avoid.\(^\text{61}\) At least in a case where AI activity has, from the perspective of a reasonable person, unforeseeably caused harm, individuals would not generally face criminal liability, as this would not even meet the threshold for criminal negligence. In some cases, they would not even be civilly liable if their actions were not negligent under the tort standard.\(^\text{62}\)

There are several possible grounds on which the criminal law might deem AI crime to be irreducible.\(^\text{63}\)

1) *Enforcement Problems:* A bad actor is responsible for an AI crime, but the individual cannot be identified by law enforcement. For example, this might be the case where the creator of a computer virus has managed to remain anonymous.

2) *Practical Irreducibility:* It would be impractical for legal institutions to seek to reduce the harmful AI conduct to individual human actions, because of the number of people involved, the difficulty in determining how they contributed to the AI’s design, or because they were active far away or long ago. Criminal law inquiries do not extend indefinitely for a variety of sound reasons.\(^\text{64}\)

3) *Legal Irreducibility:* Even if the law could reduce the AI crime to a set of individual human actions, it may be bad criminal law policy to do so. For example, unjustified risks might not be substantial enough to warrant being criminalized. Perhaps multiple individuals acted carelessly in insubstantial ways, but their acts synergistically led to AI causing significant harm. In such cases, the law might deem the AI’s conduct to be irreducible for reasons of criminalization policy.

We will largely set aside enforcement-based reasons for irreducibility as less interesting from a legal design perspective. Enforcement problems exist without AI. Other forms of irreducibility may exist, such as moral irreducibility, but we will not focus on these here because they are controversial and undertheorized.\(^\text{65}\)

---

\(^{59}\) Model Penal Code, 2.02(2)(a)-(c) (defining purpose and recklessness).

\(^{60}\) Larry Alexander & Kimberly Fezan, Crime and Culpability (2009) (arguing there should be no criminal liability for negligence).

\(^{61}\) Model Penal Code, 2.02(2)(d) (defining negligence).

\(^{62}\) Developers may be civilly liable other than under tortious negligence. For instance, if it were a defective commercial “product,” its supplier might be subject to strict products liability. Abbott, supra note 47 at 13–16 (discussing product liability law). However, such liability only applies in limited situations. *Id.*

\(^{63}\) See Part IV.A.2.

\(^{64}\) See Section IV.B.1 (discussing the reducibility challenge as applied to corporate liability).

\(^{65}\) It’s conceivable the law might adopt a “moral irreducibility” view. That is, law might deem an AI (perhaps incorrectly) to be a full-blown moral agent, i.e., genuinely responsible for its conduct, and for that reason the law might regard its conduct as irreducible. However, while this might be logically possible, we have doubts about it—especially if sufficient creativity is used to identify bad human behavior nearby. *See infra* Part IV.B.1.
Instead, our analysis will focus on what we take to be less controversial forms of irreducibility: those where it is not practically feasible to reduce the harmful AI conduct to human actors or where the harmful AI conduct was just the result of human misconduct too trivial to penalize. In these instances, AI can be seen as autonomously committing crimes in irreducible ways, where there is no responsible person. This is what we refer to as “Hard AI Crime” and what we take to provide the strongest case for holding AI criminally liable in its own right.

C. A Mainstream Theory of Punishment

To anchor our analysis, this section introduces a theory of punishment that reflects the broad consensus in the literature.66 We use the term “punishment” roughly as defined by HLA Hart in terms of five elements:

(i) It must involve pain or other consequences normally considered unpleasant.
(ii) It must be for an offense against legal rules.
(iii) It must be of an actual or supposed offender for his offence.
(iv) It must be intentionally administered by human beings other than the offender.
(v) It must be imposed and administered by an authority constituted by a legal system against which the offence is committed.67

Thus, “punishment” requires a conviction for a legally recognized offense following accepted procedures.68 Imprisonment, fines or asset forfeiture carried out in response to a proper conviction for a specified offense would count as punishment, but a range of other activities that most people might consider “punishment” in a loose sense would not.69 For instance, harsh treatment by private citizens for violating informal social norms, preventative detention of the insane on grounds of their being a danger to themselves or others, or asset forfeiture carried out in advance of a conviction would not count as punishment.70 Civil penalties, while violations of legal norms, do not count as an “offense” for criminal law purposes, as the criminal law seeks to condemn egregious categories of conduct.71

67 HLA Hart, PUNISHMENT AND RESPONSIBILITY, 2nd ed. 4-5 (2008).
68 This is likewise supported by the principle of legality, built into any well-functioning legal system. This principle (nulla poena sin lege) provides that it is not legally permitted to penalize someone for an action without a law prohibiting that conduct. Douglas N. Husak and Craig A. Callender, Wilful Ignorance, Knowledge, and the “Equal Culpability” Thesis: A Study of the Deeper Significance of the Principle of Legality, 1994 WIS. L. REV. 29 (1994).
69 Id. at 5.
70 It remains open, on this definition, whether mere arrest is itself a form of punishment. It is properly carried out only in response to a suspected criminal offense, although it is in advance of a conviction and therefore it is unclear whether it satisfies the “imposed for an offense” requirement in (i). As Hart notes, these may be punishment-like in some respects, but do not fall within the core concept of punishment. Hart, supra note 67 at 5.
71 To deem some conduct an offense is to condemn it, to mark it out as culpable and to label the one who commits it an offender. See Duff, infra note 96. The expression of condemnation and declaring someone convicted of a crime to be an offender who is guilty is a feature of core punishment, but not civil liability. Because Hart’s definition is couched in terms like “offense” and “offender,” carrying as they do connotations of culpability and condemnation, civil liability would not qualify as punishment.
Punishment is justified only if its affirmative justifications outweigh its costs and it does not otherwise offend applicable negative limitations. Affirmative justifications are the positive benefits that punishment might produce like harm reduction, increased safety, enhanced wellbeing, or expressing a commitment to core moral or political values. Such benefits can give reason to criminalize certain types of conduct and impose sanctions on actors who perform acts of those types. Affirmative justifications are distinct from negative limitations on punishment. These are commonly associated with culpability-focused retributivist views of the criminal law. For example, it is widely held to be unjust to punish the innocent—or punish wrongdoers in excess of what they deserve in virtue of their culpability—even if this would promote aggregate wellbeing in society. This so-called “desert constraint” imposes a limitation, grounded in justice, on promoting social welfare through punishment.

1) Affirmative Reasons to Punish

It is common to be a pluralist about the reasons for punishment. U.S. Federal law refers to the most widely acknowledged benefits, including the need to “afford deterrence to criminal conduct,” to “protect the public from further crimes of the defendant,” to “provide the defendant with” rehabilitative treatment of various kinds, as well as to reflect “the seriousness of the offense” which covers the culpability of the act and the desert of the actor.

For simplicity, the affirmative aims of punishment can be grouped into two main categories: a) consequentialist aims and b) retributivist aims. Some theorists also mention c) expressive aims, though these largely reduce to other aims in the first two categories. Consequentialist benefits cover the good consequences that punishment can bring about, usually understood as enhancing the aggregate well-being of the members of society by reducing harm. The main type of consequentialist benefit of punishment is preventive, in that punishment can reduce crime.

Punishment can reduce crime several ways. The simplest is incapacitation: when the offender is locked up, he or she is physically limited from committing further crimes while incarcerated. The next and arguably most important way punishment prevents harm is through deterrence—namely by threatening negative consequences for the commission of a crime that give would-be offenders reasons to refrain from prohibited conduct. Deterrence comes in two forms: i) specific deterrence and ii) general deterrence. Specific deterrence is the process whereby punishing a specific individual discourages that person from committing more crime in

---

72 See infra notes 90-92.
73 Id.
74 Michael T. Cahill, Punishment Pluralism, in RETRIBUTIVISM: ESSAYS ON THEORY AND POLICY, Mark D. White ed., (2011) (“many have proposed a hybrid model of ‘limiting retributivism’ that explicitly purports to combine aspects of both the canonical theories” of consequentialism and retributivism; suggesting that “the ascendant view of punishment is more openly pluralistic about its purposes”); Berman, infra note 66 at 141-42 (noting “conver[gence] on a desert-constrained pluralism” about the justifications of punishment; describing it as “something approaching a consensus” view).
75 18 U.S.C. § 3553.
76 Berman, infra note 66 at 148 (discussing whether expressivism reduces to consequentialism).
78 Id.
the future. General deterrence occurs when punishing an offender discourages other would-be offenders from committing crimes. It is a matter of punishing an offender in order to “send a message” to other potential offenders. There can be affirmative benefits to punishing the insane because it may deter sane individuals from committing crimes and attempting to rely on an insanity defense.

These are not the only kinds of consequentialist benefits that can support punishment. Besides incapacitation and deterrence, punishment can reduce harm through 3) rehabilitation of the offender. Insofar as punishment helps the offender to see the error of his or her ways, or training or skills are provided during incarceration, this, too, can help prevent future crimes.

Besides crime prevention, there also may be non-consequentialist benefits that can provide additional affirmative grounds for punishment. Most importantly, it may be intrinsically valuable to give culpable actors what they deserve in a way that does not just reduce to the value of harm reduction. In other words, the idea is that retribution, giving offenders what they are due in virtue of the culpability of what they did, is intrinsically valuable or fitting. Retribution matters, for example, because it allows society to sufficiently distance itself from the offender’s wrongdoing and prevents it from being complicit (or overly tolerant) of culpable wrongdoing.

While virtually everyone agrees that the good consequences of preventing crime must be a major part of what justifies punishment, there is more debate about whether retributivist reasons also exist. While retributivist reasons for punishment are worth taking seriously, here we assume that the lion’s share of the affirmative case in favor of punishment will involve harm reduction and similar desirable consequences.

One last group of affirmative reasons that merit mention are expressive reasons. Punishment involves the communication of society’s collective commitment to certain core reasons.

79 Berman, infra note 66 at 145 (discussing types of deterrence).
80 Id.
81 Hart offered this response to Bentham’s argument that because children and the insane are not deterrable, they should not be punished. Hart argues more soberly that “though…the threat of punishment could not have operated on [children or the insane], the actual infliction of punishment on those persons, may secure a higher measure of conformity to law on the part of normal persons.” Hart, supra note 67 at 19. While there are other reasons for not punishing children and the insane (i.e. reduced culpability), Bentham’s “undeterrability” argument is not a convincing reason.
82 Berman, infra note 66 (discussing rehabilitation).
83 To illustrate, suppose punishing a murderer will do absolutely nothing to prevent future crime or reduce harm to others. Maybe the offense was committed decades ago and the offender is now too infirm to reoffend. Suppose the punishment is guaranteed to remain a complete secret from the public. Punishment would thus not result in specific or general deterrence, but there may still be a retributive reason to punish. The reason would stem from the value (if any) inherent in giving offenders what they are due in virtue of their culpability.
84 John Cottingham, Varieties of Retribution, PHILOSOPHICAL QUARTERLY 29, 1979; Victor Tadros, THE END OF HARM 26 (2011) (identifying retributivism with the claim that it is “intrinsically valuable” that offenders suffer in proportion to their desert); Doug Husak, Retributivism in Extremis, 32 L. & Philosophy 3 (2013) (defending broader versions of retributivism).
86 Tadros, supra note 84 (ch. 2).
87 See e.g., id. ch. 4.
values. The state, through punishment, conveys official condemnation of culpable conduct through the mechanism of a criminal conviction. This can benefit victims psychologically to see the state reaffirm their rights which were violated by a criminal act. Officially expressing condemnation of culpable conduct may also affect behavior and attitudes in general by reinforcing positive social values.\(^9\)

Some question whether expressive benefits are a distinct category of reason to punish, or whether they simply reduce to consequentialist or retributivist reasons.\(^9\) After all, many of the benefits in the expressive category center around harm prevention, such as the deterrent effects of signaling that society will not stand for seriously culpable conduct. Expressive reasons might also reduce also to retributivist reasons insofar as the value of expressing condemnation is that it involves giving offenders their due. In what follows, we assume expressive benefits reduce to consequentialist or retributive reasons, but not much turns on it. Our arguments are also compatible with the contrary view.

2) Negative (Retributive) Limitations

Punishment also should not violate deeply held normative commitments such as justice or fairness. The most important of these limitations focus on the culpability of those subject to the criminal law. One such limitation on punishment is the desert constraint, which figures into most retributivist views.\(^9\) The desert constraint is the claim that an offender may not, in justice, be punished in excess of his or her desert. Desert, in turn, is understood mainly in terms of the culpability one incurs in virtue of one’s conduct. The main effect of the desert constraint is to rule out punishments that go beyond what is proportionate to one’s culpability.\(^9\) Thus, it would be wrong to execute someone for jaywalking even if doing so would ultimately save lives by reducing illegal and dangerous pedestrian crossings.

What supports the desert constraint? Intuition, for one thing. It seems unjust to punish someone who is completely innocent even if it would produce significant benefits through general deterrence.\(^9\) Similarly, it seems unjust to impose a very severe punishment on someone who only committed a minor crime. Beyond its intuitive plausibility, the desert constraint is also supported by the argument—tracing back at least to Kant—that it is wrong to use people merely as a means to one’s ends without also treating them as ends in themselves (i.e., without respecting their value as persons).\(^9\) Punishing the innocent to obtain broader social benefits is a paradigmatic example of treating people merely as means. It fails to show individuals the respect they are due. Under some Kantian views, the desert constraint is absolute. It is not appropriate to treat someone merely as a means to one’s ends regardless of the costs of respecting their value as

---

\(^9\) See Robinson, infra note 112.
\(^9\) Berman, infra note 66 at 148.
\(^9\) See id. at 144 (on retributivism, punishment is justified if, but only to the extent that, “it is deserved or otherwise fitting, right or appropriate, and not [necessarily because of] any good consequences” it may have”); see also id. at 151 (discussing desert-constrained consequentialism).
\(^9\) Negative retributivism is the view that the desert of the offender only prohibits punishing in excess of desert (even if it has good consequences). Positive retributivism says that the offender’s desert provides an affirmative reason for punishment.
\(^9\) Cf. Hart, supra note 67 at 11 (discussing the famous example of the small southern town).
\(^9\) See, e.g., Tadros, supra note 86, ch. 6 (defending a version of the means principle).
persons. Others have a more nuanced view, such that violating a negative limitation could be overall justified if the benefits were sufficiently weighty.\footnote{Hart, supra note 67 at 12 (“In extreme cases many might think it right to resort to these expedients but we should do so with the sense of sacrificing an important principle.”).}

There are limitations on punishment other than the desert constraint. Most importantly, the criminal law requires certain prerequisites, such as capacity for culpability, that defendants must meet in order to be properly subject to punishment. It is a fundamental aim of the criminal law to condemn culpable wrongdoing, and it is the default position in criminal law doctrine that punishment may only be properly imposed in response to culpable wrongdoing.\footnote{See MODEL PENAL CODE, §1.02(c) (1962) (declaring that one of the “general purposes” of the Code is “to safeguard conduct that is without fault from condemnation as criminal”); Nicola Lacey and Hanna Pickard, To Blame or to Forgive? Reconciling Punishment and Forgiveness in Criminal Justice, 35 OXFORD J. LEGAL STUDIES 665 (2015), https://doi.org/10.1093/ojls/gqv012 (observing that on retributivist theories, “punishment is only justified if the condition of responsible agency is met”); ANTONY DUFF, THE REALM OF THE CRIMINAL LAW 19-20 (2019) (“censure…is essential to a criminal conviction”; a legal system “that criminalizes conduct that is not even alleged to be or portrayed as being wrongful is, necessarily, a perversion of criminal law”).} Without the requisite capacities of deliberation and agency, an entity is not an appropriate subject for criminal punishment—as can be seen from the fact that lacking such capacities altogether can give rise to an incapacity defense.\footnote{MPC § 4.01 (1962) (outlining incapacity defense based on mental defect causing him to be unable “either to appreciate the criminality…of his conduct or to conform [it to] the law”).} Thus, capacity for culpability is an eligibility requirement for being aptly subject to regulation by the criminal law.

3) Alternatives to Punishment

For punishment to be justified, it is not enough for it to have affirmative benefits and to be consistent with the negative limitations for punishment. In addition, there cannot be better, feasible alternatives, including doing nothing. This is an obvious point that is built into policy analysis of all kinds.\footnote{Sven Hansson, Philosophical Problems in Cost-Benefit Analysis, 23 ECON, & PHILOSOPHY 163, 164(2007) (“In cost-benefit analysis, an alternative is not evaluated by itself but in comparison to other alternatives (or, at least, in comparison to not choosing that alternative).”).}

Even if punishing AI has affirmative benefits, and even if the practice did not seriously violate any negative limitations, it still would not be justified if, for example, civil liability, licensure, or industry standards provide a better solution. It is often claimed that when seeking to exert social control, criminal law should be a tool of last resort.\footnote{Doug Husak, The Criminal Law as Last Resort, 24 OXFORD J. LEGAL STUD. 207 (2004).} After all, criminal law sanctions are the harshest form of penalty society has available, involving as they do both the possible revocation of personal freedom as well as the official condemnation of the offender. Thus, the third requirement for a given punishment to be justified is the absence of better alternatives.

4) Putting the Pieces Together

Determining whether a given punishment is appropriate requires investigation of three questions:
a) **Affirmative Benefits:** Are there sufficiently strong affirmative reasons in favor of punishment? This chiefly concerns consequentialist benefits of harm reduction but may also include retributive and expressive benefits.

b) **Culpability-Focused Negative Limitations:** Would punishment be consistent with applicable negative limitations? This primarily concerns culpability-focused principles like the desert constraint as well as basic prerequisites of apt criminal punishment such as capacity for culpability.

c) **Feasible Alternatives:** Is punishment a better response to the harms or wrongs in question than alternatives like civil liability or regulation or doing nothing?

In what follows, we will apply this theory to investigate whether the direct punishment of AI is justified. We will begin in Part III with the question of Affirmative Benefits, consider Negative Limitations in Part IV, then Feasible Alternatives in Part V.

### III. THE AFFIRMATIVE CASE

This Part considers the affirmative benefits that might be adduced to support punishing AI. The discussion focuses primarily on consequentialist benefits. Even if retribution can also count in favor of punishment, we assume that such benefits would be less important than consequentialist considerations centering on harm reduction. This Part does not aim to completely canvass the benefits of punishing AI. Instead, it argues that punishing AI could produce at least some significant affirmative benefits.

#### A. Consequentialist Benefits

Recall that, arguably, the paramount aim of punishment is to reduce harmful criminal activity through deterrence. Thus, a preliminary objection to punishing AI is that it will not produce any affirmative harm-reduction benefits because AI is not deterrable. Peter Asaro argues that “deterrence only makes sense when moral agents are capable of recognizing the similarity of their potential choices and actions to those of other moral agents who have been punished for the wrong choices and actions—without this…recognition of similarity between and among moral agents, punishment cannot possibly result in deterrence.”

The idea is that if AIs, given current designs, are not able to detect and respond to criminal law sanctions in a way that renders them deterrable, there would be nothing to affirmatively support punishing AI. It is likely true that AI, as currently operated and envisioned, will not be responsive to punishment, although responsive AI is theoretically possible.

---

100 Tadros, *supra* note 84.


102 AIs could in principle be programmed to follow court orders or adapt their behavior in response to a conviction. This may be a less effective way to ensure AI lawfulness, however, than programming the AI ex ante not to run afoul of criminal law. This will be more challenging with criminal laws that are *standards* rather than simple rules. It is comparatively easy to program a self-driving car not to run a red light than it is to not run a red light *except* in unspecified emergency conditions.
The answer to the undeterrability argument requires distinguishing specific deterrence from general deterrence. Specific deterrence involves incentivizing a particular defendant not to commit crimes in the future. By contrast, general deterrence involves incentivizing other actors besides the defendant from committing crimes. We must further distinguish two types of general deterrence: deterring others from committing offenses of the same type the defendant was convicted of, offense-relative general deterrence, and deterring others from committing crimes in general, unrestricted general deterrence.

The crucial point, then, is that punishing AI could provide general deterrence. Presumably, it will not produce offense-relative general deterrence to other AIs, as such systems are not designed to be sensitive to criminal law prohibitions and sanctions. Nonetheless, AI punishment could produce unrestricted general deterrence. That is to say, direct punishment of AI could provide unrestricted general deterrence as against the developers, owners, or users of AI and provide incentives for them to avoid creating AIs that cause especially egregious types of harm without excuse or justification. Depending on the penalty associated with punishment, such as destruction of an AI, what Mark Lemley and Brian Casey have termed the “robot death penalty,” punishing AI directly could deprive such developers, owners or users of the financial benefits of the systems they would otherwise obtain, thus incentivizing them to modify their behavior in socially desirable ways. The deterrence effect may be stronger if capitalization requirements are associated with some forms of AI in the future, or if penalties associated with punishment are passed on to, for example, an AI’s owner.

B. Expressive Considerations

Punishment of AI may also have expressive benefits. Expressing condemnation of the harms suffered by the victims of an AI could provide these victims with a sense of satisfaction and vindication. Christina Mulligan has defended the idea that punishing robots can generate victim-satisfaction benefits, arguing that, “taking revenge against wrongdoing robots, specifically, may be necessary to create psychological satisfaction in those whom robots harm.” On her view, “robot punishment—or more precisely, revenge against robots—primarily advances…the creation of psychological satisfaction in robots’ victims.” Punishment conveys a message of official condemnation that could reaffirm the interests, rights and ultimately the value of the victims of the harmful AI. This, in turn, could produce an increased sense of security among victims and society in general.

103 Hart, supra note 67 at 19.
104 See supra notes 79-80.
106 Analogous considerations could apply to provide support for punishing inanimate objects and corporations.
107 Christina Mulligan, Revenge Against Robots, 69 S. Carolina. L. Rev. 579, 580 (2018); cf. David Lewis, The Punishment that Leaves Something to Chance, 18 PHILO. & PUB. AFF. 53, 54 (1989) (discussing but rejecting the idea of defending puzzling features of the criminal law on the ground that when harm results the population may “demand blood”).
108 Mulligan, supra note 107 at 593
109 Duff, supra note 88; Guyora Binder, Victims and the Significance of Causing Harm, 28 PACE L. REV. 713, 733 (2008) (“We punish not only in order to admonish the offender…but also…to show the victim our own respect. If so, we are punishing harm for a purpose that transcends doing justice to the offender.”); Jack Boeglin and Zachary Shapiro, A Theory of Differential Punishment, 70 VANDERBILT L. REV. 1499 (2017) (arguing that victims’ interests should be taken “into account when determining how severely criminal offenders should be punished”).
This sort of expressivist argument in favor of punishing AI may seem especially forceful in light of empirical work demonstrating the human tendency to anthropomorphize and attribute mentality to artificial persons like corporations.\textsuperscript{110} The same sorts of tendencies are likely to be even more powerful for AI-enabled robots that are specifically designed to seem human enough to elicit emotional responses from humans.\textsuperscript{111} In the corporate context, some theorists argue that corporations should be punished because the law should reflect lay perceptions of praise and blame, “folk morality,” or else risk losing its perceived legitimacy.\textsuperscript{112} This sort of argument, if it succeeds for corporate punishment, is likely to be even more forceful as applied to punishing AI, which often are deliberately designed to piggy-back on the innate tendency to anthropomorphize.\textsuperscript{113} Were the law to fail to express condemnation of robot-generated harms despite robots being widely perceived as blameworthy (even if this is ultimately a mistaken perception), this could erode the perception of the legitimacy of the criminal law. Thus, a number of benefits could be obtained through the expressive function of punishment.\textsuperscript{114}

Nonetheless, there are a range of prima facie worries about appealing to expressive benefits like victim satisfaction in order to justify the punishment of AI. Punishing AI to placate those who want retaliation for AI-generated harms would be akin to giving in to mob justice. Legitimizing such reactions could enable populist calls for justice to be pressed more forcefully in the future. The mere fact that punishing AI might be popular would not show the practice to be just. As David Lewis observed, if it is unjust for the population to “demand blood” in response to seeing harm, then satisfying such demands through the law would itself be unjust—even if “it might be prudent to ignore justice and do their bidding.”\textsuperscript{115} Simply put, the popularity of a practice does not automatically justify it, even if popularity could be relevant to its normative justification. Moreover, punishing AI for expressivist purposes could lead to further

\textsuperscript{110} Mihailis Diamantis, \textit{Corporate Criminal Minds}, 91 NOTRE DAME L. REV. 2049, 2078 (2016) (arguing that “When groups exhibit high levels of coherence, as do most corporations, humans perceive them as possessing many of the attributes traditionally associated with individuals,” thus rendering “‘blame and punishment [of] these groups…psychologically sensible and sustainable’”); id. at 2077-79 (collecting psychology sources).


\textsuperscript{112} Diamantis, \textit{supra} note 110 at 2088-89 (a “criminal legal system that is more responsive to society’s perceptions of blameworthiness may foster forces, like respect for and confidence in the law, that ultimately increase compliance by individuals. Conversely, ignoring lay perceptions of blameworthiness...threatens to undermine the broader effectiveness of the criminal law in preventing crime”); see generally PAUL ROBINSON, INTUITIONS OF JUSTICE AND THE UTILITY OF DESERT 176-88 (2013).


\textsuperscript{114} Some might worry that expressive benefits just are consequentialist reasons to punish AI. While conceptually interesting, not much of practical importance turns on this issue for our purposes.

\textsuperscript{115} Lewis, \textit{supra} note 107 at 229.
bad behavior that might spill over to the way other humans are treated. Thus, Kate Darling has argued robots should be protected from cruelty in order to reflect moral norms and prevent undesirable human behavior.

Further, expressing certain messages through punishment may also carry affirmative costs which should not be omitted from the calculus. Punishing AI could send the message that AI is itself an actor on par with a human being, which is responsible and can be held accountable through the criminal justice system. Such a message is concerning, as it could entrench the view that AI has rights to certain kinds of benefits, protections and dignities that could restrict valuable human activities.

In sum, punishing AI may have affirmative benefits. It could result in general deterrence for developers, owners and users, as well as produce expressive benefits (if also potential costs). Whether these benefits would provide sufficient justification for punishing AI when compared to the feasible alternatives will be discussed in Part V. Before that, we turn to another kind of threshold question: whether punishing AI violates the culpability-focused negative limitations on punishment.

IV. RETRIBUTIVE LIMITATIONS

This Part considers retributivist (culpability-focused) limitations on punishment. Section A asks whether AI can be seen as being the right kind of entity to be eligible for punishment—what we call The Eligibility Challenge. Where the criminal law’s fundamental prerequisites are not satisfied, its sanctions are not be legitimately deployed. Section B considers two further retributivist objections to the punishment of AI. The Reductionist Challenge insists that any apparent AI culpability is fully reducible to the actions of persons who are better targets for punishment. This challenge purports to show that there is no need for the direct punishment of AI. Finally, the Spillover Objection insists it would be unjust to punish AI if this would predictably harm innocent people who develop, own, or use such systems.

We argue that all three of these objections to the punishment of AI admit of answers. Along the way, we draw on the literature on corporate punishment and strict liability, where analogs of these same objections have been explored.

A. The Eligibility Challenge

---

116 This is similar to Kant’s point that although he thought animals are not strictly speaking moral persons, there are still good reasons to discourage the mistreatment of animals because it could embolden people to mistreat other human beings. Lori Gruen, The Moral Status of Animals, in STANFORD ENCYCLOPEDIA OF PHILOSOPHY (2017) https://plato.stanford.edu/entries/moral-animal/ (discussing Kant’s view of ethical treatment of animals; quoting Kant’s Lectures in Ethics as stating that if one unfairly harms a dog “he does not fail in his duty to the dog…but…[he] must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men” ([1784-5] 1997: 212 [Ak 27: 459])).

The Eligibility Challenge is simple to state: AI, like other inanimate objects, is not the right kind of thing to be punished. It lacks mental states and the deliberative capacities needed for culpability, so it cannot be punished without sacrificing core commitments of the criminal law. The issue is not that AI punishment would be unfair to AI. AIs are not conscious, do not feel (at least in the phenomenal sense), and do not possess interests or wellbeing. Therefore, there is no reason to think AI gets the benefit of the protections of the desert constraint, which prohibits punishment in excess of what culpability merits. The Eligibility Challenge does not derive from the desert constraint.

Instead, the Eligibility Challenge, properly construed, comes in one narrow and one broad form. The narrow version is that, as a mere machine, AI lacks mental states and thus cannot fulfill the mental state (mens rea) elements built into most criminal offenses. Therefore, convicting AI of crimes requiring a mens rea like intent, knowledge or recklessness would violate the principle of legality. This principle stems from general rule of law values and holds that it would be contrary to law to convict a defendant of a crime unless it is proved (following applicable procedures and by the operative evidentiary standard) that the defendant satisfied all the elements of the crime. If punishing AI violates the principle of legality, it threatens the rule of law and could weaken the public trust in the criminal law.

The broad form of the challenge holds that because AI lacks the capacity to deliberate and weigh reasons, AI cannot possess broad culpability of the sort the criminal law aims to respond to. A fundamental purpose of the criminal law is to condemn culpable wrongdoing, as it is at least the default position in criminal law doctrine that punishment may be properly imposed only in response to culpable wrongdoing. The capacity for culpable conduct thus is a general prerequisite of the criminal law, and failing to meet it would remove the entity in question from the ambit of proper punishment—a fact that is encoded in law, for example, in incapacity defenses like infancy and insanity. Thus, the broad version of the Eligibility Challenge holds that because AI lacks the practical reasoning capacities needed for being culpable, AI does not fall within the scope of the criminal law. Punishing AI despite its lack of capacity would not only be conceptually confused, but would fail to serve the retributive aims of the criminal law—namely, to mark out seriously culpable conduct for the strictest public condemnation.

Here we develop three answers to the Eligibility Challenge.

---

119 Id. (discussing the hard problem of consciousness).
120 Supra notes 90-92.
121 See Husak, supra note 68.
122 See Doug Husak, ‘Broad’ Culpability and the Retributivist Dream, 9 OHIO ST. J. CRIM. L. 449 (2012) (distinguishing narrow culpability as merely mens rea categories from broad culpability, which is the underlying normative defect that the criminal law aims to respond to).
123 See supra note 96; see also MICHAEL MOORE, PLACING BLAME 35 (1997) (arguing for a presumption in favor of punishing “all and only those who are morally culpable in the doing of some morally wrongful action”); Duff, supra note 96 (a legal system “that criminalizes conduct that is not even alleged to be…wrongful is, necessarily, a perversion of criminal law”). While strict liability crimes exist, these are only justified in exceptional circumstances and are otherwise unjust. W. Robert Thomas, On Strict Liability Crimes: Preserving a Moral Framework for Criminal Intent in an Intent-Free Moral World, 110 MICH. L. REV., 647, 647-50 (2012).
1. Answer 1: Respondeat Superior

The simplest answer to the Eligibility Challenge has been deployed with respect to corporations. Corporations are artificial entities that might also be thought ineligible for punishment because they are incapable of being culpable in their own right. However, even if corporations cannot literally satisfy mens rea elements, the criminal law has developed doctrines that allow culpable mental states to be imputed to corporations. The most important such doctrinal tool is respondeat superior, which allows mental states possessed by an agent of the corporation to be imputed to the corporation itself provided that the agent was acting within the scope of her employment and in furtherance of corporate interests. Some jurisdictions also tack on further requirements. Since imputation principles of this kind are well-understood and legally accepted, thus letting actors guide their behavior accordingly, respondeat superior makes it possible for corporations to be convicted of crimes without violating the principle of legality.

If this kind of legal construction of mental states is a promising mechanism by which corporations can be brought back within the ambit of proper punishment and avoid the Eligibility Challenge, the same legal device could be used to make AI eligible for punishment. The culpable mental states of AI developers, owners, or users could be imputed to the AI under certain circumstances pursuant to a respondeat superior theory.

It may be more difficult to use respondeat superior to answer the Eligibility Challenge for AI than for corporations—at least in cases of Hard AI Crime. Unlike a corporation, which is literally composed of the humans acting on its behalf, an AI is not guaranteed to come with a ready supply of identifiable human actors whose mental states can be imputed. This is not to say there will not also be many garden-variety cases where an AI does have a clear group of human developers. Most AI applications are likely to fall within this category and so respondeat superior would at least be a partial route to making AI eligible for punishment. Of course, in many of these cases when there are identifiable people whose mental states could be imputed to the AI—such as developers or owners who intended the AI to cause harm—criminal law will

---

124 See e.g. Albert Alschuler, Two Ways to Think About the Punishment of Corporations, 46 AM. CRIM. L. REV. 1359 (2009) (arguing against corporate punishment).
125 Ashley S. Kircher, Corporate Criminal Liability versus Corporate Securities Fraud Liability: Analyzing the Divergence in Standards of Culpability, 46 AM. CRIM. L. REV. 157, 157 (2009) ("respondeat superior has been the most traditionally accepted method of imputing criminal liability to a corporation"); Eli Lederman, Models for Imposing Corporate Criminal Liability: From Adaptation and Imitation Toward Aggregation and the Search for Self-Identity, 4 BUFF. CRIM. L. REV. 641, 654-55 (2000) (under respondeat superior, “a corporation is liable for the deeds of any of its agents or employees...as long as...[t]he agent was acting within the course and scope of his or her employment, having the authority to act for the corporation...at least in part in furtherance of the corporation’s business interests” (internal alterations and quotation marks omitted)).
126 MODEL PENAL CODE § 2.07(1)(c) (adopting respondeat superior but restricting it to the mental states of high corporate officials).
127 Granted, this is a legal fiction. But the principle of legality does not obviously require that corporations literally—as opposed to legally—satisfy the mens rea element. Paul Robinson, Imputed Criminal Liability, 93 YALE L. J. 609 (1984). Even if one thinks imputation principles are in tension with the principle of legality, strictly construed, the costs we normally fear from violating it—like weakening public trust in the criminal law—are not likely to be very serious. So even if corporations’ literal lack of mental states were to remain a formalistic problem for corporate punishment, it would be a very weighty one.
128 See Hallevy, supra note 12 at 201 (arguing “there is no substantive legal difference between the idea of criminal liability imposed on corporations and on AI entities”).
already have tools at its disposal to impose liability on these culpable human actors. In these cases, there is less likely to be a need to impose direct AI criminal liability.

Thus, while respondeat superior can help mitigate the Eligibility Challenge for AI punishment in many cases, this unlikely to be an adequate response in cases of Hard AI Crime.

2. Answer 2: Strict Liability

A different sort of response to the Eligibility Challenge is to look for ways to punish AI despite its not literally possessing culpable mental states. That is not to simply reach for a consequentialist justification of the conceptual confusion or inaptness involved in applying criminal law to AI. Within the criminal law, we take this to be a justificatory strategy of last resort—especially given the blunt form of consequentialism it relies on. Rather, what is needed is a method of cautiously extending the criminal law to AI that would not entail weighty violations of the principle of legality.

One way to do this would be to establish a range of new strict liability offenses specifically for AI crimes—i.e., offenses that an AI could commit even in the absence of any mens rea like intent to cause harm, knowledge of an inculpatory fact, reckless disregard of a risk or negligent unawareness of a risk. In this sense, the AI would be subject to liability without “fault.” This would permit punishment of AI in the absence of mental states. Accordingly, strict liability offenses may be one familiar route by which to impose criminal liability on an AI without sacrificing the principle of legality.

Many legal scholars are highly critical of strict liability offenses. As Duff argues, strict criminal liability amounts to unjustly punishing the innocent:

That is why we should object so strongly (...): the reason is not (only) that people are then subjected to the prospect of material burdens that they had no fair opportunity to avoid, but that they are unjustly portrayed and censured as wrongdoers, or that their conduct is unjustly portrayed and condemned as wrong.129

Yet this normative objection applies with greatest force to persons. The same injustice does not threaten strict criminal liability offenses for AI because AI does not obviously enjoy the protections of the desert constraint130 (which prohibits punishment in excess of culpability).131

This strategy is not without problems. Even to be guilty of a strict liability offense, defendants still must satisfy the voluntary act requirement.132 LaFave’s criminal law treatise observes that “a voluntary act is an absolute requirement for criminal liability.”133

---

129 Duff, supra note 96 at 19.
130 There may be unfairness to adjacent innocent people who own or rely on the AI, but that is a separate problem which afflicts any punishment. See Part IV.B.
131 Matters would be different if AIs, perhaps like many animals, could experience pleasure and pain, or were conscious or otherwise in possession morally salient interests. It would indeed seem unfair to subject animals to extreme suffering just for general deterrent benefits (if not as unfair as for a human being).
132 WAYNE LAFAVE, 1 SUBST. CRIM. L. § 6.1(c) (3d ed.) (“criminal liability requires that the activity in question be voluntary”).
133 Id.
Penal Code, for example, holds that a “person is not guilty of an offense unless his liability is based on conduct that includes a voluntary act or the omission to perform an act of which he is physically capable.” 134 Behaviors like reflexes, convulsions or movements that occur unconsciously or while sleeping are expressly ruled out as non-voluntary. 135 To be a voluntary act, “only bodily movements guided by conscious mental representations count”. 136 If AI cannot have mental states and is incapable of deliberation and reasoning, it is not clear how any of its behavior can be deemed to be a voluntary act.

There are ways around this problem. The voluntary act requirement might be altered (or outright eliminated) by statute for the proposed class of strict liability offenses that only AI can commit. Less dramatically, even within existing criminal codes, it is possible to define certain absolute duties of non-harmfulness that AI defendants would have to comply with or else be guilty by omission of a strict liability offense. The Model Penal Code states that an offense cannot be based on an omission to act unless the omission is expressly recognized by statute or “a duty to perform the omitted act is otherwise imposed by law.” 137 A statutory amendment imposing affirmative duties on AI to avoid certain kinds of harmful conduct is all it would take to enable an AI to be strictly liable on an omission theory.

Of course, this may also carry costs. Given that one central aim of the criminal law is usually taken to be responding to and condemning culpable conduct, if AI is punished on a strict liability basis, this might risk diluting the public meaning and value of the criminal law. 138 That is, it threatens to undermine the expressive benefits that supposedly help justify punishing AI in the first place. 139 This is another potential cost to punishing AI that must be weighed against its benefits.

3. Answer 3: A Framework for Direct Mens Rea Analysis for AI

The last answer is the most speculative. A framework for directly defining mens rea terms for AI—analogous to those possessed by natural persons—could be crafted. This could require an investigation of AI behavior at the programming level and offer a set of rules that courts could apply to determine when an AI possessed a particular mens rea—like intent, knowledge or recklessness—or at the very least, when such a mens rea could be legally constructed. 140 This inquiry could draw on expert testimony about the details of the AI’s code, though it need not. By way of analogy, juries assess mental states of human defendants by using common knowledge about what mental states (intentions, knowledge, etc.) it takes to make a person behave in the

134 Model Penal Code, § 2.01(1).
135 Id. § 2.01(2).
137 MPC, § 2.01(3).
138 See Duff, supra notes 88 and 96 at 19-20.
139 See supra, Part III.
140 In IV.A, we discussed respondeat superior as a mode of taking an existing human mental state and carrying it over to an AI. This section, by contrast, discusses possible methods of legally constructing AI mental states that no person already possesses. Cf. infra note 156 (discussing the collective knowledge doctrine for corporations).
observed fashion. Similarly in AI cases, experts might need only to testify in broad terms about how the relevant type of AI (say, a neural network) functions and how its information-processing architecture could have generated the observed behavior. Thus, direct mens rea analysis for AI could, but need not, require “looking under the hood” at the details of the code. Instead, it would be enough to simply guide the legal determination of what mens rea the AI can be deemed to possess.

Towards this end a framework is needed to steer decisionmakers in conducting direct mens rea analysis for AI, and it must consist of two parts. First, to answer the broad Eligibility Challenge, we need a general conception of what it would mean for AI to be culpable in its own right. Second, to answer the narrow version of the challenge, we need to offer a set of rules for when an AI may be deemed to possess a given mens rea.

To begin with, a coherent concept of AI culpability could be legally constructed in the following way. The prevailing theory has it that one is criminally culpable for an action to the extent that it manifests insufficient regard for legally protected interests or values. These protected interests and values provide legally recognized reasons bearing on how to behave. Insufficient regard is a form of ill will or indifference that produces mistakes in the way one recognizes, weighs and responds to the applicable legal reasons for action. The criminal law typically does not demand that we are motivated by respect for others, or even respect for law; all it demands is that we do not put our disrespect on display by acting in ways that are inconsistent with attaching proper weight to protected interests and values. Thus, criminal culpability can be seen as being more about what one’s behavior manifests and less about the nuances of one’s private motivations, thoughts and feelings. There are good institutional-design reasons—such as clarity, the need for the law to be able to guide the conduct of normal citizens, and the demand for the law not to intrude too heavily into the private sphere—for the criminal law not to be overly concerned with the specific motives or private mental states.

---

141 Peter Carruthers, Mindreading in Infants, 28 Mind & Language 141 (2013) (discussing how infants attribute beliefs and intentions to others); David Premack & Guy Woodruff, Does the chimpanzee have a theory of mind?, 1 Behavioral and Brain Sciences, 515 (1978) (defining “theory of mind” as the system whereby “the individual imputes mental states to himself and to others” and noting that it is “not directly observable [but] can be used to make predictions… about the behavior of other organisms”).

142 LARRY ALEXANDER AND KIMBERLY FERZAN, CRIME AND CULPABILITY 67-68 (2009) (“insufficient concern [is] the essence of culpability”). Peter Westen, An Attitudinal Theory of Excuse, 25 LAW & PHILOSOPHY 289, 373-74 (2006) (“a person is normatively blameworthy for engaging in conduct that a statute prohibits if he was motivated by an attitude of disrespect for the interests that the statute seeks to protect”). See Yaffe, supra note 143; Gideon Yaffe, Intoxication, Recklessness, and Negligence, 9 Ohio St. J. CRIM. L. 545, 552-53 (2012); VICTOR TADROS, CRIMINAL RESPONSIBILITY 250 (2005) (“if [a defendant] is convicted of a serious offence, the state communicates…that [his] behaviour manifested an inappropriate regard for other citizens and their interests”).

143 See, e.g., GIDEON YAFFE, ATTEMPTS 38 (2011) (an action is culpable to the degree that “it is a product of a faulty mode of recognition or response to reasons for action”). Note that legal culpability may or may not be the same as moral blameworthiness. Compare Mark Dsouza 2015, p. 453 (distinguishing moral from legal culpability) and Sarch supra note 17 with Michael Moore, Choice, character, and excuse, 7 Social Philosophy and Policy 29, 30-31 (1990) (taking moral and legal culpability to be presumptively the same). We are agnostic on how to understand moral blameworthiness, which may be more fine-grained and searching of one’s inner mental states than legal culpability. Compare Pete Graham, A Sketch of A Theory of Blameworthiness, 88 Philosophy & Public Affairs, 388, 403 (2014) (“[W]hat people are truly blameworthy for are the motivations from which [their] actual actions spring, rather than the actions themselves”) with ARPALY AND SCHROEDER, IN PRAISE OF DESIRE 170 (2014) (defending a notion of blameworthiness that is similar to criminal culpability as described here). Our focus here, regardless, is legal culpability.

144 See Sarch, supra note 17.
involved in law breaking. Thus, as long as one crosses the line and has no affirmative defense, we may treat the presumption that one’s illegal action manifests insufficient regard as being unrebutted—i.e., as legally conclusive.

By way of analogy, this notion of culpability can account for corporate culpability. If only the legal notion of criminal culpability is required for proper punishment, then eligibility for punishment requires being capable of behaving in ways that manifest insufficient regard for the legally recognized reasons. All that avoiding legal culpability requires is to abstain from actions that are reasonably interpreted as disrespectful forms of conduct stemming from a legally deficient appreciation of the legal reasons. This provides a recipe for how to regard corporations as being criminally culpable in their own right. They possess information-gathering, reasoning and decision-making procedures in virtue of the hierarchy of employees they are made up of. Thus, corporations can be seen as having the capacity for criminal culpability. Through their members, they weigh and act on the reasons that the criminal law demands not displaying insufficient regard for in action. Corporations can engage in conduct that puts on display their insufficient regard for the legally recognized interests of others. For example, if a corporation learns, through its employees, that its manufacturing processes generate dangerous waste that is seeping into the drinking water in the nearby town, this is a legally recognized reason for altering its conduct. If the corporation continues its activities unchanged, this demonstrates that it—through its information-sharing and decision-making procedures—did not end up attaching sufficient weight to the legally recognized reasons against continuing its dangerous manufacturing activities. This is paradigmatic criminal culpability.

AI could qualify as criminally culpable in an analogous manner. Sophisticated AI may have built-in goals with a greater or lesser autonomy to determine the means of completing those goals. AI may gather information, process it and determine the most efficient means to accomplishing its goals. Accordingly, the law might deem some AIs to possess the functional equivalent of sufficient reasoning and decision-making abilities to manifest insufficient regard. If the AI is programmed to be able to take account of the interests of humans and consider legal requirements, but ends up behaving in a way that is inconsistent with taking proper account of these legally recognized interests and reasons, then the AI can be reasonably seen as manifesting insufficient regard—which is to say, be deemed in law to be criminally culpable.

---

145 Gideon Yaffe, The Age of Culpability (2018) (developing evidentialist account of manifestation of insufficient regard); see also, Sarch supra note 17.

146 See W. Robert Thomas, The Ability and Responsibility of Corporate Law to Improve Criminal Punishment, 78 Ohio St. L.J. 601, 612-13 (2017) (“[C]orporations are able to function competently in a range of normatively laden activities. (...) They can deliberate and act consistent with their self-identified interests and separate from outside pressures. Corporations are willing participants in...our normative practices, even if they may not be objects of moral consideration in...themselves. For example, through contract law, corporations routinely participate in a normatively laden practice akin to promising.”); Christian List & Phillip Pettit, Group Agency 158-63 (2011) (arguing that corporations can have decision-making structures that satisfy the main preconditions for responsibility).

147 One might object that a corporation’s practical reasoning and decision-making capacities merely derive from, or are composed out of, those of the corporation’s members. However, this is merely a worry about reducibility, discussed below. See Part IV.B. It does not undermine corporations’ threshold eligibility for punishment. Thomas, supra note 146 (if “corporate attitudes derive from the contributions of individuals who themselves are uncontroversially moral agents...it would be surprising that every emergent corporate attitude would be stripped of normative content”).
This gives a flavor of how criminal culpability might broadly be understood for AI, but we still need a framework for determining when sophisticated AIs can be said to possess a functional analog of a standard mens rea like purpose or knowledge. We do not attempt here to formulate necessary and sufficient conditions for an AI mens rea, but rather to sketch some possible approaches.

Work in philosophy of action characterizing the functional roles of an intention in a person could be extended to AI. Consider Bratman’s well-known account. On his view, actors who intend (i.e. act with the purpose) to bring about an outcome “guide [their] conduct in the direction of causing” that outcome. This means that “[i]n the normal case, one [who intends an outcome] is prepared to make adjustments in what one is doing in response to indications of one’s success or failure in promoting” that outcome. So if the actor is driving with the intention to hit a pedestrian, should the actor detect that conditions have changed so as to require behavioral adjustments to make this outcome more likely, an actor with this intention will be disposed to make these adjustments. Moreover, an actor with this intention is disposed to monitor the circumstances to find ways to increase the likelihood of the intended outcome. Merely foreseeing the outcome, but not intending it, does not entail these same forms of guiding one’s behavior to promote the outcome (i.e., to make it more likely). This conception of intention could be applied to AI.

One conceivable way to argue that an AI (say, an autonomous vehicle) had the intention (purpose) to cause an outcome (to harm a pedestrian) would be to ask whether the AI was guiding its behavior so as to make this outcome more likely (relative to its background probability of occurring). Is the AI monitoring conditions around it to identify ways to make this outcome more likely, and is the AI then disposed to make these behavioral adjustments to make the outcome more likely (either as a goal in itself or as a means to accomplishing another goal)? If so, then the AI may plausibly said to have the purpose of causing that outcome. Carrying out this sort of inquiry will of course require extensive and technically challenging expert testimony regarding the nature of the programming—and could thus be prohibitively difficult or expensive. But it does not seem impossible in principle even if difficult questions remain.

---

149 Bratman, supra 148 at 141.
150 Id.
151 For example, suppose the autonomous vehicle is actually aiming not to harm pedestrians by hitting them, but with something that merely correlates with hitting pedestrians—such as reducing the amount of shadows objects cast on the streets (as fewer shadows increases other metrics of reliable driving, which is the car’s primary goal). Should this be construed as intentionally hitting the pedestrians, or merely hitting them knowingly? This is a familiar problem from criminal law theory and philosophy of action. (See e.g., Adam Feltz & Joshua May, The Means/Side-Effect Distinction in Moral Cognition: A Meta-Analysis, 166 COGNITION 314 (2017).) We need not resolve this difficult question here to establish our main point that it is possible to make progress on extending mens rea terms to AI. Nonetheless, by analogy, we suspect this case would be plausibly be construed as intentionally hitting the pedestrian as a means to the self-driving car’s other goals. If the AI regulates its conduct to make hitting pedestrians more likely, this is not simply a “foreseen byproduct” of the AI behavior, but something it pursues as a means to accomplishing its deeper aims. Intending harm as a means suffices for showing purpose in the criminal law. If you kill a relative merely as the means to getting your inheritance, the killing still is purposeful. Alternatively, perhaps the “intended as a means/foreseen as a side-effect” distinction should be jettisoned as unworkable.
Similar strategies may be developed for arguing that an AI possessed other mens rea like knowledge. For example, on dispositional theories, knowledge may be attributed to an actor when the actor has a sufficiently robust set of dispositions pertaining to the truth of the proposition—such as the disposition to assent to the proposition if queried, to express surprise and update one’s plans if the proposition is revealed to be false, to behave consistently with the truth of the proposition or to depend on it carrying out one’s plans. In the criminal law, knowledge is defined as practical certainty. Thus, if we extend the above dispositional theory to AI, there is an argument for saying an AI knows a fact, F, if the AI displays a sufficiently robust set of dispositions associated with the truth of F—such as the disposition to respond affirmatively if queried (in a relevant way) whether F is practically certain to be true, or the disposition to revise plans upon receiving information showing that F is not practically certain, or the disposition to behave as if F is practically certain to be true. If enough of these dispositions are proven, then the knowledge that F could be attributed. One could take a similar approach to arguing that recklessness is present as well, as this requires only awareness that a substantial risk of harm is present—i.e. knowledge that the risk has a mid-level probability of materializing (below practical certainty).

Finally, as an alternative to direct arguments for showing AI mens rea, one could develop new imputation rules for AI. For example, one might follow the model of the collective knowledge doctrine, which identifies culpable interference with the flow of information within an organization and uses this as the basis for pretending as if the organization itself “knew” the facts it prevented itself from learning. The idea as applied here would be to take culpable conduct by the AI’s developers and use this as the basis for pretending the AI possessed a culpable mens rea itself. For example, if AI developers were reckless (or negligent) in their design, testing or production, and the AI goes on to cause harm, this could provide an argument for treating the AI itself as if it were reckless (or negligent) as to the harm caused.

Although much more needs to be said for such arguments to be workable, it at least suggests that it may be possible to develop a set of legal doctrines by which courts could deem AIs to possess the mens rea elements of crimes.

---

152 Eric Schwitzgebel, Belief: Dispositionalism, STANFORD ENCYCLOPEDIA OF PHILOSOPHY, https://plato.stanford.edu/entries/belief/#1.2 (Traditional dispositional views of belief assert that for someone to believe some proposition P is for [her] to possess [relevant] behavioral dispositions pertaining to P. Often cited is the disposition to assent to utterances of P in [appropriate] circumstances. Other relevant dispositions might include the disposition to exhibit surprise should the falsity of P [become] evident, the disposition to assent to Q if shown that P implies Q, and to depend on P’s truth in [acting]. [More generally, this amounts to] being disposed to act as though P is the case.

153 MPC, 2.02(2)(b) (defining knowledge as practical certainty).

154 See Eric Schwitzgebel, In-between Believing, 51 PHIL. QUART., 76 (2001) (defending this approach to determining when to attribute beliefs to humans).

155 MPC 2.02(2)(c) (defining recklessness).

156 ALEX SARCH, CRIMINALLY IGNORANT, ch. 9 (2019) (defending the collective knowledge doctrine as an equal culpability imputation rule for corporations); United States v. Bank of New England, 821 F.2d 844, 856 (1st Cir. 1987) (embracing one version of collective knowledge doctrine).

157 Among other problems there may not be deterrence benefits to punishing autonomous vehicles that hit pedestrians due to code that could be reconstructed as embodying a culpable maxim (like “if you flip me off then I run you over”), but withholding such punishment from unexplainable machine learning code that results in the same thing. Why the later should not generate independent liability while the former would seems to be a distinction without a normative difference.
B. Other Retributivist Challenges: Reducibility and Spillover

Even assuming AI is eligible for punishment, two further culpability-focused challenges remain. The first concerns the reducibility of any putative AI culpability, while the second concerns spillover of AI punishment onto innocent people nearby. This Section offers answers to both.

1. Reducibility

One might object that there is never a genuine need to punish AI because any time an AI seems criminally culpable, this culpability can always be reduced to that of nearby human actors—such as developers, owners, and users. The law could target the relevant culpable human actors instead.

This objection has been raised against corporate punishment too. Skeptics argue that corporate culpability is always fully reducible to culpable actions of individual humans. Any time a corporation does something intuitively culpable—like causing a harmful oil spill through insufficient safety procedures—this can always be fully reduced to the culpability of the individuals involved: the person carrying out the safety checks, the designers of the safety protocols, the managers pushing employees to cut corners in search of savings. For any case offered to demonstrate the irreducibility of corporate culpability, a skeptic may creatively find additional wrongdoing by other individual actors further afield or in the past to account for the apparent corporate culpability.

This worry may not be as acute for AI as it is for corporations. AI seems able to behave in ways that are more autonomous from its developers than corporations are from their members. Corporations, after all, are simply composed out of their agents (albeit organized in particular structures). Also, AI may sometimes behave in ways that are less predictable and foreseeable than corporate conduct.

In any case, there are ways to block the reducibility worry for corporate culpability as well as AI. The simplest response is to recall that it is legal culpability we are concerned with, not moral blameworthiness. Specifically, it would be bad policy for the criminal law to always allow any putative corporate criminal culpability to always be fully reduced to individual criminal liability. To ensure that corporate criminal culpability can always be reduced to individual criminal

---


159 Consider Pettit and List’s notion of a “responsibility deficit.” Petit & List, supra note 146. Perhaps “the individuals are blamelessly ignorant [or] act under such felt pressure that they cannot be held fully responsible for their contribution to a bad outcome; they can each argue that the circumstances mitigate their personal…responsibility.” Id. at 161. If the individuals have lowered culpability, then the total culpability for the group harm might seem greater than the sum of individual culpability. Whether such responsibility deficits can really arise, however, remains debatable. (After all, when the individuals are excused, might that lower the total amount of blame to be attributed for the group harm?)

160 Id.
culpability would require criminalizing very minute bits of individual misconduct—momentary lapses of attention, the failure to perceive emerging problems that are difficult to notice, tiny bits of carelessness, mistakes in prioritizing time and resources, not being sufficiently critical of groupthink, and so on. Mature legal systems should not criminalize infinitely fine-grained forms of misconduct, but rather will focus on broader and more serious categories of directly harmful misconduct that can be straightforwardly defined, identified and prosecuted. Criminalizing all such small failures—and allowing law enforcement to investigate them—would be invasive and threatening to values like autonomy and freedom of expression and association.\(^1\) It would also increase the risk of abuse of process. Instead, we should expect “culpability deficits”\(^2\) in any well-designed system of criminal law, and this in turn creates a genuine need for corporate criminal culpability as an irreducible concept.

Similar reasoning could be employed for AI culpability. There is reason to think it would be a bad system that encouraged law enforcement and prosecutors, any time an AI causes harm, to invasively delve into the internal activities of the organizations developing the AI in search of minute individual misconduct—perhaps even the slightest negligence or failure to plan for highly unlikely exigencies. It would be a disturbingly invasive criminal justice system that creates a sufficient number of individual offenses to ensure that any potential AI culpability can always be fully reduced to individual crimes. Hence, where AI is concerned, we do not think the Reducibility Challenge—at least as applied to legal culpability—imposes a categorical bar to punishing AI.

2. **Spillover**

A final retributivist challenge to punishing AI is the “spillover problem,” again familiar from the corporate context.\(^3\) Because corporate punishments (usually in the form of fines) amount to a hit to the corporation’s bottom line, these punishments inevitably spill over onto innocent shareholders.\(^4\) This might seem to violate the desert constraint against the state harming people in excess of their desert. The same objection has been raised against punishing AI. Mulligan worries that “[o]ne could…imagine situations where the notion of separating a rogue robot from its owner [or damaging or restricting the robot in punishing it] would create a disproportionate burden on the owner, for example if a robot was unique, unusually expensive relative to the harm caused, or difficult to replace.”\(^5\) This is just a version of the spillover problem. If the AI system unforeseeably causes harm, it may seem unfair or disproportionate to its innocent owner or operator to damage the AI system in punishment.

There are familiar responses to the spillover objection for corporations. First, one might contend that spillover does not qualify as punishment because it is not imposed on a shareholder for her offense.\(^6\) Nonetheless, this definitional answer is somewhat unsatisfying, as there

---

\(^1\) Cf. Hart, *supra* note 67 (ch. 1).
\(^2\) See Pettit & List, *supra* note 159 (defending “responsibility deficits” as creating a need for irreducible corporate accountability).
\(^3\) Thomas, *supra* note 146.
\(^4\) Glanville Williams, *Criminal Law: The General Part* 863 (2d ed. 1961) (noting that “a fine imposed on the corporation is in reality aimed against shareholders who are not…responsible for the crime, i.e., is aimed against innocent persons”).
\(^6\) See *supra* note 67 at 4-5.
clearly are strong reasons for the state not to knowingly harm innocent bystanders even if it does not strictly count as punishment.

A better answer is that spillover is not a special problem for corporate or AI punishment. Most forms of punishment—including punishment of individual wrongdoers—has the potential to harm the innocent, as when a convicted person has dependent children. Spillover objections may simply expose general problems with criminal law. The fact that punishment tends to harm the innocent suggests a need to reform criminal law as well as prisons, reentry programs and similar initiatives to lessen the collateral consequences of punishment of all types. In the corporate context, some have recently responded to the spillover objection by defending reforms to corporate punishments so the “pain” they impose is more accurately distributed to the culpable actors within the company who contributed to the crime. For example, Will Thomas argues that managers found to have contributed to a crime by the corporation should have their incentive compensation clawed back to satisfy the criminal fines that were levied against the corporation in the first instance.

When it comes to AI punishment, similar thinking applies. To the extent spillover is a concern, AI punishments should be narrowly tailored. Destroying an AI, for example, would be a blunt remedy that is more likely to harm the innocent. More tailored remedies might be implemented instead, such as reprogramming the AI, or civil remedies directed at responsible persons. In such ways, the punishment of AI systems could be crafted to minimize the spillover effects. Further, spillover may be less of a concern in the case of Hard AI Crime, where there may be little nexus between AI punishment and harm to innocent individuals. Even here, spillover could be largely addressed through well-designed mechanisms like the ex-ante creation of a financially responsible party or creation of a fund to cover criminal liability as a condition of operation the AI system (akin to criminal liability insurance). We explore such implementation ideas further in the next Part. The spillover problem thus is not an absolute bar to AI punishment. It is an omnipresent problem with criminal punishment, which must be addressed for any novel mode of criminal punishment—whether for corporations or AI.

V. FEASIBLE ALTERNATIVES

We have argued that punishing AI could have benefits and that it is not ruled out by the negative limitations and retributive preconditions of punishment. But this does not yet show the punishment of AI to be justified. This requires addressing the third main question in our theory of punishment: Would the benefits of punishing AI outweigh the costs, and would it be better than alternative solutions? These solutions might involve doing nothing, or relying on civil liability and regulatory responses, perhaps together with less radical or disruptive changes to criminal laws that target individuals.

Section A focuses Hard AI Crime, and finds that existing criminal law coverage will likely fall short. Section B argues that AI punishment has significant costs that suggest alternative approaches may be preferable. In Sections C and D, we map out some alternative approaches to managing AI crime. In particular, we examine moderate expansions of the criminal law as well

---

167 Thomas, surpa note 146.
168 Id. at 80-84.
as the tools available within civil law, and we argue that they have the resources to provide preferable solutions to the problem of Hard AI Crime.

A. First Alternative: the Status Quo

In considering the alternatives to direct punishment of AI, we begin with asking whether it would be preferable to simply do nothing. This section answers that existing criminal law falls short: there is an AI criminal gap that requires a response.

1) What the AI criminal gap is not: reducible harmful conduct by AI

We begin by setting aside something that will not much concern us: cases where responsibility for harmful AI conduct is fully reducible to the culpable conduct of individual human actors. A clear example would be one where a hacker uses AI to steal funds from individual bank accounts. There is no need to punish AI in such cases, because existing criminal offenses, like fraud or computer crimes, are sufficient to respond to this type of behavior.169

Even if additional computer-related offenses must be created to adequately deter novel crimes implemented with the use of AI, the criminal law has further familiar tools at its disposal, involving individual-focused crimes, which provide other avenues of criminal liability when AI causes foreseeable harms. For example, as Hallevy observes, cases of this sort could possibly be prosecuted under an innocent “agency model” (assuming AI can sensibly be treated as meeting the preconditions of an innocent agent, even if not of a fully criminally responsible agent in its own right).170 Under the innocent agency doctrine, criminal liability attaches to a person who acts through an agent who lacks capacity—such as a child or an insane person. For instance, if an adult uses a five-year old child to deliver illegal drugs, the adult rather than the child would generally be criminally liable.171 This could be analogous to a person programming a sophisticated AI to break the law: the person has liability for intentionally causing the AI to bring about the external elements of the offense.172

This doctrine requires intent (or at least the knowledge) that the innocent agent will cause the prohibited result in question.173 This means that in cases where someone does not intend or

---

169 18 U.S.C. § 1030(a)(1)-(7) (defining offenses such as computer trespass and computer fraud); 18 U.S.C. § 1343 (wire fraud statute).
171 Sanford Kadish, Complicity, Cause and Blame: A Study in the Interpretation of Doctrine, 73 CAL. L. REV. 323, 372-73 (1985) (“Most criminal actions can readily be committed through the instrumentality of another person”).
172 One might have doubts about this model of liability, too. After all, if AI is merely a tool, one would simply prosecute the user of the AI on a direct liability model. However, if AI is to be analogized to some kind of autonomous actor, which could break the chain of causation, akin to a child perhaps, then the innocent agency model would seem more apt. In any case, we argued in Part IV that AI might plausibly count as an agent at least for legal purposes. Therefore, we think it is not ruled out at least in principle that the innocent agency model of liability could be applied to actors who cause AI to produce criminally prohibited results.
173 Peter Alldridge, The Doctrine of Innocent Agency, 2 CRIM. L. FORUM 45 (1990); 18 U.S.C. § 2(b) (“Whoever willfully causes an act to be done which [is a crime] is punishable as a principal”). This intent requirement for innocent agency is similar to complicity liability, used where the actor assists or encourages another full-fledged agent with capacity to do a crime, which also requires intent or knowledge by the accomplice that the principal actor
foresee that the AI system being used will cause harm, the innocent agency model does not provide a route to liability. In such cases, one could instead appeal to recklessness or negligence liability if AI creates a foreseeable risk of a prohibited harm. For example, if the developers or users of AI foresee a substantial and unjustified risk that an AI will cause the death of a person these human actors could be convicted of reckless homicide. If such a risk was merely reasonably foreseeable (but not foreseen), then lower forms of homicide liability would be available. Similar forms of recklessness or negligence liability could be adopted where the AI’s designers or users actually foresaw, or should have foreseen, a substantial and unjustified risk of other kinds of harms as well—such as theft or property damage.

Hallevy also discusses this form of criminal liability for AI-generated harms, calling it the “natural and probable consequences model” of liability. This is an odd label, however, since the natural and probable consequences doctrine generally applies only when the defendant is already an accomplice to—i.e., intended—the crime of another. More specifically, the “natural and probable consequences” rule provides that where A intentionally aided B’s underlying crime C1 (say theft), but then B also goes on to commit a different crime C2 (say murder), then A would be guilty of C2 as well provided C2 was reasonably foreseeable.

Despite his choice of label, Hallevy seems alive to this complication and correctly observes that there are two ways in which negligence liability could apply to AI-generated harms that are reasonably foreseeable. He writes:

the natural-probable-consequence liability model [applied] to the programmer or user differ in two different types of factual cases. The first type of case is when the programmers or users were negligent while programming or using the AI entity but had no criminal intent to commit any offense. The second type of case is when the programmers or users programmed or used the AI entity knowingly and willfully in order to commit one offense via the AI entity, but the AI entity deviated from the plan and committed some other offense, in addition to or instead of the planned offense.

In either sort of scenario, there would be a straightforward basis for applying existing criminal law doctrines to impose criminal liability on the programmers or users of an AI that causes reasonably foreseeable harms. Thus, no AI criminal gap exists here.

---

Footnotes:

174 MPC § 2.02(2)(c)-(d) (defining recklessness and negligence).
175 MPC § 210.3(a) (recklessly causing death suffices for manslaughter).
176 MPC § 210.4 (negligent homicide).
177 MPC 220.1(2) (reckless burning or exploding); 220.2(2) (risking catastrophe); 220.3 (criminal mischief).
178 Hallevy, supra note 12 at 18-19.
179 The rule holds that the aider and abettor ‘of an initial crime…is also liable for any consequent crime committed by the principal, even if he or she did not abet the second crime, as long as the consequent crime is a natural and probable consequence of the first crime.’ Baruch Weiss, What Were They Thinking? The Mental States of the Aider and Abettor and the Causer Under Federal Law, 70 FORDHAM L. REV. 1341, 1424 (2002); United States v. Barnett, 667 F.2d 835, 841 (9th Cir. 1982) (adopting natural or probable consequences doctrine).
180 Hallevy, supra note 12 at 19.
A slightly harder scenario involves reducible harms by AI that are not foreseeable, but this is still something criminal law has tools to deal with. Imagine hackers use an AI to drain a fund of currency, but this ends up unforeseeably shutting down an electrical grid which results in widespread harm. The hackers are already guilty of something—namely, the theft of currency (if they succeed) or the attempt to do so (if they failed). Therefore, our question here is whether the hackers can be convicted of any further crime in virtue of their causing harm through their AI unforeseeably taking down an electrical grid.\footnote{Compare this case to the one where some kids are illegally using fireworks in their back yard, and this causes a massive forest fire destroying many homes. Sure, they can be convicted of any offenses, if any, related to illicitly using the fireworks. But can they also be convicted of offenses related to the massive forest fire and destroyed homes?}

At first sight, it might seem that the hackers would be in the clear for the electrical grid. They could argue that they did not proximately cause those particular harms. Crimes like manslaughter or property damage carry a proximate cause requirement under which the prohibited harm must at least be a reasonably foreseeable type of consequence of the conduct that the actors intentionally carried out.\footnote{See, e.g., MPC 2.03(2) (characterizing proximate or legal causation requirement using a “scope of the risk” test).} But in this case, taking down the electrical grid and causing physical harm to human victims was assumed to be entirely unforeseeable even to a reasonable actor in the defendant’s shoes.

The criminal law has tools to deal with this kind of scenario too. This comes in the form of so-called constructive liability crimes. These are crimes that consist of a base crime which require a mens rea, but where there then is a further result element as to which no mens rea is required. Felony murder is a classic example.\footnote{WAYNE LAFAVE, 2 SUBST. CRIM. L. § 14.5 (2d ed.) (explaining felony murder as the doctrine that “one whose conduct brought about an unintended death in the commission or attempted commission of a felony was guilty of murder”).} Suppose one breaks into a home one believes to be empty in order to steal artwork. Thus, one commits the base crime of burglary.\footnote{MPC 221.1 (defining burglary).} However, suppose further that the home turns out not to be empty, and the burglar startles the homeowner who has a heart attack and dies. This could make the burglar guilty of felony murder.\footnote{LaFave, supra note 183.} This is a constructive liability crime because the liability for murder is constructed out of the base offense (burglary) plus causing the death (even where this is unforeseeable). According to the leading theory of constructive liability crimes, they are normatively justifiable when the base crime in question (burglary) typically carries at least the risk of the same general type of harm as the constructive liability element at issue (death).\footnote{Andrew Simester, CRIMINAL LAW, chapter 15.5 (manuscript on file with authors) (arguing that constructive liability as to a result is justified when it is “intrinsic to the culpability of the defendant”). To the extent one has normative qualms about the inclusion of such strict liability elements, one could mitigate this worry by requiring the mens rea of negligence as to the further harm element—though that would prevent this kind of crime from being of any use when the further harm is unforeseeable, as it is stipulated to be in the cases in question here.}

This tool, if extended to the AI case, provides a familiar way to hold the hackers criminally liable for unforeseeably taking down the electrical grid and causing physical harm to human victims.

It may be beneficial to create a new constructive liability crime that takes a criminal act like the attempt to steal currency using AI as the base offense, and then taking the further harm to the

---

\footnote{Compare this case to the one where some kids are illegally using fireworks in their back yard, and this causes a massive forest fire destroying many homes. Sure, they can be convicted of any offenses, if any, related to illicitly using the fireworks. But can they also be convicted of offenses related to the massive forest fire and destroyed homes?}

\footnote{See, e.g., MPC 2.03(2) (characterizing proximate or legal causation requirement using a “scope of the risk” test).}

\footnote{WAYNE LAFAVE, 2 SUBST. CRIM. L. § 14.5 (2d ed.) (explaining felony murder as the doctrine that “one whose conduct brought about an unintended death in the commission or attempted commission of a felony was guilty of murder”).}

\footnote{MPC 221.1 (defining burglary).}

\footnote{LaFave, supra note 183.}

\footnote{Andrew Simester, CRIMINAL LAW, chapter 15.5 (manuscript on file with authors) (arguing that constructive liability as to a result is justified when it is “intrinsic to the culpability of the defendant”). To the extent one has normative qualms about the inclusion of such strict liability elements, one could mitigate this worry by requiring the mens rea of negligence as to the further harm element—though that would prevent this kind of crime from being of any use when the further harm is unforeseeable, as it is stipulated to be in the cases in question here.}
electrical grid, or other property or physical harm, as the constructive liability element, which requires no mens rea (not even negligence) in order to be guilty of the more serious crime. This constructive liability offense, in a slogan, could be called *Causing Harm Through Criminal Uses of AI*.

New crimes could be adopted to the extent there are not already crimes on the books that fit this mold. Indeed, in the present example, one might think there are already some available constructive liability crimes. Perhaps felony murder insofar as attempting to steal currency may be a felony, and this conduct caused fatalities. However, this tool would be of no avail in respect to the property damage caused. This is why a new crime like *Causing Harm Through Criminal Uses of AI* would seem to be necessary. In any case, no AI criminal gap is present here because the criminal law has familiar tools available for dealing with unforeseeable harms of this kind.

2) What the AI Criminal gap is: irreducible criminal conduct by AI

Consider a case of irreducible AI crime inspired by RDS. Suppose an AI is designed to purchase class materials for incoming Harvard students, but, through being trained on data from online student discussions regarding engineering projects, the AI unforeseeably “learns” to purchase radioactive material on the darkweb and has it shipped to student housing. Suppose the programmers of this “Harvard Automated Shopper” did nothing criminal in designing the system and they had entirely lawful aims. Nonetheless, despite the reasonable care taken by the programmers—and subsequent purchasers and users of the AI (*i.e.*, Harvard)—the AI caused student deaths.

In this hypothetical, there are no upstream actors who could be held criminally liable. Innocent agency is blocked as a mode of liability because the programmers, users and developers of the AI did not have the intent or foresight that any prohibited or harmful results would ensue—as is required for innocent agency to be available. Moreover, if the risk of the AI purchasing the designer drugs was not reasonably foreseeable, then criminal negligence would also be blocked. Finally, constructive liability is not available in cases of this sort because there is no “base crime”—no underlying culpable conduct by the programmers and users of the AI—out of which their liability for the unforeseeable harms the AI causes could be constructed.

One could imagine various attempts to extend existing criminal law tools to provide criminal liability for developers or users. Most obviously, new negligence crimes could be added for developers that make it a crime to develop systems that foreseeably could produce a risk of *any* serious harm or unlawful consequence, even if a specific risk was unforeseeable. The trouble is that this does not seem to amount to individually culpable conduct, as all activities and technologies involve some risks of some harm. So this expansion of the criminal law would stifle innovation and beneficial commercial activities. If there were such a crime, all the early developers of the internet would be guilty of it.

B. The Costs of Punishing AI

---

187 See *supra* note 173.
188 For related reasons, we would reject proposals to impose strict criminal liability on developers of AI that autonomously causes harms. Strict liability crimes for designers amounts to punishing the innocent. See *supra* note 129 and accompanying text.
Earlier, we discussed some of the potential costs of AI punishment, including conceptual confusion, expressive costs, and spillover. Even aside from these, punishment of AI would entail serious practical challenges as well as substantial changes to criminal law. Begin with a practical challenge: the mens rea analysis.\footnote{See supra section IV.A.1.} For individuals, the mens rea analysis is generally how culpability is assessed. Causing a given harm with a higher mens rea like intent is usually seen as more culpable than causing the same harm with a lower mens rea like recklessness or negligence.\footnote{See, e.g., Kenneth Simons, Should the Model Penal Code's Mens Rea Provisions Be Amended?, 1 OHIO ST. J. CRIM. L. 179, 195-96 (2003) (“MPC views its four basic mental states or culpability terms as hierarchically ordered”).} But how do we make sense of the question of mens rea for AI?

Section IV considered this problem, and argued that for some AI, as for corporations, the mental state of an AI’s developer, owner, or user could be imputed under something like the respondeat superior doctrine. But for cases of Hard AI Crime that is not straightforwardly reduced to human conduct—particularly where the harm is unforeseeable to designers and there is no upstream human conduct that is seriously unreasonable to be found—nothing like respondeat superior would be appropriate. Some other approach to AI mens rea would be required.

A regime of strict liability offenses could be defined for AI crimes. However, this would require a legislative work-around so that AI are deemed capable of satisfying voluntary act, applicable to all crimes.\footnote{See supra note 132-137.} This would require major revisions to the criminal law and a great deal of concerted legislative effort. It is far from an off-the-shelf solution. Alternately, a new legal fiction of AI mens rea, vaguely analogous to human mens rea, could be developed, but this too is not currently a workable solution. This approach could require expert testimony to enable courts to consider in detail how the relevant AI functioned to assess whether it was able to consider legally relevant values and interests but did not weight them sufficiently, and whether the programming has the relevant behavioral dispositions associated with mens rea like intention or knowledge. We tentatively sketched several types of argument that courts might use to find various mental states to be present in an AI. However, much more theoretical and technical work is required here, and we do not regard this as a first best option.

Mens rea, and similar challenges related to the voluntary act requirement, are only some of the practical problems to be solved in order to make AI punishment workable. For instance, there may be enforcement problems with punishing an AI on a blockchain. Such AIs might be particularly difficult to effectively combat or deactivate.

Even assuming the practical issues are resolved, punishing AI would still require major changes to criminal law. Legal personality is necessary to charge and convict an AI of a crime, and conferring legal personhood on AIs would create a whole new mode of criminal liability, much the way that corporate criminal liability constitutes a new such mode beyond individual criminal liability.\footnote{Thomas Bernard, The Historical Development of Corporate Criminal Liability, 22 CRIMINOLOGY 3, 3-4 (1984) (describing criminal law as having “always had as its primary concern the regulation of relationships between} There are problems with implementing such a significant reform.
Over the years, there have been many proposals for extending some kind of legal personality to AI. Perhaps most famously, a 2017 report by the European Parliament called on the European Commission to create a legislative instrument to deal with “civil liability caused by robots.” It further requested the Commission to consider “a specific legal status for robots”, and “possibly applying electronic personality” as one solution to tort liability. Even in such a speculative and tentative form this proposal proved highly controversial. More than 150 AI experts subsequently sent an open letter to the European Commission warning that, “[f]rom an ethical and legal perspective, creating a legal personality for a robot is inappropriate whatever the legal status model.”

Full-fledged legal personality for AIs equivalent to that afforded to natural persons, with all the legal rights that natural persons enjoy, would clearly be inappropriate. To take a banal example, allowing AI to vote would undermine democracy, given the ease with which anyone looking to determine the outcome of an election could create AIs to vote for a particular candidate. However, legal personality comes in many flavors, even for natural persons such as children who lack many adult rights and obligations. Crucially, no artificial person enjoys all of the same rights and obligations as a natural person. The best-known class of artificial persons, companies, have long enjoyed only a limited set of rights and obligations that allows them to sue and be sued, enter contracts, incur debt, own property, and be convicted of crimes. However, they do not receive protection under constitutional provisions such as the Equal Protection clause of the Fourteenth Amendment, they cannot bear arms, run for or hold public office, marry, or enjoy other fundamental rights that are enjoyed by natural persons.

individual persons,” while for practical reasons the “legal fiction” of corporate personality—and later corporate crime—developed).

193 See, e.g., Lawrence B. Solum, Legal Personhood for Artificial Intelligences, 70 N.C. L. REV. 1231 (1992); Samir Chopra and Lawrence White, A Legal Theory for Autonomous Artificial Agents (2011) (arguing that AI could and should be given legal personality in near future); See Asaro, supra note 101 at 169 (proposing a Turing test to decide if an AI agent that caused harm is legally fit to stand trial for a criminal offense); Amanda Wurah, We Hold These Truths to Be Self-Evident, That All Robots Are Created Equal, J. OF FUTURE STUD., DOI:10.6531/JFS.2017.22(2).A61.


195 Id.


197 Indeed, even without the right to vote, AI may have been used to attempt to undermine democracy. Bots have been employed to influence election outcomes, inflate online follower counts, spread fake news, or intimidate users expressing particular opinions. Nicole Radziwill & Morgan Benton, Evaluating Quality of Chatbots and Intelligent Conversational Agents, https://arxiv.org/pdf/1704.04579.pdf. More generally, much online content is generated by AI. Id.


artificial persons with legal personality, such as maritime vessels, have even fewer rights.\footnote{201} Thus, granting legal personality to AI to allow it to be punished would not require AI to receive the rights afforded to natural persons, or even those afforded to companies. AI legal personality could consist solely of obligations.

Even so, any sort of legal personhood for AIs would be a dramatic legal change that could prove problematic. As discussed earlier, providing legal personality to AI could result in increased anthropomorphisms. People anthropomorphizing AI expect it to adhere to social norms and have higher expectations regarding AI capabilities.\footnote{202} This is problematic where such expectations are inaccurate and the AI is operating in a position of trust. Especially for vulnerable users, such anthropomorphisms could result in “cognitive and psychological damages to manipulability and reduced quality of life.”\footnote{203} These outcomes may be more likely if AI were held accountable by the state in ways normally reserved only for human members of society. Strengthening questionable anthropomorphic tendencies regarding AI could also lead to more violent or destructive behavior directed at AI, such as vandalism or attacks.\footnote{204} Further, punishing AI could also affect human well-being in less direct ways, such as by producing anxiety about one’s own status within society due to the perception that AIs are given a legal status on a par with human beings.

Finally, and perhaps most worryingly, conferring legal personality on AI may lead to rights creep. Even if AIs are given few or no rights initially when they are first granted legal personhood, they may gradually acquire rights over time. Granting legal personhood to AI may thus be an important step down a slippery slope. In a 1933 Supreme Court opinion, for instance, Justice Brandeis warned about rights creep, and that granting companies an excess of rights could allow them to dominate the State.\footnote{205} Eighty years after that decision, Justice Brandeis’ concerns were prescient in light of recent Supreme Court jurisprudence such as Citizen’s United and Hobby Lobby, which significantly expanded the rights extended to corporations.\footnote{206} Such rights, for corporations and AI, can restrict valuable human activities and freedoms.

C. Second Alternative: Minimally Extending the Criminal Law

---

\footnote{204}{See Diamantis, supra note 110 at 2088–89.}  
\footnote{205}{Louis K. Liggett Co. v. Lee, 288 U.S. 517, 549 (1933) (Brandeis, J., dissenting).}  
\footnote{206}{Citizens United v. Federal Election Commission, 558 U.S. 310 (2010) at A.1 (curtailing government’s ability to restrict political speech by companies). Citizens United held that the free speech clause of the First Amendment prohibits the government from restricting independent expenditures for communications by companies. See id. (“[t]he Court has recognized that First Amendment protection extends to corporations… The Court has thus rejected the argument that political speech of corporations or other associations should be treated differently under the First Amendment simply because such associations are not ‘natural persons.’”); Burwell v. Hobby Lobby, 573 U.S. ___ (2014) (recognizing a for-profit company’s claim to religious belief).}
There are alternatives to direct AI punishment other than doing nothing. The problem of Hard AI crime would more reasonably be addressed through minimal extensions of existing criminal law. The most obvious would be to define new crimes for individuals. Just as the Computer Fraud and Abuse Act criminalizes gaining unauthorized access or information using personal computers, an AI Abuse Act could criminalize malicious or reckless uses of AI. In addition, such an Act might criminalize the failure to responsibly design, deploy, test, train and monitor the AIs one contributed to developing. These new crimes would target individual conduct that is culpable along familiar dimensions, so they may be of limited utility with regard to Hard AI Crimes that do not reduce to culpable actors. Accordingly, a different way to expand the criminal law seems needed to address Hard AI Crime.

In cases of Hard AI Crime, a designated adjacent person could be punished who would not otherwise be directly criminally liable—what we call a Responsible Person. This could involve new forms of criminal negligence for failing to discharge statutory duties (perhaps relying on strict criminal liability) in order to make a person liable in cases of Hard AI Crime. It could be a requirement for anyone creating or using an AI to ex ante register a Responsible Person for the AI. It could be a crime to design or operate AI capable of causing harm without designating a Responsible Person. This would be akin to the offense of driving without a license. The registration system might be maintained by a federal agency. However, a registration scheme is problematic because it is difficult to distinguish between AI capable of criminal activity and AI not capable of criminal activity, especially when we are dealing with unforeseeable criminal activity. Even simple and innocuous seeming AI could end up causing serious harm. Thus, it might be necessary to designate a Responsible Person for any AI. Registration might involve substantial administrative burden and, given the increasing prevalence of AI, the costs associated with mandatory registration might outweigh any benefits.

A default rule rather than a registration system might be preferable. The Responsible Person could be the AI’s manufacturer or supplier if it is a commercial product. If it is not a commercial product, the Responsible Person could be the AI’s owner, developer if no owner exists, or user if no developer can be identified. Even non-commercial AIs are usually owned as property, although that may not always be the case, for instance, with some open-source software. Similarly, all AI has human developers, and in the event an AI ever autonomously creates another AI, responsibility for the criminal acts of an AI-created AI could reach back to the original AI’s owner. In the event an AI’s developer cannot be identified, or potentially if there are a large number of developers, again in the case of some open-source software, responsibility could attach to an AI’s user. Though, this would fail to catch the rare, perhaps only hypothetical, case of the non-commercial AI with no owner, no identifiable developer, and no user. To the

---

208 A new criminal offense—akin to driving without a license—could be imposed for cases where programmers, developers, owners or users have unreasonably failed to designate a Responsible Person for an AI.
209 The Responsible Person should also be liable for harms caused by an AI where the AI, if a natural person, would be criminally liable together with another individual. Otherwise, there is a risk that sophisticated AI developers could create machines that cause harm but rely co-conspirators to escape liability.
210 There is precedent for such a Responsible Person registration scheme. In the corporate context, executives may be required to attest to the validity of some SEC filings and held strictly liable for false statements even where they have done nothing directly negligent. If the Responsible Person is a person at a company where a company owns the AI, it would have to be an executive to avoid the problem of setting up a low-level employee as “fallguy.” The SEC for this reason requires a C-level executive to attest to certain statements on filings.
extent that a non-commercial AI owner, developer and user working together would prefer a different responsibility arrangement, they might be permitted to agree to a different ex ante selection of the Responsible Person.\footnote{It might also be likely that parties with more negotiating power would attempt to offload their liability. For instance, AI suppliers might attempt to shift liability to consumers. At least in the case of commercial products it should not be possible for suppliers to do this.} That might be more likely to occur with sophisticated parties where there is a greater risk of Hard AI Crime. The Responsible Person could even be an artificial person such as a corporation.\footnote{This raises potential concerns about corporations with minimal capital being used to avoid liability. However, this same concern exists now with human activities, where thinly capitalized corporations are exploited as a way to limit the liability of individuals. Still, there are familiar legal tools to block this sort of illicit liability avoidance. To the extent a bad actor is abusing the corporate form, courts can, for instance, pierce the corporate veil.}

It would be possible to impose criminal liability on the Responsible Persons directly in the event of Hard AI Crime. For example, if new statutory duties of supervision and care were defined regarding the AI for which the Responsible Person is answerable, criminal negligence liability could be imposed on the Responsible Person should he or she unreasonably fail to discharge those duties. Granted, this would not be punishment for the harmful conduct of the AI itself. Rather, it would be a form of direct criminal liability imposed on the Responsible Person for his or her own conduct.

More boldly, if this does not go far enough to address Hard AI Crime, criminal liability could also be imposed on the Responsible Person on a strict liability basis—particularly if the relevant punishments are only fines rather than incarceration. Generally strict liability crimes are restricted to minor infractions or regulatory offenses or “violations”;\footnote{Model Penal Code 2.05(1).} though some examples of more serious strict criminal liability can also be found (such as statutory rape in some jurisdictions).\footnote{See, e.g., N.Y. Penal Law § 130.25-50 (defining statutory rape offenses); Funari v. City of Decatur, 563 So. 2d 54, 55 (Ala. Cr. App. 1990) (holding that Alabama “statute which prohibits the selling of alcohol to minors does not contain any language requiring knowledge or intent,” and “the very purpose of the statute clearly indicates a legislative intent to impose strict liability”).} This could be defended by claiming that there is a special duty owed to society at large to provide \textit{special assurances} that certain especially serious risks will be mitigated as much as possible.\footnote{\textit{Cf.} Duff, supra note 88 at 170 (suggesting in the mala prohibitum context that “we owe it to each other not merely to ensure that we act safely, but to assure each other that we are doing so, in a social world in which we lack the personal knowledge of others that could give us that assurance”).} A Responsible Person accepting strict criminal liability could serve this function. Especially in the case of AI where user trust is critical to realizing the benefits of AI, this approach could be warranted to combat the perception that unsafe AI is being employed. Accordingly, AI could become another context in which strict criminal liability on the Responsible Person is imposed.

Yet we have serious reservations about strict liability crimes applied to persons.\footnote{Kenneth W. Simons, \textit{When is Strict Criminal Liability Just?}, 87 J. CRIM. L. \& CRIMINOLOGY, 1075 (1997) (discussing retributive views that denounce strict liability; noting that “[s]trict liability appears to be a straightforward case of punishing the blameless, an approach that might have consequential benefits but is unfair on any retrospective theory of just deserts”).} If at all, they can only be justifiably used as a last resort in exigent circumstances—as in cases of unusually dangerous activities. However, it is not obvious that the use of AI qualifies as
unusually dangerous. To the contrary, in many areas of activity it would be unreasonable not to use AI, as when safety can be improved over human actors such as may soon be the case with self-driving cars.217 Most bad human actors using AI systems to commit crimes will still be caught under existing criminal laws, and so far there have not been high-profile cases of Hard AI Crimes. As a result, we are not yet convinced that Hard AI Crime is a significant enough social problem to merit the use of strict criminal liability.

At the end of the day, a Responsible Person regime accompanied by new statutory duties, which carry criminal penalties if these duties are negligently or recklessly breached, provides an attractive approach to dealing with Hard AI Crime. While it is only a minimal expansion of the criminal law, by expressing condemnation through a criminal conviction of the Responsible Person, much of the expressive benefit from a direct conviction AI can be achieved—but without as serious a loss of public trust as the legal fictions needed to punish AI directly would create.

D. Third Alternative: Moderate Changes to Civil Liability

A further alternative to dealing with Hard AI Crime is to look to the civil law, primarily tort law, as a method of both imposing legal accountability and deterring harmful AI. Some AI crime will no doubt already result in civil liability, however, if existing civil liability falls short, new liability rules could be introduced. A civil liability approach could even be used in conjunction with expansions to criminal liability.

While it is beyond the scope of this Article to canvas gaps in civil liability for AI crime, it is worth noting that existing civil liability comes with built-in limitations. Very few laws specifically address AI-generated harms, which means civil liability must usually be established under a traditional negligence or product liability framework or under contractual liability.218 Negligence generally requires a person to act carelessly, so where this cannot be established there may be no recovery. Product liability may require both that an AI is a commercial product (e.g., this may not apply where AI is just software or the use of AI is a “service”), and that there be a defect in the product (or that its properties are falsely represented).219 In the case of complex AI, it may be difficult to prove a defect, and AI may cause harm without a “defect” in the product liability sense. For these reasons, the European Commission has created Expert Groups to determine whether new technologies necessitate a revision of the Product Liability Directive, which harmonizes product liability across the European Union, and whether even more ambitious changes are needed.220 Civil liability may also derive from contractual relationships, but this usually only applies where there is privity of contract between parties, and it may also have significant limitations.221

To the extent there is inadequate civil liability for Hard AI Crimes, the Responsible Person proposal sketched above could be repurposed so that the Responsible Person might only be civilly liable. The case against a Responsible Person could be akin to a tort action if brought by an individual or a class of plaintiffs, or a civil enforcement action if brought by a government

217 Abbott, supra note 47.
218 Id.
219 Id.
220 See, e.g., European Commission, Register of Commission expert groups and other similar entities, EUROPA (Mar. 9, 2018), http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3592
221 Abbott, supra note 217.
agency tasked with regulating AI. At trial, an AI would not be treated like a corporation, where the corporation itself is held to have done the harmful act and the law treats the company as a singular acting and “thinking” entity. Rather, the question for adjudication would be whether the Responsible Person discharged his or her duties of care in respect of the AI in a reasonable way—or else civil liability could also be imposed on a strict liability basis (a less troubling prospect than it is within the criminal law).

A Responsible Person scheme is not the only solution to inadequate civil liability for Hard AI Crimes. An insurance scheme is another approach. Owners, developers or users of AI, or just certain types of AI, could pay a tax into a fund to ensure adequate compensation for victims of Hard AI Crime. The cost of this tax would be relatively minor compared to the financial benefits of AI. This could either replace the Responsible Person solution or apply to cases whether no appropriate Responsible Person exists. An AI compensation fund could operate like the National Vaccine Injury Compensation Program (VICP). Vaccines result in widespread social benefit but are known in rare cases to cause serious problems. VICP is a no-fault alternative to traditional tort liability that compensates individuals injured by a VICP-covered vaccine. It is funded by a tax on vaccines which is paid by users. Other models for an insurance scheme exist, such as the Price Anderson Act for nuclear power.

E. Concluding Thoughts

This Article has argued that, confronted with the growing possibility of Hard AI Crime, we should not overreact and reach for the radical tool of punishing AI. Alternative approaches could provide substantially similar benefits and would avoid many of the pitfalls and difficulties involved in punishing AI. A natural alternative, we argued, involves modest expansions to the criminal law, including, most importantly, new negligence crimes centered around the improper design, operation and testing of AI applications as well as possible criminal penalties for designated parties who fail to discharge statutory duties. This could be supplemented by expanded civil liability.

We took a careful look at how a criminal law regime that punished AI might be constructed and defended. In so doing, we showed that it is all too easy to underestimate the ability of criminal law theory to accommodate substantial reforms. We explored the ways in which the criminal law can—and, where corporations are involved, already does—appeal to elaborate legal fictions to provide a basis within the defensible boundaries of criminal law theory for punishing some artificial entities. We showed what a system of punishment for AI might look like and showed how some hasty arguments against it can be answered.

222 Indeed, New Zealand has replaced tort law with a publicly funded insurance scheme to compensate victims of accidents. See, e.g., Peter H. Schuck, Tort Reform, Kiwi-Style, 27 YALE L. & POL’Y REV. 187, 187-90 (observing New Zealand “abolished the most important areas of tort law more than three decades ago” in favor of an insurance scheme that awards compensation to victims on a no-fault basis).


224 Health Resources and Services Administration, About the National Vaccine Injury Compensation Program, H.R.S.A, https://www.hrsa.gov/vaccine-compensation/about/index.html

The use of legal fictions to solve difficult conceptual questions or practical problems—such as how to conceptualize or prove particular sorts of mental elements for AI or misbehavior by its developers—gives criminal law theory impressive plasticity. Legal fictions help turn the criminal law into a pragmatic tool for solving social problems. Nonetheless, legal fictions must be used with caution, as their over-use risks eroding public trust and weakening the rule of law. Moreover, allowing legal fictions to proliferate unchecked can lead to widespread injustice either through punishing the innocent or by punishing more harshly than one’s culpability calls for. While some legal fictions can be justified, they must be used judiciously. For this reason, there is and should be an onerous burden to meet before we can be confident that a particular legal fiction—such as legal personality for AI or the invention of culpable mental states for AI—is adopted. Embracing legal fiction without meeting this justificatory burden would be tantamount to believing in science fiction.

---

Sarch, supra note 156 ch. 5 (defending certain restricted uses of particular legal fictions based on culpably preserving ignorance).