Punishing Artificial Intelligence: Legal Fiction or Science Fiction

Ryan Abbott[†*] *and Alex Sarch*[**]

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 of a guilty mind.

Drawing on analogies to corporate and strict criminal liability, as well as familiar imputation principles, we show how 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.

TABLE OF CONTENTS

* Ryan Abbott, Professor of Law and Health Sciences, University of Surrey School of Law and Adjunct Assistant Professor of Medicine, David Geffen School of Medicine at University of California, Los Angeles.

** Alex Sarch, Reader (Associate Professor) in Legal Philosophy, University of Surrey School of Law. Thanks to Antony Duff, Sandra Marshall, Mark D'Souza, and Steve Bero for their insightful comments.

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.[1] 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.[2]

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.[3] If RDS had been a natural person located in the United States, it could be criminally prosecuted under U.S. law.[4] For that matter, entities involved in this activity other than RDS could also be criminally prosecuted, such as those supplying the bitcoin and hosting the exhibition.[5] Luckily for RDS and crew, the Swiss authorities were art fans.[6]

Cases like this will pose new challenges, including for criminal law doctrine.[7] The RDS case may be relatively straightforward, but programs exist that are autonomous, decentralized, and "unstoppable."[8] What if

---

[1] Arjun Kharpal, *Robot with $100 Bitcoin Buys Drugs, Gets Arrested*, CNBC (Apr. 22, 2015, 5:09 AM), https://www.cnbc.com/2015/04/21/robot-with-100-bitcoin-buys-drugs-gets-arrested.html.

[2] *See id.*

[3] *See* Matt Novak, *Was This the First Robot Ever Arrested?*, GIZMODO (Feb. 18, 2014, 12:00 PM), 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,* SCI. ALERT (Sept. 20, 2016), https://www.sciencealert.com/a-robot-was-just-arrested-by-russian-police.

[4] *See* 21 U.S.C. § 841(a)(1) (2019) (criminalizing distribution and possession with intent to distribute a controlled substance).

[5] *See* 18 U.S.C. § 2(a) (2019) (criminalizing aiding and abetting offenses).

[6] Random Darknet Shopper was eventually returned to its creators together with all of the purchases except the Ecstasy. *See* Kharpal, *supra* note 1 (noting the 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. *See id.*

[7] *See* Christopher Markou, *We Could Soon Face a Robot Crimewave. . .The Law Needs to Be Ready*, CONVERSATION (Apr. 11, 2017, 9:36 AM), https://theconversation.com/we-could-soon-face-a-robot-crimewave-the-law-needs-to-be-ready-75276.

[8] *See infra* Part I.A (discussing The Decentralized Autonomous Organization ("The DAO")).

RDS had been open source software that individuals from around the world independently helped 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 user content that some necessities could be purchased at lower cost from other websites, and that a broader understanding of "necessities" exists? If Random Shopper autonomously buys Ecstasy in a manner not reasonably foreseeable to its developers, should those individuals be criminally liable? For that matter, who 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.[9] 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."[10] 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."[11] Drawing on the analogy to corporations,[12] 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."[13] "Modern times," he contends, "warrant modern legal

[9] *See, e.g.*, Gabriel Hallevy, *The Punishibility of Artificial Intelligence Technology*, *in* LIABILITY FOR CRIMES INVOLVING ARTIFICIAL INTELLIGENCE SYSTEMS 185-229 (2014); J.K.C. Kingston, *Artificial Intelligence and Legal Liability*, *in* RESEARCH AND DEVELOPMENT IN INTELLIGENT SYSTEMS XXXIII: INCORPORATING APPLICATIONS AND INNOVATIONS IN INTELLIGENT SYSTEMS XXIV 269 (Max Bramer & Miltos Petridis eds., 2016), https://arxiv.org/pdf/1802.07782.pdf; Christina Mulligan, *Revenge Against Robots*, 69 S.C. L. REV. 579, 580 (2018); Jeffrey Wale & David Yuratich, *Robot Law: What Happens If Intelligent Machines Commit Crimes?*, CONVERSATION (July 1, 2015, 8:06 AM), http://theconversation.com/robot-law-what-happens-if-intelligent-machines-commit-crimes-44058; *infra* Part I.A (discussing The DAO).

[10] Gabriel Hallevy, *The Criminal Liability of Artificial Intelligence Entities — From Science Fiction to Legal Social Control*, 4 AKRON INTELL. PROP. J. 171, 191 (2010).

[11] *Id.* at 199.

[12] *See id.* at 200 (asking why AI entities should be treated "different from corporations").

[13] *Id.* at 200-01.

measures."[14] More recently, Ying Hu has subjected the idea of criminal liability for AI to philosophical scrutiny and made a case "for imposing criminal liability on a type of robot that is likely to emerge in the future," insofar as they may employ morally sensitive decision-making algorithms.[15] Her arguments likewise draw heavily on the analogy to corporate criminal liability.[16]

In contrast to AI punishment expansionists like Hallevy and Hu, 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 concepts of the criminal law like culpability — a "guilty mind" that is characterized by insufficient regard for legally protected values[17] — would be misplaced. 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,

---

[14] *Id.* at 199.

[15] Ying Hu, *Robot Criminals*, 52 MICH. J.L. REFORM 487, 531 (2019); *see also id.* at 490 ("[A]n argument can be made for robot criminal liability, provided that the robot satisfies three threshold conditions . . . . [T]he robot must be (1) equipped with algorithms that can make nontrivial morally relevant decisions; (2) capable of communicating its moral decisions to humans; and (3) permitted to act on its environment without immediate human supervision.").

[16] *See* Ying Hu, *Robot Criminal Liability Revisited*, *in* DANGEROUS IDEAS IN LAW 494, 497-98 (Jin Soo Yoon, Sang Hoon Han & Seong Jo Ahn eds., 2018) (arguing that corporations are "structurally similar" to "robots that are equipped with machine learning algorithms to determine the appropriate course of actions in specific circumstances," and concluding that "if there is reason to treat corporations as moral agents, there is reason to treat sophisticated robots as moral agents as well"); Hu, *supra* note 15, at 520-21 ("One may argue that a smart robot can act intentionally in the same way that a corporation can. A robot's moral algorithms are functionally similar to a corporation's internal decision structure . . . . By analogy, . . . any act made pursuant to a smart robot's moral algorithms is an act done for the robot's own reasons and would therefore amount to an intentional action."). Unlike Hu, we do not argue that AIs have genuine moral responsibility. We focus on the legal notion of culpability, which involves institutional design constraints that allow it to diverge from moral responsibility or blameworthiness.

[17] Alexander Sarch, *Who Cares What You Think? Criminal Culpability and the Irrelevance of Unmanifested Mental States*, 36 L. & PHIL. 707, 709 (2017) [hereinafter *Who Cares*].

whose actions resemble those of other subjects of the criminal law, especially corporations. These harms may be *irreducible* where, for a variety of reasons, they are not directly attributable to the activity of a particular person or persons.[18] Corporations similarly can directly face criminal charges when their defective procedures generate condemnable harms[19] — particularly in scenarios where structural problems in corporate systems and processes are difficult to reduce to the wrongful actions of individuals.[20]

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 into criminal law doctrine, and to consider the alternatives. Our primary focus is not what form AI punishment would take, which could directly target AIs through censure, deactivation, or reprogramming, or could involve negative outcomes directed at natural persons or companies involved in the use or creation of AI.[21] Rather, our focus is the prior question of whether the doctrinal and theoretical commitments of the criminal law can be reconciled with criminal liability for AI.

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 practicably 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.[22] In

---

[18] *See infra* Part II.B.

[19] *See* MODEL PENAL CODE § 2.07 (AM. LAW INST. 1962) (outlining conditions under which a corporation could be convicted of an offense).

[20] *See* William S. 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); *infra* notes 166–168 and accompanying text (discussing ways to defend the irreducibility of corporate culpability).

[21] *See* Hu, *supra* note 15, at 529-30 (discussing the question of how a robot should be punished, and proposing "a range of measures [that] might be taken to ensure that the robot commits fewer offenses in the future"); Mark A. Lemley & Bryan Casey, *Remedies for Robots* 86 U. CHI. L. REV. 1311, 1316, 1389-93 (2019) (discussing the "robot death penalty").

[22] *See infra* Part III.

the end, our conclusion is this: While a coherent theoretical case can be made for punishing AI, 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 I provides a brief background of AI and "AI crime." It then provides a framework for justifying punishment that considers affirmative benefits, negative limitations, and feasible alternatives. Part II considers potential benefits to AI punishment, and argues it could provide general deterrence and expressive benefits. Part III 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 IV considers feasible alternatives to AI punishment. It argues the status quo is or will be inadequate for properly addressing AI crime. While direct AI punishment is a solution, this would require problematic changes to 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.

I. ARTIFICIAL INTELLIGENCE AND PUNISHMENT

A. Introduction to Artificial Intelligence

We use the term "AI" to refer to a machine that is capable of completing tasks otherwise typically requiring human cognition.[23] 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).

[23] AI lacks a standard definition, but its very first definition in 1955 holds up reasonably well: "[T]he artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving." J. MCCARTHY ET AL., A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE (1955), http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

AI is rapidly improving, driven by advances in software, computing power, and big data.[24] 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.[25] This was considered an important feat in the AI community, because of the sheer complexity of the game.[26] There are more possible Go board configurations than there are atoms in the universe.[27] 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.[28] Go was the last traditional board game in which people had been able to outperform machines.[29]

In some areas, AI already makes significant practical contributions. For instance, Google Translate supports more than 100 languages, including 37 by photo input, 32 by voice input, and 27 in "augmented reality mode."[30] 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.[31]

A few features of AI are important to highlight. First, AI has the potential to act unpredictably.[32] 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.[33] This is one means by which AI can

---

[24] *See* Ryan Abbott, *Everything Is Obvious*, 66 UCLA L. REV. 2, 23-28 (2019).

[25] *See id.* at 24.

[26] *See id.*

[27] *See id.*

[28] *See id.*

[29] *See id.*

[30] GOOGLE TRANSLATE, https://translate.google.com/intl/en/about/languages/ (last visited Oct. 9, 2019).

[31] *See, e.g.*, Daisuke Wakabayashi, *Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam*, N.Y. TIMES (Mar. 19, 2018), https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

[32] *See, e.g.*, Taha Yasseri, *Never Mind Killer Robots — Even the Good Ones Are Scarily Unpredictable*, PHYS.ORG (Aug. 25, 2017), https://phys.org/news/2017-08-mind-killer-robots-good-scarily.html; *Why Did the Neural Network Cross the Road?*, AI WEIRDNESS (2018), http://aiweirdness.com/post/174691534037/why-did-the-neural-network-cross-the-road (describing a programmer who made her machine learning algorithm attempt to tell jokes).

[33] *See, e.g.*, Davide Castelvecchi, *Can We Open the Black Box of AI?*, NATURE (Oct. 5, 2016), https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

engage in activities its original programmers may not have intended or foreseen.[34]

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.[35] This has led to some AIs being described as "black box" systems.[36] For instance, an algorithm may refuse a credit application but not be able to articulate why the application was rejected.[37] 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.[38] 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 iterative 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 in harmful ways 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.[39]

Fourth, while AI can already outperform people in spectacular fashion in some domains, like playing board games, in other domains AI is not even competitive with toddlers.[40] That is because all AI is

[34] 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, *The Risk of Machine-Learning Bias (and How to Prevent It)*, MIT SLOAN MGMT. REV. (Mar. 26, 2018), https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/.

[35] *See, e.g.*, Castelvecchi, *supra* note 33.

[36] *Id.*

[37] *See id.*

[38] *See id.*

[39] "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-71e6a8a551ee. The concept was to deploy an entity that could no longer be controlled by its creators and would act without further direction. The DAO would operate through smart contracts, 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. *See id.*

[40] *See* Abbott, *supra* note 24, at 40.

designed to perform "narrow" or "specific" tasks.[41] 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.[42] By contrast, the holy grail of computer science research is developing "general" AI that would be able to perform any task that a person could perform.[43] Experts are divided on whether, and when, general AI will be developed. For now, the weight of expert opinion holds there will likely be no general AI for at least a couple of decades.[44]

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.

Finally, general AI, and even super- or ultra-intelligent AI,[45] 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 a human-like abilities to cognitively reason and to be morally culpable for their actions.[46] Today, even the prospect of such machines is safely within the realm of science fiction.[47] We will not consider punishment of strong AI.[48]

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 had performed a similar act. Machines have caused harm since ancient times, and robots have caused fatalities since at least the 1970s.[49] However, besides

[41] *Id.* at 25.

[42] *See id.* at 24-25.

[43] *Id.* at 25.

[44] *See generally* Vincent C. Müller & Nick Bostrom, *Future Progress in Artificial Intelligence: A Survey of Expert Opinion*, *in* FUNDAMENTAL ISSUES OF ARTIFICIAL INTELLIGENCE 555 (Vincent C. Müller ed., 2016) (describing a survey finding that experts think AI superintelligence will not be a reality for at least a few decades).

[45] *See* Abbott, *supra* note 24, at 23-28 (describing super- and ultra-intelligent AI).

[46] *See* Jesus Rodriguez, *Gödel, Consciousness and the Weak vs. Strong AI Debate*, TOWARDS DATA SCI. (Aug. 23, 2018), https://towardsdatascience.com/g%C3%B6del-consciousness-and-the-weak-vs-strong-ai-debate-51e71a9189ca.

[47] *See id.*

[48] If and when such machines come into existence, we will certainly enjoy reading their works on AI criminal liability.

[49] *See* Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1, 8 (2018) [hereinafter *The Reasonable Computer*];

machines being intentionally used to inflict harm (as when a person runs someone over with their 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).[50] Such harms are not *mere* accidents; rather, they are accidents that implicate criminal law.[51] Nonetheless, even in such cases, criminal law is not usually deployed against the harmful machines themselves (at least outside of some intriguing but archaic prosecutions of inanimate objects).[52] 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 few essential ways that make the direct application of criminal law more worthy of consideration. Specifically, AI can behave in ways that display high degrees of *autonomy* and *irreducibility*.[53] In terms of autonomy, AI is capable of acting largely independently of human control. AI can receive sensory input, set targets, assess outcomes against criteria, make decisions and adjust behavior to increase its likelihood of success — all without being directed by human orders.[54] Reducibility is also critical

Bryan Young, *The First 'Killer Robot' Was Around Back in 1979*, HOW STUFF WORKS (Apr. 9, 2018), https://science.howstuffworks.com/first-killer-robot-was-around-back-in-1979.htm.

50 *See* Clifford Krauss & John Schwartz, *BP Will Plead Guilty and Pay Over $4 Billion*, N.Y. TIMES (Nov. 15, 2012), https://www.nytimes.com/2012/11/16/business/global/16iht-bp16.html.

51 *See BP Exploration and Production Inc. Agrees to Plead Guilty to Felony Manslaughter, Environmental Crimes and Obstruction of Congress Surrounding Deepwater Horizon Incident*, U.S. DEP'T JUSTICE (Nov. 15, 2012), https://www.justice.gov/opa/pr/bp-exploration-and-production-inc-agrees-plead-guilty-felony-manslaughter-environmental (outlining BP's guilty plea to criminal offenses).

52 *See* 1 WILLIAM BLACKSTONE, COMMENTARIES *302; OLIVER WENDELL HOLMES, THE COMMON LAW 7, 24 (1881) (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); Albert W. Alschuler, Comment, *Ancient Law and the Punishment of Corporations: Of Frankpledge and Deodand*, 71 B.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.").

53 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. *See, e.g.*, A.M. Turing, *Computing Machinery and Intelligence*, 59 MIND 433 (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* David J. Chalmers, *Facing Up to the Problem of Consciousness*, 2 J. CONSCIOUSNESS STUD. 200, 216 (1995) (distinguishing intellectual capacities from phenomenal consciousness).

54 *See supra* notes 24–37 and accompanying text.

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.[55] 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.

Almost all 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.[56] 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 the harm. Even when AI causes unforeseeable harm, it 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 that disrupts hospital care. This is a familiar problem in criminal law.[57] If someone commits a robbery and in so doing injures bystanders in unforeseeable ways (imagine a tripped bank alarm startles the animals in a neighboring zoo and they break loose and trample pedestrians), criminal law has doctrinal tools by which liability could still be imposed.[58]

Sometimes, however, it may be difficult to reduce AI crime to an individual due to AI autonomy, complexity, or lack of explainability. 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.[59] 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

---

[55] *See infra* Part IV.A.

[56] *See* Jeffrey K. Gurney, *Driving into the Unknown: Examining the Crossroads of Criminal Law and Autonomous Vehicles*, 5 WAKE FOREST J.L. & POL'Y 393, 433 (2015) (discussing crimes applicable to this scenario).

[57] *See infra* Part IV.

[58] *See infra* Part IV.A (discussing constructive liability offenses).

[59] In 2017, for instance, more than 4,500 Microsoft employees contributed to open source software hosted on GitHub. *See* Matt Asay, *Who Really Contributes to Open Source*, 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. *See* Frederic Lardinois & Ingrid Lunden, *Microsoft Has Acquired GitHub for $7.5B in Stock*, TECHCRUNCH (June 4, 2018, 6:08 AM), https://techcrunch.com/2018/06/04/microsoft-has-acquired-github-for-7-5b-in-microsoft-stock/.

machine has learned how to behave based on accessing millions or billions of data points from heterogeneous sources.[60] Thus, it may be more difficult to assign fault to individuals where AI is concerned versus a conventional product such as a car, where just one component is faulty. In fact, it may be *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 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.[61] Sometimes, although more controversially,[62] 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.[63] 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.[64]

There are several possible grounds on which criminal law might deem AI crime to be irreducible.[65]

---

[60] *See* Lothar Determann & Bruce Perens, *Open Cars*, 32 BERKELEY TECH. L.J. 915, 988 (2017).

[61] *See* MODEL PENAL CODE § 2.02(2)(a)-(c) (AM. LAW INST. 1962) (defining purpose and recklessness).

[62] *See* LARRY ALEXANDER & KIMBERLY KESSLER FERZAN, CRIME AND CULPABILITY: A THEORY OF CRIMINAL LAW 70 (2009) (arguing there should be no criminal liability for negligence).

[63] *See* § 2.02(2)(d) (defining negligence).

[64] 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 product liability. Abbott, *The Reasonable Computer*, *supra* note 49, at 13-16 (discussing product liability law). However, such liability only applies in limited situations. *See id.*

[65] *See infra* Part III.B.1.

> 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.[66]
>
> 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.[67]
>
> 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.[68]

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

---

[66] The chance of being prosecuted for a cyberattack in the United States is estimated at a mere 0.05% versus 46% for a violent crime. *See* William Dixon, *Fighting Cybercrime — What Happens to the Law When the Law Cannot Be Enforced?*, WORLD ECON. FORUM (Feb. 19, 2019), https://www.weforum.org/agenda/2019/02/fighting-cybercrime-what-happens-to-the-law-when-the-law-cannot-be-enforced/. Incidentally, cybercrime is predicted to cost the global economy $6 trillion by 2021. *See id.*

[67] *See infra* Part III.B.1 (discussing the reducibility challenge as applied to corporate liability).

[68] It's conceivable the law might adopt a "moral irreducibility" view. That is, the 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 generally infra* Part III.B.1.

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.[69] We use the term "punishment" roughly as defined by H.L.A. 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 offense;
>
> (iv) It must be intentionally administered by human beings other than the offender; and
>
> (v) It must be imposed and administered by an authority constituted by a legal system against which the offense is committed.[70]

Thus, "punishment" requires *a conviction for a legally recognized offense following accepted procedures*.[71] Under this definition, 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.[72] For instance, harsh

---

[69] *See generally* Mitchell N. Berman, *The Justification of Punishment*, *in* THE ROUTLEDGE COMPANION TO PHILOSOPHY OF LAW 141, 144-45 (Andrei Marmor ed., 2012) (noting the convergence on this sort of theory of punishment).

[70] H.L.A. HART, PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 4-5 (2d ed. 2008).

[71] 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. *See* Douglas N. Husak & 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, 30 (1994).

[72] *See* HART, *supra* note 70, at 5.

treatment by private citizens for violating informal social norms, preventative detention of people suffering from mental illnesses 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.[73] Civil penalties, while violations of legal norms, do not count as an "offense" for criminal law purposes, as criminal law seeks to condemn egregious categories of conduct.[74]

Punishment is justified only if its affirmative justifications outweigh its costs and it does not otherwise offend applicable negative limitations on punishment. Affirmative justifications are the positive benefits that punishment might produce like harm reduction, increased safety, enhanced well-being, 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 those types of acts. Affirmative justifications are distinct from *negative limitations on punishment*, which are commonly associated with culpability-focused retributivist views of criminal law. For example, it is widely held to be unjust to punish the innocent — or to punish wrongdoers in excess of what they deserve in virtue of their culpability — even if this would promote aggregate well-being in society.[75] This so-called "desert constraint" imposes a limitation, grounded in justice, on promoting social welfare through punishment.[76]

1. Affirmative Reasons to Punish

It is common to be a pluralist about the benefits of punishment.[77] U.S. federal law refers to the most widely acknowledged benefits, including

[73] 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. *See id.*

[74] 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* R A DUFF, THE REALM OF CRIMINAL LAW 19-20 (2018) [hereinafter REALM]. 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.

[75] *See infra* notes 94–95 and accompanying text.

[76] *See id.*

[77] *See* Berman, *supra* note 69, at 141-42 (noting the "converg[ence] on a desert-constrained pluralism" about the justifications of punishment, describing it as

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.[78]

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 are largely reducible to the aims in the first two categories.[79] 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.[80] 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.[81] 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 the future.[82] General deterrence occurs when punishing an offender discourages *other* would-be offenders from committing crimes.[83] It is a matter of punishing an offender in order to "send a message" to other potential offenders. There can be affirmative

"something approaching a consensus" view); Michael T. Cahill, *Punishment Pluralism*, *in* RETRIBUTIVISM: ESSAYS ON THEORY AND POLICY 25, 25 (Mark D. White ed., 2011) ("[M]any 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 . . . .").

78 18 U.S.C. § 3553 (2019).

79 *See* Berman, *supra* note 69, at 148 (discussing whether expressivism reduces to consequentialism).

80 *See Consequentialist Accounts*, STAN. ENCYCLOPEDIA PHIL. (July 18, 2017), https://plato.stanford.edu/entries/legal-punishment/#PurConPun ("It is commonly suggested that punishment can help to reduce crime by deterring, incapacitatiing [sic], or reforming potential offenders . . . .").

81 *See id.*

82 *See* Berman, *supra* note 69, at 145 (discussing types of deterrence).

83 *See id.*

benefits to punishing those who qualify for an insanity defense because it may deter sane individuals from committing crimes and attempting to rely on an insanity defense.[84]

These are not the only kinds of consequentialist benefits that can support punishment. Besides incapacitation and deterrence, punishment can reduce harm through rehabilitation of the offender.[85] 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.[86] 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.[87] 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.[88]

---

[84] 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 70, 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.

[85] *See* Berman, *supra* note 69, at 145 (discussing rehabilitation).

[86] 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.

[87] *See* VICTOR TADROS, THE ENDS OF HARM: THE MORAL FOUNDATIONS OF CRIMINAL LAW 26 (2011) [ENDS OF HARM] (identifying retributivism with the claim that it is "intrinsically valuable" that offenders suffer in proportion to their desert); John Cottingham, *Varieties of Retribution*, 29 PHIL. Q. 238, 238 (1979); Doug Husak, *Retributivism in Extremis*, 32 L. & PHIL. 3, 4-5 (2013) (defending broader versions of retributivism).

[88] *See* Leora Dahan Katz, *Response Retributivism: Defending the Duty to Punish* 16-17 (Oct. 10, 2018) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3264139.

While virtually everyone agrees that the good consequences of preventing crime must be a major part of what justifies punishment,[89] there is more debate about whether *retributivist reasons* also exist.[90] 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*.[91] Punishment involves the communication of society's collective commitment to certain core values. The state, through punishment, conveys official condemnation of culpable conduct through the mechanism of a criminal conviction. Victims may benefit 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.[92]

Some question whether expressive benefits are a distinct category of reason to punish, or whether they simply reduce to consequentialist or retributivist reasons.[93] 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 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.[94] The desert constraint claims that

---

[89] *See* TADROS, ENDS OF HARM, *supra* note 87, at 21.

[90] *See, e.g.*, *id.* at 60.

[91] *See* R A DUFF, ANSWERING FOR CRIME: RESPONSIBILITY AND LIABILITY IN THE CRIMINAL LAW 140-46 (2007) [hereinafter ANSWERING].

[92] *See* PAUL H. ROBINSON, INTUITIONS OF JUSTICE AND THE UTILITY OF DESERT 161-62 (2013).

[93] *E.g.*, Berman, *supra* note 69, at 148.

[94] *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

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.[95] 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.[96] 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).[97] The idea is that punishing the innocent to obtain broader social benefits is a paradigmatic example of treating people merely as means, which 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 persons. Others have a more nuanced view, such that violating a negative limitation could be overall justified if the benefits were sufficiently weighty.[98]

There are limitations on punishment other than the desert constraint. Most importantly, criminal law requires certain prerequisites, such as a capacity for culpability, that defendants must meet in order to be properly subject to punishment. It is a fundamental aim of criminal law to condemn culpable wrongdoing, and it is the default position in criminal law doctrine that punishment may only be properly imposed

because of] any good consequences" it may have); *see also id.* at 151 (discussing desert-constrained consequentialism).

[95] 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.

[96] *Cf.* HART, *supra* note 70, at 12 (discussing the famous example of the small southern town).

[97] *See, e.g.*, TADROS, ENDS OF HARM, *supra* note 87, at 114 (defending a version of the means principle).

[98] *See, e.g.*, HART, *supra* note 70, 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.").

in response to culpable wrongdoing.[99] 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.[100] Thus, capacity for culpability is an eligibility requirement for being aptly subject to regulation by 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.[101]

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.[102] 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.

---

[99] *See* MODEL PENAL CODE § 1.02(c) (AM. LAW INST. 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 & Hanna Pickard, *To Blame or to Forgive? Reconciling Punishment and Forgiveness in Criminal Justice*, 35 OXFORD J. LEGAL STUD. 665, 666 (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"); DUFF, REALM, *supra* note 74, at 19-20 (noting that "censure . . . is essential to a criminal conviction" and that 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").

[100] *See* MODEL PENAL CODE § 4.01 (outlining the incapacity defense based on mental defect as when a person is unable "either to appreciate the criminality . . . of his conduct or to conform [it to] the law").

[101] *See* Sven Ove Hansson, *Philosophical Problems in Cost-Benefit Analysis*, 23 ECON. & PHIL. 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).").

[102] *E.g.*, Doug Husak, *The Criminal Law as Last Resort*, 24 OXFORD J. LEGAL STUD. 207, 235 (2004) (discussing the view that "[a]mong those modes of social control we are likely to deem acceptable, the criminal law should be used only as a last resort").

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) *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, compared to alternatives like civil liability, regulation, or doing nothing?

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

II. 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.[103] 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

[103] *See* TADROS, ENDS OF HARM, *supra* note 87, at 25-28.

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."[104] The idea is that if AIs cannot 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.[105]

The answer to the undeterrability argument requires distinguishing specific deterrence from general deterrence.[106] Specific deterrence involves incentivizing a particular defendant not to commit crimes in the future.[107] 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*.

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

---

104 Peter M. Asaro, *A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 169, 181 (Patrick Lin et al. eds., 2012). Asaro is ultimately skeptical of punishing robots because of questions about how to make AI be directly responsive to punishments. *See id.* at 182.

105 It is conceivable that AIs could 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 compared to programming it not to run a red light *except* in unspecified emergency conditions. *Cf.* Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557 (1992) (arguing that standards are more costly to interpret and apply).

106 *See* HART, *supra* note 70, at 19.

107 *See* Berman, *supra* note 69, at 145.

termed the "robot death penalty,"[108] punishing AI directly could deprive such developers, owners or users of the financial benefits of the systems. This penalty may thereby incentivize such human parties 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.[109] 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."[110] On her view, "robot punishment — or more precisely, revenge against robots — primarily advances . . . the creation of psychological satisfaction in robots' victims."[111] Punishment conveys a message of official condemnation that could reaffirm the interests, rights, and ultimately the value of the victims of the harmful AI.[112] This, in turn, could produce an increased sense of security among victims and society in general.

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

---

[108] Lemley & Casey, *supra* note 21, at 100.

[109] Analogous considerations could apply to provide support for punishing inanimate objects and corporations.

[110] Christina Mulligan, *Revenge Against Robots*, 69 S.C. L. REV. 579, 580 (2018); *cf.* David Lewis, *The Punishment That Leaves Something to Chance*, 18 PHIL. & PUB. AFF. 53, 54 (1989) (discussing but rejecting the idea of defending puzzling features of criminal law on the ground that when harm results the population may demand blood).

[111] Mulligan, *supra* note 110, at 593.

[112] *See* DUFF, ANSWERING, *supra* note 91, at 114; 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 & Zachary Shapiro, *A Theory of Differential Punishment*, 70 VAND. L. REV. 1499, 1503 (2017) (arguing that victims' interests should be taken "into account when determining how severely criminal offenders should be punished").

persons like corporations.[113] 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.[114] 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.[115] 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.[116] 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 criminal law. Thus, a number of benefits could be obtained through the expressive function of punishment.[117]

Nonetheless, there are a range of *prima facie* worries about appealing to expressive benefits like victim satisfaction in order to justify the

---

[113] *See* Mihailis E. Diamantis, *Corporate Criminal Minds*, 91 NOTRE DAME L. REV. 2049, 2078 (2016) (arguing that "[w]hen 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).

[114] *See* Matthias Scheutz, *The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 205, 205-22 (Patrick Lin et al. eds., 2012); Sherry Turkle*, In Good Company? On the Threshold of Robotic Companions*, *in* CLOSE ENGAGEMENTS WITH ARTIFICIAL COMPANIONS 3, 3 (Yorick Wilks ed., 2010); Luisa Damiano & Paul Dumouchel, *Anthropomorphism in Human-Robot Co-evolution*, FRONTIERS IN PSYCHOL. (Mar. 26, 2018), https://doi.org/10.3389/fpsyg.2018.00468 (discussing "social robotics" which sees anthropomorphism not as "a cognitive error" but as a useful tool "to facilitate social interactions between humans and . . . social robots"); *see* Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 538 (2015).

[115] *See* Diamantis, *supra* note 113, 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* ROBINSON, *supra* note 92, at 176-88 (discussing lay observers' ideas as they bear on the legitimacy of the criminal justice system).

[116] *See, e.g.*, Margaret Rhodes, *The Touchy Task of Making Robots Seem Human — But Not Too Human,* WIRED (Jan. 19, 2017, 7:00 AM), https://www.wired.com/2017/01/touchy-task-making-robots-seem-human-not-human/.

[117] 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.

punishment of AI. First, 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."[118] Simply put, the popularity of a practice does not automatically justify it, even if popularity could be relevant to its normative justification. Second, punishing AI for expressivist purposes could lead to further bad behavior that might spill over to the way other *humans* are treated.[119] Thus, Kate Darling has argued robots should be protected from cruelty in order to reflect moral norms and prevent undesirable human behavior.[120]

Third, 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 IV. Before that, we turn to another kind of threshold question: whether punishing AI violates the culpability-focused negative limitations on punishment.

---

[118] Lewis, *supra* note 110, at 54.

[119] 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. *See The Moral Status of Animals*, STAN. ENCYCLOPEDIA PHIL. (Aug. 23, 2017) https://plato.stanford.edu/entries/moral-animal/ (discussing Kant's view of ethical treatment of animals: 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" (quoting Kant's *Lectures on Ethics*)).

[120] *See* Kate Darling, *Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects*, *in* ROBOT LAW 213, 228 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016). Relatedly, in the United Kingdom, laws criminalizing animal cruelty exist to disapprove of offensive human conduct. *See* Animal Welfare Act 2006, c. 45 (Gr. Brit.), http://www.legislation.gov.uk/ukpga/2006/45/contents.

III. RETRIBUTIVE AND CONCEPTUAL LIMITATIONS

This Part considers retributivist (culpability-focused) limitations on punishment. Section A asks whether AI is the right kind of entity to be eligible for punishment — what we call *The Eligibility Challenge*. Where criminal law's fundamental prerequisites are not satisfied, its sanctions are not 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. Finally, Section C considers the conceptual objection that AI punishment is not actually punishment at all.

A. The Eligibility Challenge

The Eligibility Challenge is simple to state: AI, like inanimate objects, is not the right kind of thing to be punished. AI lacks mental states and the deliberative capacities needed for culpability, so it cannot be punished without sacrificing the core commitments of the criminal law. The issue is not that AI punishment would be unfair to AI. AIs are not conscious and do not feel (at least in the phenomenal sense),[121] and they do not possess interests or well-being.[122] 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.[123] 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.[124] If punishing AI

[121] *See* Chalmers, *supra* note 53, at 201 (describing phenomenal experiences as those personally felt or experienced).

[122] *See id.* (discussing the hard problem of consciousness).

[123] *See supra* notes 94–96 and accompanying text.

[124] *See* Husak & Callender, *supra* note 71, at 32-33.

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 that criminal law aims to respond to.[125] 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.[126] The capacity for culpable conduct thus is a general prerequisite of 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 criminal law. Punishing AI despite its lack of capacity would not only be conceptually confused, but would fail to serve the retributive aims of criminal law — namely, to mark out seriously culpable conduct for the strictest public condemnation.

Here we develop three answers to the Eligibility Challenge.

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.[127] However, even if corporations cannot literally satisfy mens rea elements, criminal law has

---

[125] *See generally* Douglas Husak, *"Broad" Culpability and the Retributivist Dream*, 9 OHIO ST. J. CRIM. L. 449, 456-57 (2012) (distinguishing narrow culpability as merely mens rea categories from broad culpability, which is the underlying normative defect that criminal law aims to respond to).

[126] *See* MODEL PENAL CODE § 1.02(c) (AM. LAW INST. 1962); *see also* MICHAEL S. MOORE, PLACING BLAME: A GENERAL THEORY OF THE CRIMINAL LAW 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, REALM, *supra* note 74, at 20 (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. *See* 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).

[127] *See, e.g.*, Albert W. Alschuler, *Two Ways to Think About the Punishment of Corporations*, 46 AM. CRIM. L. REV. 1359, 1367-69 (2009) (arguing against corporate punishment).

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.[128] Some jurisdictions also tack on further requirements.[129] 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.[130]

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.[131]

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

---

[128] *See* 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) ("[R]espondeat 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) (explaining that 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)).

[129] *See* MODEL PENAL CODE § 2.07(1)(c) (AM. LAW INST. 1962) (adopting respondeat superior but restricting it to the mental states of high corporate officials).

[130] 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. *See* Paul H. Robinson, *Imputed Criminal Liability*, 93 YALE L.J. 609, 611-12 (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 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 not be a very weighty one.

[131] *See* Hallevy, *supra* note 10, at 201 (arguing "there is no substantive legal difference between the idea of criminal liability imposed on corporations and on AI entities").

ready supply of identifiable human actors whose mental states can be imputed.[132] 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 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 is 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 lack of a culpable mental state. That is not to simply reach for a consequentialist justification[133] of the conceptual confusion or inaptness involved in applying criminal law to AI. Within 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 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.

---

[132] Although, Shawn Bayern has argued that existing American LLC statutes allow an organization not to be legally associated with human members. Shawn Bayern, *Are Autonomous Entities Possible?*, 114 NW. U. L. REV. ONLINE 23, 26 (2019). He argues on this basis that organizational statutes are thus flexible enough to give something like legal personhood to software systems, because an AI can also direct the activities of an organization. *Id.*

[133] *See supra* notes 77–82 (explaining the idea of justifying punishment based on its good consequences).

Many legal scholars are highly critical of strict liability offenses. For example, 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.[134]

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 constraint[135] (which prohibits punishment in excess of culpability).[136]

This strategy is not without problems. Even to be guilty of a strict liability offense, defendants still must satisfy the *voluntary act requirement*.[137] LaFave's criminal law treatise observes that "a voluntary act is an absolute requirement for criminal liability."[138] The Model 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."[139] Behaviors like reflexes, convulsions or movements that occur unconsciously or while sleeping are expressly ruled out as non-voluntary.[140] To be a voluntary act, "only bodily movements guided by *conscious* mental representations count".[141] 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

---

134 DUFF, REALM, *supra* note 74, at 19.

135 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 infra* Part III.B.2.

136 Matters would be different if AIs, perhaps like many animals, could experience pleasure and pain, or were conscious or otherwise in possession of 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).

137 *See* WAYNE R. LAFAVE, 1 SUBSTANTIVE CRIMINAL LAW § 6.1(c) (3d ed. 2018) ("[C]riminal liability requires that the activity in question be voluntary.").

138 *Id.*

139 MODEL PENAL CODE § 2.01(1) (AM. LAW INST. 1962).

140 *See id.* § 2.01(2).

141 Gideon Yaffe, *The Voluntary Act Requirement*, *in* THE ROUTLEDGE COMPANION TO THE PHILOSOPHY OF LAW 174, 175 (Andrei Marmor ed., 2012).

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."[142] 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 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.[143] That is, it threatens to *undermine the expressive benefits* that supposedly help justify punishing AI in the first place.[144] 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.[145] 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 observed fashion.[146] Similarly, in AI cases, experts might need only

---

142 MODEL PENAL CODE § 2.01(3) (AM. LAW INST. 1962).

143 *See* DUFF, REALM, *supra* note 74, at 19-20.

144 *See supra* Part II.B.

145 In Part III.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 164 (discussing the collective knowledge doctrine for corporations).

146 *See generally* Peter Carruthers, *Mindreading in Infancy*, 28 MIND & LANGUAGE 141, 143 (2013) (discussing how infants attribute beliefs and intentions to others); David Premack & Guy Woodruff, *Does the Chimpanzee Have a Theory of Mind?*, 1 BEHAV. BRAIN SCI. 515, 515 (1978) (defining "theory of mind" as the system whereby "the individual

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 decision-makers 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 holds that a person is criminally culpable for an action to the extent that he or she *manifests insufficient regard* for legally protected interests or values.[147] 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.[148] The criminal

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").

147 *See* ALEXANDER & FERZAN, *supra* note 62, at 67-68 (2009) ("[I]nsufficient concern [is] the essence of culpability."); Peter K. Westen, *An Attitudinal Theory of Excuse*, 25 L. & PHIL. 289, 374-75 (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 also* VICTOR TADROS, CRIMINAL RESPONSIBILITY 250 (2005) ("[I]f [a defendant] is convicted of a serious offence, the state communicates . . . that [his] behaviour manifested an inappropriate regard for other citizens and their interests . . . ."); Gideon Yaffe, *Intoxication, Recklessness, and Negligence*, 9 OHIO ST. J. CRIM. L. 545, 552-53 (2012).

148 *See, e.g.*, GIDEON YAFFE, ATTEMPTS 38 (2010) (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, *Criminal Culpability After the Act*, 26 KING'S L.J. 440, 453 (2015) (distinguishing moral from legal culpability), *and* Sarch, *Who Cares*, *supra* note 17, at 710, *with* Michael S. Moore, *Choice, Character, and Excuse*, 7 SOC. PHIL. & POL'Y 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 PHIL. & PHENOMENOLOGICAL RES. 388, 403 (2014) ("[W]hat people are truly blameworthy for are the motivations from which [their] actions spring, rather than the actions themselves."), *with* NOMY ARPALY & TIMOTHY SCHROEDER, IN PRAISE OF DESIRE 170 (2014) (defending a notion of

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.[149] 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 criminal law not to be overly concerned with the specific motives or private mental states 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.[150] 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 criminal law demands not displaying insufficient regard for in action.[151]

blameworthiness that is similar to criminal culpability as described here). Our focus here, regardless, is *legal* culpability.

149 *See* Sarch, *Who Cares*, *supra* note 17, at 709-10.

150 *See* GIDEON YAFFE, THE AGE OF CULPABILITY: CHILDREN AND THE NATURE OF CRIMINAL RESPONSIBILITY (2018) (developing an evidentialist account of manifestation of insufficient regard); *see also* Sarch, *Who Cares*, *supra* note 17, at 727-33.

151 *See* CHRISTIAN LIST & PHILIP PETTIT, GROUP AGENCY 158-63 (2011) (arguing that corporations can have decision-making structures that satisfy the main preconditions for responsibility); W. Robert Thomas, *The Ability and Responsibility of Corporate Law to Improve Criminal Punishment*, 78 OHIO ST. L.J. 601, 612-13 (2017) ("Corporations have free will in a narrow sense: 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.").

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 the corporation to alter its conduct. If the corporation continues its manufacturing activities unchanged, this demonstrates — through its information-sharing and decision-making procedures — that it did not attach sufficient weight to the legally recognized reasons against continuing its dangerous activities. This is paradigmatic criminal culpability.[152]

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.[153] AI may gather information, process it, and determine the most efficient means to accomplish its goals.[154] 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.[155]

---

[152] 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. *See infra* Part III.B. It does not undermine corporations' threshold eligibility for punishment. *See* Thomas, *supra* note 151, at 613 (noting that 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").

[153] *See supra* notes 31–48.

[154] *See id.*

[155] This idea is similar in some respects to Hu's argument that robots may be punished only if they possess "moral algorithms." *See* Hu, *supra* note 15, at 496. These are "algorithms that are capable of making nontrivial morally relevant decisions," (i.e., ones that "concern[] a choice between or among two or more courses of actions that might be considered right or wrong by ordinary members of our society."). *Id.* As Hu notes, these could be taken on the basis of strict rules, guiding principles or an effort to weigh the competing interests at play in order to determine what would maximize some expected utility function. *See id.* at 497-98; *see also id.* at 505 (considering the analogy between corporate moral responsibility and similar responsibility for AI). Unlike Hu, our arguments do not assume or require that AI are moral persons or have moral responsibility. We are concerned with legal culpability, the demands of which are less exacting than true moral responsibility.

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 the Philosophy of Action characterizing the functional role of human intentions could be extended to AI. On Bratman's well-known account,[156] actors who intend (i.e., act with the purpose) to bring about an outcome "guide [their] conduct in the direction of causing" that outcome.[157] This means that in 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.[158] Suppose an actor is driving with the intention to hit a pedestrian. In that case, if the actor detects that conditions have changed so that behavioral adjustments are required to make this outcome more likely, then the actor will be disposed to make these adjustments. Moreover, actors with this intention will be disposed to monitor the circumstances to find ways to increase the likelihood of the desired outcome. Merely foreseeing the outcome, but not intending it, does not similarly entail that one will *guide* one's behavior in these ways to promote the outcome in question (i.e., 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? 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 plausibly may be 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.

156 *See* Michael E. Bratman, *What Is Intention?*, *in* INTENTIONS IN COMMUNICATION 15, 23-27 (Philip R. Cohen et al. eds., 1990); *see also* Alex Sarch, *Double Effect and the Criminal Law*, 11 CRIM. L. & PHIL. 453, 467-68 (2015).

157 Bratman, *supra* note 156, at 26.

158 *See id.*

But it does not seem impossible in principle, even if difficult questions remain.[159]

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.[160] In criminal law, knowledge is defined as practical certainty.[161] 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

[159] For example, suppose the autonomous vehicle is actually aiming not to harm pedestrians by hitting them, but rather aims for 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 314-17 (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 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.

[160] *See Belief*, STAN. ENCYCLOPEDIA PHIL. (June 3, 2019), 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 the disposition to depend on *P*'s truth in [acting]. [More generally, this amounts to] being disposed to act as though *P* is the case.").

[161] *See* MODEL PENAL CODE § 2.02(2)(b) (AM. LAW INST. 1962) (defining knowledge as practical certainty).

of these dispositions are proven, then knowledge that F could be attributed to the AI.[162] 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).[163]

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.[164] 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,[165] 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.

B. Further 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.

---

[162] *Cf.* Eric Schwitzgebel, *In-Between Believing*, 51 PHIL. Q. 76 (2001) (defending this approach to determining when to attribute beliefs to humans).

[163] *See* MODEL PENAL CODE § 2.02(2)(c) (AM. LAW INST. 1962) (defining recklessness).

[164] *See* United States v. Bank of New England, 821 F.2d 844, 856 (1st Cir. 1987) (embracing one version of collective knowledge doctrine); ALEXANDER SARCH, CRIMINALLY IGNORANT 246 (2019) (defending the collective knowledge doctrine as an equal culpability imputation rule for corporations).

[165] 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 latter should not generate independent liability while the former would seems to be a distinction without a normative difference.

1. Reducibility

One might object that there is never a genuine need to punish AI because any time an AI seems criminally culpable in its own right, 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.[166] 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, or the managers pushing employees to cut corners in search of savings. For any case offered to demonstrate the irreducibility of corporate culpability,[167] a skeptic may creatively find additional wrongdoing by other individual actors further afield or in the past to account for the apparent corporate culpability.[168]

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* of their agents (albeit organized in particular structures). Also, AI may sometimes behave in ways that are less predictable and foreseeable than corporate conduct.

Nonetheless, 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 criminal law

---

166 *See, e.g.*, Andras Szigeti, *Are Individualist Accounts of Collective Responsibility Morally Deficient?*, *in* INSTITUTIONS, EMOTIONS, AND GROUP AGENTS: CONTRIBUTIONS TO SOCIAL ONTOLOGY 329 (2014) (Anita Konzelmann Ziv & Hans Bernhard Schmid eds., 2013) (arguing that the individualist analysis does not leave any responsibility-deficit that would require a genuine group culpability).

167 Consider List and Pettit's notion of a "responsibility deficit." LIST & PETTIT, *supra* note 151, at 165. 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.* 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?

168 *See id.* at 158.

to always allow any putative corporate criminal culpability to be reduced to individual criminal liability. This would require criminalizing very minute portions 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 should 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 the freedom of expression and association.[169] It would also increase the risk of abuse of process. Instead, we should expect "culpability deficits"[170] 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 to, any time an AI causes harm, 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. The criminal justice system would be disturbingly invasive if it had to create 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.[171] Because corporate punishments (usually in the form of fines) amount to a hit to the corporation's bottom line, these punishments inevitably spill over

169 *Cf.* HART, *supra* note 70, at 1-27.

170 *See* LIST & PETTIT, *supra* note 151, at 165 (defending "responsibility deficits" as creating a need for irreducible corporate accountability).

171 *See* Thomas, *supra* note 151, at 619.

onto innocent shareholders.[172] 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."[173] 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.[174] Nonetheless, this definitional answer is somewhat unsatisfying, as there clearly are strong reasons for the state not to knowingly harm innocent bystanders even if the harm 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.[175] For example, Will Thomas argues that managers found to have contributed to a crime committed 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.[176]

---

[172] *See* 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").

[173] Mulligan, *supra* note 9, at 594.

[174] *See* HART, *supra* note 70, at 4-5.

[175] *See* Thomas, *supra* note 151, at 647.

[176] *See id.* at 647-49.

Similar thinking applies to AI punishment, which likewise 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 should be addressed for any novel mode of criminal punishment — whether for corporations or AI.

C. Not Really Punishment?

We end this Part by considering another challenge to AI punishment — that AI cannot be truly "punished." Even if an AI was convicted of an offense and subject to negative treatment — such as being reprogrammed or terminated — this may not be punishment under our working definition. On Hart's definition introduced in Part I.C, punishment "must involve pain or other consequences normally considered unpleasant."[177] However, AI cannot experience things as being painful or unpleasant.[178]

A first response is to argue that AI punishment *does* satisfy Hart's definition because prong (i) requires only that the treatment in question must be "normally considered unpleasant" — not that it be actually unpleasant or unwelcome to a convicted party. This is what allows Hart's definition to accommodate people who, for idiosyncratic reasons, do not experience their sentence as unpleasant or bad and to still regard this as punishment. The mere fact that a convicted party overtly wants to be imprisoned, like the Norwegian mass murderer Anders Bering Breivik, who wanted to be convicted and imprisoned to further his political agenda, does not mean that doing so pursuant to a conviction

[177] *See* HART, *supra* note 70, at 4.

[178] *See supra* note 53 and accompanying text.

ceases to be punishment.[179] Something similar might be said for AI as well as defendants who may be physically or psychologically incapable of experiencing pain or distress. Having one's actions frozen or being terminated really are the kind of thing that can "normally be considered unpleasant."

This response can be developed further. Why might punishment need to be *normally* regarded as unpleasant? Why does it still seem to be punishment, for example, to imprison a person who in no way experiences it as unpleasant or unwelcome? The answer may be that defendants can have interests that are *objectively* set back even when they do not experience these setbacks as painful, unpleasant or unwelcome.[180] Some philosophers argue it is intrinsically bad for humans to have their physical or agential capacities diminished — regardless of whether this is perceived as negative.[181] If correct, this suggests that what prong (i) of Hart's definition, properly understood, requires is that punishment involve events that *objectively* set back interests, and negative subjective experiences are merely one way to objectively set back interests.

Can an AI have interests that are capable of being set back? AI is not conscious in the phenomenal sense of having subjective experiences and thus cannot experience anything as painful or unpleasant.[182] However, one could maintain that being incapacitated or destroyed is *objectively* bad for AIs even if the AI does not experience it as such — in much the same way that things like nutrition, reproduction, or physical damage can be said to be good or bad for biological entities like plants or animals.[183] Some philosophers argue that it is in virtue of

[179] *See Anders Breivik Found Sane: The Verdict Explained*, TELEGRAPH (Aug. 24, 2012, 10:15 AM), https://www.telegraph.co.uk/news/worldnews/europe/norway/9496641/Anders-Breivik-found-sane-the-verdict-explained.html (discussing why Breivik would "want to be sent to prison" rather than getting the benefit of an insanity defense).

[180] *See* Guy Fletcher, *A Fresh Start for the Objective-List Theory of Well-Being*, 25 UTILITAS 206, 206 (2013) (defending objective theories of well-being from familiar objections); Alexander Sarch, *Multi-Component Theories of Well-Being and Their Structure*, 93 PAC. PHIL. Q. 439, 439-41 (defending a partially objective theory of well-being, where both subjective experiences and some objective components can impact well-being).

[181] *See* Elizabeth Harman, *Harming As Causing Harm*, *in* HARMING FUTURE PERSONS 137, 139 (Melinda A. Roberts & David T. Wasserman eds., 2009) (arguing that an event harms an agent when that event causes P to be in an intrinsically bad state, where such states include "pain, mental or physical discomfort, disease, deformity, disability, or death").

[182] *See supra* note 53 and accompanying text.

[183] *See* PHILIPPA FOOT, NATURAL GOODNESS 26 (2001) (noting that "features of plants and animals have what one might call an 'autonomous', 'intrinsic', or as I shall say

something's having identifiable *functions* that things can be good or bad for it. Most notably, Philippa Foot defends this sort of view (tracing it to Aristotle) when she argues that the members of a given species can be evaluated as excellent or defective by reference to the functions that are built into its characteristic form of life.[184] From this evaluation as flourishing or defective, facts about what is good or bad for the entity can be derived. Thus, if having interests in this broad, function-based sense is all that is required for punishment to be sensible, then perhaps AI fits the bill. AIs also have a range of functions — characteristic patterns of behavior needed to continue in good working order and succeed at the tasks it characteristically undertakes. If living organisms can in a thin sense be said to have an interest in survival and reproduction, ultimately in virtue of their biological programming, then arguably an AI following digital programming could have interests in this thin sense as well.

Other philosophers reject this view, however. They insist that only those entities capable of having beliefs and desires, or at least phenomenal experiences such as of pleasure and pain, can truly be said to have full-blooded interests that are normatively important. Legal philosopher Joel Feinberg took the capacity for cognition as the touchstone full-blooded interests, that is as a precondition for having things *really* be good or bad for us.[185] He notes "we do say that certain conditions are 'good' or 'bad' for plants" (unlike rocks), but he denies that they have full-blooded interests.[186] Although "Aristotle and Aquinas took trees [and plants] to have their own 'natural ends'" (in much the same sense that Foot argues for), Feinberg denies plants "the status of beings with interests of their own" because "an interest, however the concept is finally to be analyzed, presupposes at least rudimentary cognitive equipment."[187] Interests, he thinks, "are compounded out of *desires* and *aims,* both of which presuppose

'natural' goodness and defect that may have nothing to do with the needs or wants of the members of any other species of living thing").

184 *See id.* at 33 ("The way an individual *should be* is determined by what is needed for development, self-maintenance, and reproduction: in most species involving defence, and in some the rearing of the young . . . . Thus, evaluation of an individual living thing in its own right, with no reference to our interests or desires, is possible [by reference to the functions of the thing, as captured in] Aristotelian categoricals (life-form descriptions relating to the species)").

185 *See* Joel Feinberg, *The Rights of Animals and Unborn Generations*, *in* PHILOSOPHY AND ENVIRONMENTAL CRISIS 43, 49-51 (William T. Blackstone ed., 1974).

186 *Id.* at 51.

187 *Id.* at 52.

something like *belief,* or cognitive awareness."[188] Since AIs are not literally capable of cognitive awareness (notwithstanding the discussion in Part III.A of how mens rea might be imputed), they cannot literally possess full-blooded interests of the kind Feinberg has in mind.[189]

Thus, the pertinent question for present purposes is what sense of interest an entity must have for it to be intelligible to talk of punishing it — the thin sense of function-based interests of the kind Foot defended or the full-blooded, attitudinally-based interests Feinberg had in mind? This is ultimately a question about how to understand prong (i) of Hart's definition of punishment, and one that goes to the heart of what criminal law *is* and what it is for. We simply note that this is one possible way of defending the idea of AI punishment as sensible.

A different, perhaps stronger, type of reply is to distinguish between *conviction* and *punishment*, where the latter covers the sentence to which the convicted party is subject. Even if no form of treatment can count as punishment unless the entity in question experiences it negatively, this is not a precondition for a *conviction*. Perhaps for it to be intelligible to convict X of an offense, it is only required that X acted in ways that violated a prohibition and this can be sensibly construed as culpable (a manifestation of insufficient regard). If so, then one might accept that while *punishing* AI is not conceptually possible, applying criminal law to AIs, so they could be convicted of offenses, is. Thus, society might still benefit from AI convictions while not running afoul of the conceptual confusion that results from purporting to punish AIs.

Convicting AIs may require, or allow, subjecting other parties to punishment in place of the AI. Criminal law roundly rejects "vicarious punishment" where people are concerned[190] — not least where it risks the injustice of strict criminal liability imposed on innocent actors.[191] Corporate punishment might seem to involve vicarious punishment when officers or employees of the corporation are made to suffer due to the criminal fines imposed on the corporation. However, such cases are better understood not as vicarious liability, strictly speaking, but as convicting the corporation of an offense directly and then allowing the

188 *Id.*

189 *See id.* at 49-50 (arguing that an entity cannot have full-blooded interests if it has "no conative life: no conscious wishes, desires, and hopes; or urges and impulses; or unconscious drives, aims, and goals; or latent tendencies, direction of growth, and natural fulfillments").

190 *See* Joel Feinberg, *Collective Responsibility*, 65 J. PHIL. 674, 680 (1968) ("[S]urely there is no going back to [collective punishment] . . . . On the contrary, the changes that have come with modern times have dictated quite inevitably that [individual punishment] replace [collective punishment].").

191 *See supra* notes 133–134 and accompanying text.

sentence to be *distributed* to the different individuals out of which the corporation is made up.[192] In the case of an AI, it could be argued that if human owners or users accept responsibility for operating the AI safely, then were the AI convicted of an offense in its own right, these responsible parties would be the appropriate persons to whom the sentence could be distributed by virtue of their voluntarily undertaking such responsibility. We explore a simple version of this idea further in the next Part.

A final type of reply, always available as a last resort, is that even if applying criminal law to AIs is conceptually confused, it could still have good consequences to *call it punishment* when AIs are convicted. This would not be to defend AI punishment from within existing criminal law principles, but to suggest that there are consequentialist reasons to depart from them.

IV. FEASIBLE ALTERNATIVES

We have argued that punishing AI could have benefits and that doing so would not be ruled out by the negative limitations and retributive preconditions of punishment. But this does not yet show the punishment of AI to be justified. Doing so requires addressing the third main question in our theory of punishment: Would the benefits of punishing AI outweigh the costs, and would punishment 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.

Ideally a cost-benefit analysis would involve more than identifying various costs and benefits, and would include quantitative analysis. If only a single Hard AI Crime is committed each decade, there would be far less need to address an AI criminal gap than if Hard AI Crime was a daily occurrence. The absence of evidence suggesting that Hard AI Crime is common counsels against taking potentially more costly actions now, but this balance may change as technological advances result in more AI activity.

Section A focuses on 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 criminal law as well as tools available within civil law,

[192] *See* Thomas, *supra* note 151, at 612-13.

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. The impact of this gap is an empirical question we do not attempt to answer here.

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.[193]

Even if additional computer-related offenses must be created to adequately deter novel crimes implemented with the use of AI, 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).[194] Under the innocent agency doctrine, criminal liability attaches to a person who acts through an agent who lacks capacity — such as a child or someone with an insanity defense. 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.[195] This could be analogous to a person programming a sophisticated AI to break the law: the person has liability for

[193] *See* 18 U.S.C. § 1030(a)(1)-(7) (2019) (defining offenses such as computer trespass and computer fraud); *id.* § 1343 (wire fraud statute).

[194] Hallevy, *supra* note 10, at 179-81.

[195] *See* Sanford H. Kadish, *Complicity, Cause and Blame: A Study in the Interpretation of Doctrine*, 73 CALIF. L. REV. 323, 372-73 (1985) ("Most criminal actions can readily be committed through the instrumentality of another person.").

intentionally causing the AI to bring about the external elements of the offense.[196]

This doctrine requires intent (or at least the knowledge) that the innocent agent will cause the prohibited result in question.[197] This means that in cases where someone does not intend or 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.[198] 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.[199] If such a risk was merely reasonably foreseeable (but not foreseen), then lower forms of homicide liability would be available.[200] 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.[201]

Hallevy also discusses this form of criminal liability for AI-generated harms, calling it the "natural and probable consequences model" of liability.[202] 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

196 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 III 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.

197 *See* Peter Alldridge, *The Doctrine of Innocent Agency*, 2 CRIM. L. F. 45, 70-71 (1990); 18 U.S.C. § 2(b) (2019) ("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 will do the crime. Rosemond v. United States, 572 U.S. 65, 79-80 (2014) (clarifying mens rea for complicity).

198 *See* MODEL PENAL CODE § 2.02(2)(c)-(d) (AM. LAW INST. 1962) (defining recklessness and negligence).

199 *See id.* § 210.3(a) (recklessly causing death suffices for manslaughter).

200 *See id.* § 210.4 (negligent homicide).

201 *See, e.g.*, *id.* § 220.1(2) (reckless burning or exploding); *id.* § 220.2(2) (risking catastrophe); *id.* § 220.3 (criminal mischief).

202 *See* Hallevy, *supra* note 10, at 181-84.

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 that C2 was reasonably foreseeable.[203]

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.[204]

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.

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.[205]

---

[203] 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); *see also* United States v. Barnett, 667 F.2d 835, 841 (9th Cir. 1982) (adopting natural or probable consequences doctrine).

[204] Hallevy, *supra* note 10, at 184.

[205] 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

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.[206] But in this case, taking down the electrical grid and causing physical harm to human victims were assumed to be entirely unforeseeable even to a reasonable actor in the defendant's shoes.

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 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.[207] Suppose one breaks into a home one believes to be empty in order to steal artwork. Thus, one commits the base crime of burglary.[208] 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.[209] 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 one prominent 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).[210]

---

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?

206 *See, e.g.*, MODEL PENAL CODE § 2.03 (AM. LAW INST. 1962) (characterizing proximate or legal causation requirement using a "scope of the risk" test).

207 *See* WAYNE R. LAFAVE, 2 SUBSTANTIVE CRIMINAL LAW § 14.5 (3d 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").

208 *See* MODEL PENAL CODE § 221.1 (defining burglary).

209 *See* LAFAVE, *supra* note 207, § 14.5.

210 *See* A. P. Simester, *Is Strict Liability Always Wrong?*, *in* APPRAISING STRICT LIABILITY 21, 45 (A. P. Simester ed., 2005) (arguing that constructive liability as to a result is justified when the result is risked by the base offense) ("Where the risk [of Y] is intrinsic [to D's doing the base offense, X], there seems no difficulty about holding D responsible and culpable for Y."). 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 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 created to the extent there are not already existing crimes that fit this mold. Indeed, in the present example, one might think there are already some available constructive liability crimes. Perhaps felony murder fits the bill insofar as attempting to steal currency may be a felony, and the conduct subsequently 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 may be necessary. In any case, no AI criminal gap is present here because criminal law has familiar tools available for dealing with unforeseeable harms of this kind.

2. What the AI criminal gap <u>is</u>: 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 dark web and has it shipped to student housing. Suppose the programmers of this "Harvard Automated Shopper" did nothing criminal in designing the system and in fact 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.[211] Moreover, in the case of RDS, 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

[211] *See supra* notes 192–194 and accompanying text.

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, particularly as all activities and technologies involve some risks of some harm. This expansion of criminal law would stifle innovation and beneficial commercial activities. Indeed, if there were such a crime, most of the early developers of the internet would likely be guilty of it.[212]

B. The Costs of Punishing AI

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.[213] 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.[214] But how do we make sense of the question of mens rea for AI?

Part III 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.

---

[212] 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* notes 134–135 and accompanying text.

[213] *See supra* Part III.A (discussing the Eligibility Challenge).

[214] *E.g.*, Kenneth W. Simons, *Should the Model Penal Code's Mens Rea Provisions Be Amended?*, 1 OHIO ST. J. CRIM. L. 179, 195-96 (2003) ("The MPC views its four basic mental states or culpability terms as hierarchically ordered . . . .").

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 the voluntary act requirement, applicable to all crimes.[215] 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 program has the relevant behavioral dispositions associated with mens rea-like intention or knowledge. In Part III.A, 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 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.[216] 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.[217] Perhaps most famously, a 2017 report

---

[215] *See supra* notes 137–142 and accompanying text.

[216] *See* Thomas J. 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 individual persons," while for practical reasons the "legal fiction" of corporate personality — and later corporate crime — developed).

[217] *See, e.g.*, SAMIR CHOPRA & LAWRENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS (2011) (arguing that AI could and should be given legal personality in the near future); Asaro, *supra* note 104, at 169-86 (proposing a Turing test to decide if an AI agent that caused harm is legally fit to stand trial for a criminal offense); Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231

by the European Parliament called on the European Commission to create a legislative instrument to deal with "civil liability for damage caused by robots."[218] It further requested the Commission to consider "a specific legal status for robots," and "possibly applying electronic personality" as one solution to tort liability.[219] Even in such a speculative and tentative form this proposal proved highly controversial.[220]

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.[221] However, legal personality comes in many flavors, even for natural persons such as children who lack certain rights and obligations enjoyed by adults. Crucially, no *artificial person* enjoys all of the same rights and obligations as a natural person.[222] The best-known class of artificial persons, corporations, 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.[223] However, they do not receive protection under constitutional provisions, such as the Fourteenth Amendment's Equal Protection

(1992); Amanda Wurah, *We Hold These Truths to Be Self-Evident, That All Robots Are Created Equal*, 22 J. FUTURE STUD. 61 (2017).

[218] *Report with Recommendations to the Commission on Civil Law Rules on Robotics*, at 16 (Jan. 27, 2017), http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.pdf.

[219] *See id.* at 18.

[220] For instance, 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." *Open Letter to the European Commission Artificial Intelligence and Robotics*, ROBOTICS-OPENLETTER.EU, http://www.robotics-openletter.eu/ (last visited Oct. 13, 2019).

[221] 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. *See, e.g.*, Nicole M. Radziwill & Morgan C. Benton, *Evaluating Quality of Chatbots and Intelligent Conversational Agents*, (Apr. 15, 2017) (unpublished manuscript), https://arxiv.org/pdf/1704.04579.pdf. More generally, a lot of online content is generated by AI. *See id.*

[222] *See* S. M. Solaiman, *Legal Personality of Robots, Corporations, Idols and Chimpanzees: A Quest for Legitimacy*, 25 ARTIFICIAL INTELLIGENCE & L. 155 (2017).

[223] Elvia Arcelia Quintana Adriano, *The Natural Person, Legal Entity or Juridical Person and Juridical Personality*, 4 PENN. ST. J.L. & INT'L AFF. 363, 365 (2015). The first U.S. federal criminal conviction of a company was *United States v. N.Y. Cent. & Hudson River R.R. Co.*, 212 U.S. 509 (1909).

Clause, and they cannot bear arms, run for or hold public office, marry, or enjoy other fundamental rights that natural persons do.[224] 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 corporations. 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.[225] 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.[226] 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."[227] These outcomes may be more likely if AI were held accountable by the state in ways normally reserved 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.[228] 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*, or the tendency for an increasing number of rights to arise over time.[229] Even if AIs are given few or no rights initially when they are first granted legal personhood, they may gradually acquire rights as time progresses. Granting legal personhood to AI may thus be an important step down a slippery slope. In a 1933 Supreme

---

[224] *See* U.S. CONST., amend. XIV, § 1, cl. 2; Nw. Nat'l Life Ins. Co. v. Riggs, 203 U.S. 243, 255 (1906) ("The liberty referred to in [the 14th] Amendment is the liberty of natural, not artificial, persons."); *see also* Richard A. Epstein, *Of Citizens and Persons: Reconstructing the Privileges or Immunities Clause of the Fourteenth Amendment*, 1 N.Y.U. J.L. & LIBERTY 334, 341 (2005).

[225] *Cf.* Hu, *supra* note 15, at 527-28 (discussing whether recognizing legal personhood for "smart robots" would be harmful, and addressing a number of Ryan Calo's concerns about anthropomorphizing robots or other AI entities).

[226] *See* Jakub Zlotowski et al., *Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction*, INT'L J. SOC. ROBOTICS 347, 352 (2014).

[227] Damiano & Dumouchel, *supra* note 114, at 4.

[228] *Cf.* Diamantis, *supra* note 113, at 2078-80.

[229] *See* David S. Law & Mila Versteeg, *The Evolution and Ideology of Global Constitutionalism*, 99 CALIF. L. REV. 1163, 1170 (2011) (defining "rights creep").

Court opinion, for instance, Justice Brandeis warned about rights creep, and argued that granting corporations an excess of rights could allow them to dominate the State.[230] Eighty years after that decision, Justice Brandeis' concerns were prescient in light of recent Supreme Court jurisprudence such as *Citizens United* v. *Federal Election Commission* and *Burwell v. Hobby Lobby Stores*, which significantly expanded the rights extended to corporations.[231] Such rights, for corporations and AI, can restrict valuable human activities and freedoms.

C. Second Alternative: Minimally Extending Criminal Law

There are alternatives to direct AI punishment besides 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,[232] 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

[230] *See* Louis K. Liggett Co. v. Lee, 288 U.S. 517, 549 (1933) (Brandeis, J., dissenting).

[231] *See* Citizens United v. Fed. Election Comm'n, 558 U.S. 310, 341 (2010) (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.* at 341-43 ("The 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 Stores, Inc., 573 U.S. 682 (2014) (recognizing a for-profit company's claim to religious belief).

[232] *See* 18 U.S.C. § 1030(a) (2019).

to ex ante register a Responsible Person for the AI.[233] It could be a crime to design or operate AI capable of causing harm without designating a Responsible Person.[234] This would be akin to the offense of driving without a license.[235] 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 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 AIs have human developers, and in the event an AI 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. However, 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 extent that a non-commercial AI owner, developer, and user working together

[233] 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.

[234] 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 on co-conspirators to escape liability.

[235] 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 "fall guy." The SEC for this reason requires a C-level executive to attest to certain statements on filings.

would prefer a different responsibility arrangement, they might be permitted to agree to a different ex ante selection of the Responsible Person.[236] 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.[237]

It would be possible to impose criminal liability on the Responsible Person 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,"[238] though some examples of more serious strict criminal liability can also be found (such as statutory rape in some jurisdictions).[239] This could be defended by claiming that there is a special duty owed to society at large to provide *special assurances* that certain especially serious risks will be mitigated as much as possible.[240] A Responsible Person accepting

---

[236] 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.

[237] 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.

[238] *See* MODEL PENAL CODE § 2.05(1) (AM. LAW INST. 1962).

[239] *See, e.g.*, N.Y. PENAL LAW §§ 130.25-50 (2001) (defining statutory rape offenses); Funari v. City of Decatur, 563 So. 2d 54, 55 (Ala. Crim. App. 1990) (holding that an 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").

[240] *Cf.* DUFF, ANSWERING, *supra* note 91, 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").

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.[241] If justifiable 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.[242] 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 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 could 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.

---

[241] *See* Kenneth W. Simons, *When Is Strict Criminal Liability Just?*, 87 J. CRIM. L. & CRIMINOLOGY 1075, 1075-76 (1997) (discussing retributive views that denounce strict liability) ("Strict 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.").

[242] *See, e.g.*, Abbott, *The Reasonable Computer*, supra note 49.

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 frameworks come 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.[243] 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).[244] 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.[245] 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.[246]

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 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 criminal law).

A Responsible Person scheme is not the only solution to inadequate civil liability for Hard AI Crimes. An insurance scheme is another

---

[243] *See, e.g.*, *id.*

[244] *See, e.g.*, *id.*

[245] *See*, *e.g.*, *Register of Commission Expert Groups and Other Similar Entities: Group Details - Commission Expert Group*, EUROPEAN COMMISSION (Jun. 17, 2019), http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3592.

[246] *See* Abbott, *The Reasonable Computer*, *supra* note 49, at 15-16.

approach.[247] 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 where no appropriate Responsible Person exists. An AI compensation fund could operate like the National Vaccine Injury Compensation Program ("VICP").[248] Vaccines create widespread social benefits but are known in rare cases to cause serious medical 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 that is paid by users.[249] Other models for insurance schemes exist, such as the Price Anderson Act for nuclear power, which establishes a pool of funds to compensate victims in the event of a nuclear incident through a chain of indemnity regardless of who was ultimately at fault. [250]

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 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. Expanded civil liability could supplement this framework.

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

---

[247] 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 that 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).

[248] *See National Vaccine Injury Compensation Program*, HEALTH RESOURCES & SERV. ADMIN., https://www.hrsa.gov/vaccine-compensation/index.html (last visited Oct. 13, 2019).

[249] *See id.*

[250] *See The Price-Anderson Act*, BACKGROUND INFO. (Cent. For Nuclear Sci. & Tech. Info., La Grange Park, Ill.), Nov. 2005.

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.

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 overuse 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,[251] 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.

[251] *See* SARCH, *supra* note 164, at 141 (defending certain restricted uses of particular legal fictions based on culpably preserving ignorance).