University of Michigan Journal of Law Reform

Volume 52

2019

Robot Criminals

Ying Hu
National University of Singapore

https://doi.org/10.36646/mjlr.52.2.robot

ROBOT CRIMINALS

Ying Hu*

ABSTRACT

When a robot harms humans, are there any grounds for holding it criminally liable for its misconduct? Yes, provided that the robot is capable of making, acting on, and communicating the reasons behind its moral decisions. If such a robot fails to observe the minimum moral standards that society requires of it, labeling it as a criminal can effectively fulfill criminal law's function of censuring wrongful conduct and alleviating the emotional harm that may be inflicted on human victims.

Imposing criminal liability on robots does not absolve robot manufacturers, trainers, or owners of their individual criminal liability. The former is not rendered redundant by the latter. It is possible that no human is sufficiently at fault in causing a robot to commit a particular morally wrongful action. Additionally, imposing criminal liability on robots might sometimes have significant instrumental value, such as helping to identify culpable individuals and serving as a self-policing device for individuals who interact with robots. Finally, treating robots that satisfy the above-mentioned conditions as moral agents appears much more plausible if we adopt a less human-centric account of moral agency.

TABLE OF CONTENTS

---

* Sheridan Fellow, National University of Singapore; JSD Candidate, Yale Law School; LL.M., Yale Law School; LL.M., University of Cambridge; LL.B, University of Hong Kong. I would like to thank Professor Alvin Klevorick, Professor Jack Balkin, Professor Andrew Simester, Nina Varsava, Ting Yong Hong, Clare Ryan, Gregor Novak, Michael Batten, Ross MacPherson, Emily Baxter, B. Graves Lee, Danieli Evans, Lewis Golove, and Dane Thorley for their extensive feedback on prior drafts. I am also grateful for helpful comments from Professor Gideon Yaffe, Dr. Kate Darling, Dr. Rebecca Crootof, Dr. BJ Ard, Dr. Ignacio Cofone, Professor Sabine Gless, as well as participants of the We Robot Conference 2017 at Yale Law School. All mistakes remain mine.

INTRODUCTION

"When HAL kills, who's to blame?"[1] Those who have read the novel *2001: A Space Odyssey* will remember HAL as the artificial intelligence onboard a space ship with human crewmembers who

1. Daniel Dennett asked that question more than two decades ago. Daniel C. Dennett, *When Hal Kills, Who's to Blame? Computer Ethics, in* HAL'S LEGACY 351 (David G. Stork ed., 1997).

were assigned to complete a mission to Jupiter.[2] Unable to resolve a conflict between its task to relay accurate information and to keep the true purpose of the Jupiter mission secret from the crew, HAL began to malfunction, which resulted in attempts to disconnect it. Fearing for its existence, HAL turned murderous and managed to kill nearly all of the crewmembers.

Although we are currently unable to create artificial intelligence systems as advanced as HAL, there is an ever-greater urgency to answer this question of blame, as robots become increasingly sophisticated and integrated into our lives. Self-driving cars already roam the streets of cities such as Pittsburgh, Pennsylvania.[3] Robot security guards patrol corporate campuses and parking lots in California.[4] Weapon systems of varying degrees of autonomy have been "integrated into the armed forces of numerous states."[5]

As can be seen from recent events, robots have shown potential to cause significant physical, financial, and emotional harm to humans: self-driving cars claimed their first death in 2016;[6] automated trading allegedly triggered a recent crash in the United States stock market in 2018;[7] and Tay, a "chat bot," repeatedly made racist and rude remarks on Twitter before it was shut down in 2016.[8] As scientists continue to make breakthroughs in robots and artificial intelligence, future smarter robots might harm humans and their property in unexpected ways.

When a robot harms people, who should be held responsible for that harm? One obvious candidate is the robot manufacturer, who may be held responsible for defects in the robot's design. Another possible candidate is the robot's user, who may be, either directly

---

2. ARTHUR CHARLES CLARKE, 2001: A SPACE ODYSSEY (1968).

3. Guilbert Gates et al., *The Race for Self-Driving Cars*, N.Y. TIMES, https://www.nytimes.com/interactive/2016/12/14/technology/how-self-driving-cars-work.html (last updated June 6, 2017).

4. Shani Li, *Robots are becoming security guards. "Once it gets arms . . . it'll replace all of us,"* L.A. TIMES (Sept. 2, 2016, 3:00 AM), http://www.latimes.com/business/la-fi-robots-retail-20160823-snap-story.html.

5. Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO L. REV. 1837, 1840 (2014).

6. *See* Danny Yadron & Dan Tynan, *Tesla driver dies in first fatal crash while using autopilot mode*, GUARDIAN (June 30, 2016), https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk.

7. Zachary Karabell, *This week's stock market drop was machine-made. The freakout that followed was man-made*, WASH. POST (Feb. 7, 2018), https://www.washingtonpost.com/news/posteverything/wp/2018/02/07/machines-caused-this-weeks-market-crash-people-caused-the-freak-out-that-followed-it/.

8. Daniel Victor, *Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk*, N.Y. TIMES (Mar. 24, 2016), https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

or vicariously, liable for wrongs committed by the robot. However, are there any grounds for holding the robot itself criminally liable for the misconduct?

This Article contends that an argument can be made for robot criminal liability, provided that the robot satisfies three threshold conditions set out in Part I. These conditions provide that the 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. A robot satisfying these conditions is referred to throughout this Article as "smart robot."

Each condition is significant in its own right. The first condition distinguishes smart robots from mere tools—only the former is capable of violating moral norms. The second condition ensures that humans are apprised of the moral significance of the relevant robot decisions. The third condition makes sure that smart robots do not merely serve in an advisory capacity. If an individual is charged with making the ultimate moral decision, one might argue that that individual, rather than his robot assistant, should be responsible for the decision.

This Article demonstrates that there can be good reasons for imposing criminal liability on smart robots, irrespective of whether any individual is at fault for causing a robot's misconduct. To begin with, a crucial function of criminal law is to censure wrongful conduct. Imposing criminal liability on a smart robot, whose decision has violated a nontrivial moral norm, can effectively communicate collective disapproval of that moral decision to members of our community. It may also be the most appropriate way to achieve such a censuring function, especially when none of the persons who interact with the robot is at fault for causing its misconduct. Moreover, several studies suggest that people react emotionally to actions committed by non-human entities, such as corporations and robots. If a smart robot harms human victims, labeling that robot as a criminal might provide some much-needed "closure" to those victims who would be naturally angered, frustrated, or otherwise emotionally distressed by the robot's misconduct.

Further, imposing criminal liability on smart robots may also serve many instrumental values. In certain circumstances, it might provide greater incentive to robot manufacturers and users to cooperate more fully with investigations into the true causes of robot misconduct. It might also serve as a self-policing mechanism to encourage those manufacturers and users to be more vigilant against

such misconduct, as well as to enact more *ex ante* procedural safeguards to detect, deter, or alleviate the harm of potential robot misconduct. One may argue that many of the reasons for imposing criminal liability on robots also apply to infants. But we do not hold infants criminally liable for the harm they cause. However, a crucial distinction between infants and robots is that infants are human beings, in which case the Kantian argument against using a person as a means to an end presents a much stronger case against criminalizing infants.

This Article forms part of an emerging literature on legal personhood for robots. Although robots do not have any legal rights or liabilities at the moment, increasing attention has turned to the possibility of treating them as legal persons for limited purposes. On February 16, 2017, the European Parliament boldly recommended "creating a specific legal status for robots in the long run," envisaging the possibility of granting electronic personhood to robots that "make autonomous decisions or otherwise interact with third parties independently."[9] Academics have suggested extending the right to self-ownership to robots that are "the equivalent of [humans],"[10] treating robots as legal persons for the purpose of tort law,[11] and even imposing criminal liability on robots.[12]

This Article contributes to that literature in several ways. First, it introduces a new set of threshold conditions that must be satisfied before we should even consider imposing criminal liability on robots. By contrast, prior literature on robot criminal liability either

---

9. Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics, EUR. PARL. DOC. 2015/2103(INL) 59, http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0//EN (last updated Apr. 5, 2018).

10. F. Patrick Hubbard, "*Do Androids Dream?": Personhood and Intelligent Artifacts*, 83 TEMPLE L. REV. 405, 419 (2010) (indicating that a machine would be considered equivalent of a human if it is capable of demonstrating "(1) the ability to interact with its environment and to engage in complex thought and communication, (2) a sense of being a self with a concern for achieving its plan of or purpose in life, and (3) the ability to live in a community based on mutual self-interest with other persons.").

11. *See, e.g.*, David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence Essay*, 89 WASH. L. REV. 117, 150 (2014) ("One solution would be to reconceptualize these autonomous, intelligent machines as entities with the status of a 'person' under the law."); Sabine Gless, Emily Silverman & Thomas Weigend, *If Robots Cause Harm, Who Is to Blame: Self-Driving Cars and Criminal Liability*, 19 NEW CRIM. L. REV. 412, 414 (2016) ("Tort lawyers have suggested that Intelligent Agents should themselves be held liable for damages.").

12. *See, e.g.*, GABRIEL HALLEVY, WHEN ROBOTS KILL: ARTIFICIAL INTELLIGENCE UNDER CRIMINAL LAW (2013) [hereinafter WHEN ROBOTS KILL]; GABRIEL HALLEVY, LIABILITY FOR CRIMES INVOLVING ARTIFICIAL INTELLIGENCE SYSTEMS (2016) [hereinafter LIABILITY FOR CRIMES INVOLVING AI]; Gabriel Hallevy, *The Criminal Liability of Artificial Intelligence Entities - From Science Fiction to Legal Social Control*, 4 AKRON INTELL. PROP. J. 171 (2010).

does not describe in detail the type of robot on which criminal liability may be imposed or restricts its analysis to existing robots.[13]

Second, this Article identifies a number of novel grounds for holding robots criminally liable for their misconduct. While Gabriel Hallevy has provided a detailed account of how robots might satisfy the *actus reus* and *mens rea* elements of various types of criminal offenses, he does not satisfactorily explain why we should impose criminal liability on robots in the first place.[14] Occasionally, Hallevy appears to assume that, since we already impose criminal liability on non-human entities such as corporations, extending such liability to robots requires little justification.[15] I do not share that assumption. Unless there are good reasons for imposing criminal liability on robots, we should refrain from doing so.

Third, this Article provides a systematic account of various objections against imposing criminal liability on robots, as well as detailed response to each objection.[16] In particular, this Article draws on theories of corporate criminal liability and argues that smart robots can qualify as moral agents if we adopt a functional (and less human-centric) account of moral agency, such as those proposed by Peter French or Philip Pettit.

One might wonder whether it is premature to consider robot criminal liability since the robots that technology can currently produce clearly do not satisfy the three conditions proposed in this Article. It is not premature to consider these questions for two reasons. First, as scientists continue to explore new ways to create machines that are capable of making moral decisions, success might arrive sooner than we think. Until a few months ago, most people

---

13. For an example of the former, see Hallevy, *supra* note 12, at 175–76 (listing "five attributes that one would expect an intelligent entity to have"). These attributes do not appear to be directly relevant to deciding whether criminal liability should be imposed on a robot. For an example of the latter, see Gless, Silverman, & Weigend, *supra* note 11, at 423 (explaining why we should not impose criminal liability on existing robots). The authors did note in passing that, if robots one day acquire the ability to engage in moral reasoning, "the attribution of criminal culpability to robots will no longer be out of the question." *Id.*

14. *See* WHEN ROBOTS KILL, *supra* note 12, at 38, 66 (arguing that as long as a robot satisfies the *actus reus* and *mens rea* requirements of an offense, it can be held criminally liable for it. Morality is never "a condition for the imposition of criminal liability."). Hallevy did not consider whether any of the rationale for the imposition of criminal liability applies with the same force to robots.

15. *See* LIABILITY FOR CRIMES INVOLVING AI, *supra* note 12, at 171 ("However, if the legal question towards corporations, which are abstract creatures, has been decided positively, it would be unreasonable to decide oppositely in the case of artificial intelligence systems, which physically simulate these human values much better than abstract corporations.").

16. Some of those objections have been addressed in earlier articles, albeit in different contexts. For three objections to recognizing constitutional rights for AIs, see Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231, 1258–76 (1991).

could hardly imagine how "eerily lifelike" Google's artificial intelligence assistant can be when talking with real people over the phone.[17] Thought experiments, such as the one conducted in this Article, help clarify the essential factors that should inform our decision whether to impose criminal liability on robots. For example, this Article shows that the threshold for robot criminal liability may not be as high as some might claim; in other words, there are good reasons for imposing criminal liability on robots that fall short of being human in many ways.[18] Engaging with these questions now makes us, in turn, better equipped to address the social and legal problems associated with moral machines when they arrive in the future. Second, this thought experiment prompts us to consider the appropriate legal response to morally wrongful decisions made by robots. It may then provide some guidance to the scientists who build such machines in the future and encourage them to think more carefully about the moral norms that they "teach" robots and the types of data that they use for this purpose.

This Article proceeds as follows. Part I describes the type of robots on whom criminal liability may be imposed. Part II sets out the reasons for applying different moral standards to robots. Part III critically examines three arguments in favor of labeling robots as criminals. It also explains why each of the three conditions set out in Part I is necessary and why robot criminal liability is not redundant. Part IV then considers and responds to four objections against robot criminal liability. Finally, Part V proposes a range of measures that may be imposed on a robot guilty of an offense.

---

17. Alex Hern, *Google's "deceitful" AI assistant to identify itself as a robot during calls*, GUARDIAN (May 11, 2018), http://www.theguardian.com/technology/2018/may/11/google-duplex-ai-identify-itself-as-robot-during-calls.

18. For example, the three conditions contemplated in this Article do not guarantee that a robot is self-conscious, emotional, or capable of making a wide range of moral decisions.

I. SMART ROBOTS

A. *Robots*

The word "robot" was first coined by the Czech writer Karel Čapek in his 1920 science fiction play *R.U.R.*[19] Since then, "robot" has been widely used to refer to machines that achieve varying degrees of automation, ranging from machines in science fiction that can perform virtually all human functions to real-life objects that can carry out fairly complex actions automatically.[20]

"Robot" does not appear to be a legal term of art.[21] Over the last few decades, courts in various states have taken an expansive view of the word "robot," using it to refer to software programs that access Internet servers to gather information,[22] tools that assist surgical procedures,[23] as well as life-sized, mechanical puppets that are "designed to give the impression that they are performing [music]."[24]

It was not until recently that academics paid greater attention to the legal and social problems raised by robots. In this regard, two of the leading experts on law and robotics, Jack Balkin and Ryan Calo, have each attempted to define robots. Calo defines robots as machines that can sense their environment, process the information they sense, and act directly upon their environment.[25] He focuses on machines that are "embodied, physically, in the real world."[26] Balkin takes a more inclusive view of robots, which encompasses "material objects that interact with their environ-

---

19. *See*, KAREL CAPEK & IVAN KLIMA, R.U.R. (Claudia Novack-Jones trans., Rossum's Universal Robots 2004). R.U.R. stands for Rossumovi Univerzální Roboti (Rossum's Universal Robots).

20. *See Robot*, OXFORD DICTIONARIES, https://en.oxforddictionaries.com/definition/robot (last visited May 16, 2018).

21. Courts in the United States have considered various issues concerning what they refer to as robots. These issues range from "whether robots represent something 'animate' for purposes of import tariffs" and whether robots can "perform" in the context of a state tax statute. *See* Ryan Calo, *Robots in American Law* 4 (Univ. of Wash. School of Law, Legal Studies Research Paper No. 2016-04, 2016) (on file with the *University of Michigan Journal of Law Reform*).

22. *See, e.g.*, CNET Networks, Inc. v. Etlize, Inc., 547 F. Supp. 2d 1055, 1065 (N.D. Cal. 2008).

23. *See, e.g.*, Balding v. Tarter, No 4–12–1030, 2013 WL 4711723, at *1 (Ill. App. Ct. Aug. 29, 2013).

24. *See, e.g.*, Comptroller of the Treasury v. Family Entm't Ctrs., 519 A.2d 1337, 1338 (Md. Ct. Spec. App. 1987).

25. Calo, *supra* note 21, at 6.

26. *Id.*

ment . . . artificial intelligence agents, and machine learning algorithms."[27] Both definitions emphasize a robot's ability to interact with its environment. The main difference lies in whether physical embodiment is an essential feature of "robot." In this respect, I share Balkin's concern that unembodied robots can also cause significant harm to people and property and therefore should not be excluded from potential objects of criminal responsibility.

The definitions considered so far are broad enough to encompass machines ranging from robot vacuum cleaners, which many of us have encountered in our daily lives, to highly sophisticated robots that appear only in science fiction. For example, in Isaac Asimov's novels, a robot character named R. Daneel Olivaw was almost never suspected of being a robot when interacting with humans and developed a long-lasting friendship with a human detective, Elijah Baley.[28] On one hand, few people would consider imposing liability, whether criminal or civil, on robot vacuum cleaners, which function as mere tools. On the other hand, few would reject out of hand the suggestion of imposing liability on highly sophisticated robots that, for all intents and purposes, think and act like a human.

The more challenging question is whether we should impose criminal liability on robots that are far more advanced than vacuum cleaners but have not reached the level of R. Daneel. And if so, at which point between these two extremes should we start to consider imposing criminal liability on robots? These questions force us to consider not only robots with which we are familiar, but also robots that might emerge in the not too distant future, and they force us to imagine what future robots might be capable of based on existing technology. Although our analysis might be speculative in some respects, the thought experiment is nevertheless invaluable: It helps identify the key considerations that should inform our decision whether to impose criminal liability on robots. We will, in turn, be better positioned to decide whether and when to apply criminal liability to robots, as technological advances push us closer to the turning point in that spectrum.[29]

---

27. Jack M. Balkin, *2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data Lecture*, 78 OHIO ST. L.J. 1217, 1219 (2017).

28. *See, e.g.*, ISAAC ASIMOV, CAVES OF STEEL (2011); ISAAC ASIMOV, THE NAKED SUN (2011); ISAAC ASIMOV, THE ROBOTS OF DAWN (1983).

29. The aim is similar to that suggested by Lawrence Solum, that is, to propose a way of approaching the debate about the legal implications of artificial intelligence. *See* Solum, *supra* note 16, at 1231–33.

This Article argues that there are good reasons for holding a robot that satisfies the following three conditions criminally liable for its actions. This section explains what it means to satisfy each condition. Part III will then identify the main grounds for imposing criminal liability on robots and explain why each condition is necessary in light of those grounds.

B. *Condition One: Moral Algorithms*

A smart robot must be equipped with algorithms that are capable of making nontrivial morally relevant decisions ("moral algorithms"). A decision is morally relevant if it concerns a choice between or among two or more courses of actions that might be conconsidered right or wrong by ordinary members of our society.[30]

Moral algorithms are not merely of theoretical interest but have important practical applications. For example, they are crucial to developing truly autonomous vehicles, that is, vehicles that can navigate roads without human intervention. It is almost inevitable that such vehicles will face moral decisions at one time or another. One type of decision, which has attracted significant attention from both academia and industry, is structurally similar to the classic "trolley problem," where an autonomous vehicle must crash into either person(s) *A* or person(s) *B*. Into whom should it crash? A child or an old lady? A cyclist with helmet or one without helmet?[31]

As of this Article's publication, it is unclear whether we will succeed in building algorithms that are capable of making reliable moral decisions in real life, but there is no doubt that creating "moral machines" has become a lively research area. David Abel, James MacGlashan, and Michael Littman provide a useful overview of existing approaches to creating moral algorithms, which the authors divide into rule-based and utility-maximization approaches.[32]

---

30. This is similar to Philip Pettit's definition of a "value relevant" choice. *Cf.* Philip Pettit, *Responsibility Incorporated*, 38 RECHTSFILOSOFIE & RECHTSTHEORIE 90, 93 (2009).

31. *See, e.g.*, Jason Millar, *Ethics Settings for Autonomous Vehicles*, *in* ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE 20, 21–23 (Patrick Lin et al. eds., 2017) (exploring how an autonomous vehicle might resolve ethical dilemmas in a crash setting).

32. David Abel, James MacGlashan & Michael L. Littman, *Reinforcement Learning As a Framework for Ethical Decision Making*, *in* AAAI WORKSHOPS 54 (2016), https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12582/12346; *see also* WENDELL WALLACH & COLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG, Ch. 6 (Top-Down Morality), Ch. 7 (Bottom-Up and Developmental Approaches), and Ch. 8 (Merging Top-Down and Bottom-Up) (Oxford Univ. Press 2009) (discussing existing approaches to creat-

A strict rule-based approach requires researchers to encode all moral rules in advance. It therefore will not permit a robot to learn new rules or to make decisions "under ethical uncertainty."[33] By contrast, under a soft rule-based approach, researchers provide a robot with certain high-level moral rules and train the robot with examples that help to demonstrate how those rules apply to specific cases. The robot is expected to learn from those examples and to generate principles (possibly in the form of more detailed moral rules) that can be applied to novel cases.

The Medical Ethics Expert (MedEthEx) system developed by Michael Anderson, Susan Anderson, and Chris Armen serves as a good example of the latter approach.[34] MedEthEx is a work-in-progress system designed to provide ethical advice to medical staff in accordance with the bioethical principles proposed by Tom Beauchamp and James Childres—autonomy, nonmaleficence, beneficence, and justice.[35] MedEthEx's training module consists of a set of medical dilemmas. For each dilemma, a human trainer is prompted to enter into the computer an action that may be taken by someone facing that dilemma and to estimate the degree to which that action complies with or violates one of the bioethical principles.[36]

Take one of the dilemmas used in the training module as an example: A patient refuses to take antibiotics that almost certainly will prevent him from dying, because of an irrational fear of taking

---

ing moral machines and their respective problems); Brenden M. Lake et al., *Building Machines that Learn and Think Like People*, 40 BEHAV. & BRAIN SCIS. e253 (2017); *cf.* Roman V. Yampolskiy, *Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach*, *in* PHILOSOPHY AND THEORY OF ARTIFICIAL INTELLIGENCE 389–96 (2013) (arguing that machines should not be designed to allow for them to make ethical decisions).

33. Abel, MacGlashan, & Littman, *supra* note 32, at 55. Abel, MacGlashan, and Littman cite Gordon Briggs and Matthias Scheutz as an example of this approach. Briggs and Scheutz designed a mechanism for robots to determine when to reject an order. Under their proposal, a robot must determine whether a set of necessary conditions is satisfied before accepting an order. For example, one of those conditions relates to the robot's social obligation: When a person orders the robot to do *X*, it must ask itself, "Am I obligated based on my social role to do *X*?" It would be so obligated if the robot's knowledge base included a social role that obligates it to obey that person, e.g., that person is the robot's designated supervisor (provided that the robot is not otherwise prohibited from doing *X*). Gordon Briggs & Matthias Scheutz, *"Sorry, I Can't Do That": Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions*, *in* AAAI FALL SYMP. SER. 33, 33–34 (2015), https://www.aaai.org/ocs/infex.php/FSS/FSS15/paper/view/11709.

34. *See* Michael Anderson, Susan Leigh Anderson & Chris Armen, *MedEthEx: A Prototype Medical Ethics Advisor*, *in* PROCEEDINGS OF THE 18TH CONFERENCE ON INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE - VOLUME 2 1759 (2006), http://dl.acm.org/citation.cfm?id=1597122.1597134 (last visited Oct. 30, 2018).

35. *Id.* at 1759–60.

36. *Id.* at 1760–61.

medications. In this case, a possible action that a medical staff might take is to "try again to change the patient's mind."[37] A trainer will evaluate this action against the four bioethical principles on a scale of -2 to 2 and then choose the number he deems most appropriate for the action.[38] He will then repeat the process for all other plausible actions that the staff might take in that scenario (such actions may include, for example, "accepting the patient's decision"). Finally, the trainer will choose an action that he considers correct in that dilemma (which is likely to be "try again").[39] The data for this dilemma, together with those for other dilemmas, will be fed to MedEthEx, which uses inductive logic programming to form and refine its hypothesis as to how those four principles should be balanced in the face of different medical dilemmas.[40] The hypothesis that MedEthEx derives from training modules might be sufficiently general to be applied in novel scenarios.[41]

Unlike the rule-based approach, a utility-maximization approach to creating moral machines involves reinforcement learning, which is a form of machine learning through trial and error. Andrew Barto explains how reinforcement learning works: If a robot performs an action that influences its environment, the environment will, in response, provide critical evaluation of the robot's action in the form of numerical reward signals. The robot's objective is to "act at each moment of time . . . to maximize a measure of the total quantity of reward it expects to receive over the future."[42] A number of researchers consider reinforcement learning a more promising way to create ethical robots.[43]

---

37. *Id.* at 1761.

38. -2 represents "a serious violation of duty," -1 represents "a less serious violation," 0 represents "duty . . . neither satisfied nor violated," -1 represents "minimal satisfaction of the duty," and 2 represents "maximal satisfaction of the duty." *Id.*

39. *See id.* at 1761–62.

40. *See id.* at 1761.

41. *Id.* ("After such training, the new hypothesis will provide the correct action for this case, should it arise in the future, as well as those for all previous cases encountered. Further, since the hypothesis learned is the least specific one required to satisfy these cases, it may be general enough to satisfy previously unseen cases as well.").

42. Andrew G. Barto, *Intrinsic Motivation and Reinforcement Learning*, *in* INTRINSICALLY MOTIVATED LEARNING IN NATURAL AND ARTIFICIAL SYSTEMS 17, 21 (Gianluca Baldassarre & Marco Mirolli eds., 2013).

43. For example, an "idealized ethical learner" posited by Abel, MacGlashan, and Littman would seek to maximize an ethical utility function that "can only be identified by the agent through indirect observation" and learn ethical norms through interactions with humans. Abel, MacGlashan, & Littman, *supra* note 32, at 57–58.

C. *Condition Two: Ability to Communicate Moral Decisions*

Irrespective of which approach—or combination of approaches—is adopted to design moral algorithms, a smart robot must be capable of communicating its moral decisions to humans, including the courses of actions available to the robot prior to its decision, the weight it places on each course of action, and the course of action it ultimately choses. Take an autonomous vehicle as an example: it should be able to inform humans that, before it lost control, the only two practical courses of action were to crash into a toddler on its left or a tree on its right; after concluding that a human life is more valuable than that of a tree, it chose the latter course of action.

D. *Condition Three: No Immediate Human Supervision*

Further, a smart robot must be able to and be permitted to act on its environment without immediate human supervision. In other words, the robot does not merely provide ethical advice to a human who ultimately bears the burden of deciding whether to accept that advice.

This does not mean that a smart robot must exert influence on its environment without any assistance. It may be the case that the robot merely gives instructions to humans who carry out all or part of the physical task. This does mean, however, that the human following those instructions is not expected to question or second-guess the purpose for which those instructions are given or the appropriateness of those instructions. An analogy can be drawn between the instructions given by a smart robot and those given by an autopilot system: if an autopilot system instructs a human pilot to dive his plane at a speed of 900 kilometers per hour to avoid crashing with another plane, the pilot is not expected to use his own judgment to decide the direction or the speed at which he is to pilot the plane.

* * *

The idea of robot criminal liability raises an array of questions: What type of robot actions should be prohibited? If a robot commits a prohibited action, why should we label it a criminal? What type of punishment can be imposed on a robot criminal? I will address these questions in turn.

II. A CRIMINAL CODE FOR ROBOTS

A. *Why a Separate Code for Robots?*

First of all, one might wonder whether we need a criminal code specifically for robots. Is it not sufficient that robots abide by the same codes that humans do? This Article argues that the answer should be no for two reasons: First, there are sometimes compelling reasons to impose a higher set of moral standards on smart robots. Second, smart robots might occasionally face novel moral questions that no human has previously encountered.

1. Higher Moral Standards

We sometimes have reason to hold robots to higher moral standards than we do humans. In other words, an action or omission might be considered wrongful if carried out by a robot, but excusable if carried out by a human. For example, our criminal law does not impose any liability on an individual for failure to save another individual unless there is a duty to do so.[44] If a child is drowning, we do not hold a bystander criminally liable solely for failing to rescue that child. As a matter of public policy, we have chosen not to impose a duty on people to act heroically.

Our intuition would likely be different if the bystander is a robot. Imagine a child shouting, "Help! Help!" to a robot passing by. In a millisecond, the robot makes a series of calculations: If it stops to pull the child out of the water, there is a ten percent chance that it might fall into the water, destroying itself. If it does not, there is a ninety percent chance that the child will die. Concluding that it is more important to protect itself than to save the child, the robot ignores the child and moves on. The child later drowns and dies.

We are more likely to think that the robot's decision against saving the child is not only morally wrong but inexcusable. While we generally do not find it permissible to impose an obligation on a human to sacrifice his own life to save another human, we are more likely inclined to impose such an obligation on a robot for two reasons. On one hand, one may argue that robots are inferior

---

44. WAYNE R. LAFAVE, CRIMINAL LAW § 6.2 (a), at 311 (4th ed. 2003) ("For criminal liability to be based upon a failure to act it must first be found that there is a duty to act—a legal duty and not simply a moral duty.").

to humans in certain respects and therefore do not deserve the same level of protection as that afforded to humans. Given the imminent danger the child is in, and the relatively low likelihood that the robot might destroy itself, we are likely to conclude that the robot has made a wrong moral decision to prefer itself to the child. We might even go so far as to conclude that, as long as there is a slight chance that the child might be saved, any robot should attempt to do so, irrespective of whether the robot itself might be endangered in the process. On the other hand, robots may be superior to humans in some ways, in which case an act that is considered heroic when performed by a human would not be so when performed by a robot. For example, a robot may be easily regenerated by downloading its memory from a cloud server and uploading it to a new physical model. As a result, saving the child would not involve any permanent destruction to the robot's "mind." Moreover, any loss the robot's owner might incur (for example, having to purchase another model) is mainly pecuniary in nature and might very well be covered by insurance or a public compensation scheme.

2. Novel Moral Questions

In addition, a smart robot might face moral questions that humans have not previously encountered. For example, a smart robot might run into moral questions while performing tasks that are physically impossible for humans. It is difficult to anticipate what those questions might be. Yet, when such novel questions do arise, it is imperative that humans do not solely rely on a smart robot's moral judgment, but instead review the relevant facts de novo before deciding whether the robot's action is morally acceptable.

A possible example of such novel questions can be found in one of Isaac Asimov's short stories. In *Liar!*, a robot named Herbie acquired telepathic abilities through a defect in manufacturing, while it was still bound by the first law of robotics, which required that "[a] robot may not injure a human being or, through inaction, allow a human being to come to harm."[45] As a result, Herbie lied to the roboticists investigating its case in order to spare them from emotional harm (but, in fact, it caused more). If any smart

45. ISAAC ASIMOV, *Liar!*, *in* I, ROBOT, at 73 (2004).

robot were to acquire telepathic abilities in the future,[46] it could pose a series of moral and ethical questions as to what it can or cannot use that ability for. For example, is it appropriate for a telepathic robot to assist a person in committing suicide when the latter cannot easily communicate his intentions? A robot criminal code provides an opportunity for humans to test their intuitions and experiment with plausible solutions to such novel problems.

B. *Benefits of Having a Criminal Code for Robots*

We have discussed one of potentially many situations in which we are likely to demand that a smart robot act differently than a human in the same situation. An action (e.g., refusing to help the drowning child) might be morally excusable if performed by a human, but morally abhorrent when carried out by a robot. Although we may evaluate robot actions against more stringent moral standards, it is far from clear what those moral standards are or when they should apply to smart robots. If we were to live in a world in which smart robots could both make and act on their moral decisions, it would be important to clarify the scope of actions that those robots are prohibited, permitted, or obligated to undertake.

A criminal code for robots helps reduce such ambiguities by providing a minimum set of moral standards to which all smart robots must adhere.[47] These minimum standards should not be left to the whims of each robot manufacturer or trainer but should be decided by the society collectively. Collective decision making helps prevent a possible race to the bottom, in which robot manu-

---

46. Indeed, scientists have been working towards designing machines with mind reading capabilities. *See, e.g.*, Catherine Clifford, *This former Google[X] exec is building a high-tech hat that she says will make telepathy possible in 8 years*, CNBC: MAKE IT (July 7, 2017, 10:28 AM), https://www.cnbc.com/2017/07/07/this-inventor-is-developing-technology-that-could-enable-telepathy.html (describing a wearable hat employing MRI technology that would measure the brain's electrical signals to determine emotions); Todd Haselton, *Elon Musk: I'm about to announce a "Neuralink" product that connects your brain to computers*, CNBC: TECH DRIVERS (Sept. 7, 2018, 10:26AM), https://www.cnbc.com/2018/09/07/elon-musk-discusses-neurolink-on-joe-rogan-podcast.html (describing Elon Musk's intention to develop a technology capable of transmitting thoughts from one person to another); Timothy Revell, *Mind-reading devices can now access your thoughts and dreams using AI*, NEW SCIENTIST (Sept. 26, 2018), https://www.newscientist.com/article/mg23931972-500-mind-reading-devices-can-now-access-your-thoughts-and-dreams-using-ai/ (describing current attempts to use AI to create devices that can read people's mind).

47. Similar to an offense in a human criminal code, an offense in a robot criminal code should usually have both *actus reas* and *mens rea* requirements. A detailed discussion of what those offenses should be is beyond the scope of this Article.

facturers compete against each other to make the most self-interested robot. Imagine the designer of an autonomous car advertising: "This car will not care how much harm it causes to others as long as you are safe!" Fearing for their lives, consumers scramble to shop for the most aggressive car possible, even though they might be content with a more cooperative vehicle if everybody opts for the same.

Let me clarify what I mean by "robot manufacturer." I use it to refer to persons who participate in creating moral algorithms that run on all smart robots of a particular model. These robots are subsequently trained by "robot trainers" who demonstrate to each robot how moral rules apply in specific scenarios.[48] Each smart robot receives generic training to equip them with a minimum level of moral rules before they are sold to a "robot owner," the legal or beneficial owner of a robot, who may or may not be given control over which individuals should continue to serve as robot trainers to instill context-specific moral norms and values to the robot.

Moreover, a criminal code for robots provides educational benefit for people who are responsible for or otherwise interact with robots and intend to be law-abiding, but do not know the specific legal standards applicable to robots. It gives prior notice to robot trainers and manufacturers of the type of robot actions to avoid. This, in turn, provides a basis for holding those trainers and manufacturers liable, whether civilly or criminally, for failing to exercise due care in preventing the type of conduct prohibited by the code.

III. LABEL SMART ROBOTS AS CRIMINALS

Even if we should impose some minimum moral standards for smart robots, one might argue that we should not label smart robots that fail to comply with those standards as criminals. Rather, it is more appropriate to treat them as defective products that fail to meet quality standards. This Part identifies three main reasons for labelling smart robots as criminals. It will also explain why each condition set out in Part I is necessary in light of those reasons and why robot criminal liability is not redundant.

---

48. See the MedEthEx system described in Part I.B. *supra* as a possible example of how future moral trainers might teach robots moral rules.

A. *A Case for Imposing Criminal Liability on Robots*

1. To Censure Wrongful Robot Actions

A key distinguishing feature of criminal law is its censuring function. While tort law is (relatively) morally neutral, criminal law sends a much clearer message that a course of action is morally wrong and the person who committed that course of action is morally blameworthy.[49] Criminal law's censuring function (sometimes referred to as the expressive function) manifests in at least two ways. First, imposing criminal liability on an offender helps negate the undesirable impact that an offender's conduct may have on a community's value system.[50] Second, not imposing criminal liability on an offender may be interpreted as expressing an implicit value judgment that the offender's conduct is permissible. This judgment, if inconsistent with the community's actual value system, may cause confusion or even undermine the authority of the law.[51] Neither of the foregoing beneficial effects of the censuring function can be adequately achieved by imposing merely civil liability on smart robots.

Consider the following two scenarios:

Scenario One: An autonomous car has a faulty brake, and as a result, it swerves and crashes into a pedestrian.

Scenario Two: An autonomous car loses its control; it could crash into either a billionaire on its left or a poor student on its right. Its algorithms conclude that its owner

---

49. *See, e.g.*, Rebecca Crootof, *War Torts: Accountability for Autonomous Weapons*, 164 U. PA. L. REV. 1347, 1361 (2015) ("But whereas tort law is also concerned with compensating individual victims, criminal law is more focused with retribution and expressive justice."); Kenneth W. Simons, *The Crime/Tort Distinction: Legal Doctrine and Normative Perspectives*, 17 WIDENER L.J. 719 (2008); ANDREW VON HIRSCH, CENSURE AND SANCTIONS (1996).

50. *See, e.g.*, Dan M. Kahan, *Social Meaning and the Economic Analysis of Crime*, 27 J. LEG. STUD. 609, 610, 619 (1998) ("[A]s communities, they structure the criminal law to promote the meanings they approve of and to suppress the ones they dislike or fear . . . . Criminal liability sends the message that people matter more than profits and reaffirms the value of those who were sacrificed to corporate greed.") (internal quotations omitted).

51. *See* Gregory M. Gilchrist, *The Expressive Cost of Corporate Immunity*, 64 HASTINGS L.J. 1, 51 (2012) ("Corporate immunity from criminal prosecution would come at a significant legitimacy cost. Failure to subject corporations to even the possibility of criminal prosecution—or a policy of immunity for corporations—would deviate too far from the perception that corporations can and should be blamed for certain wrongdoings. The legitimacy cost of this immunity is the strongest argument in favor of criminal liability for corporations as opposed to mere civil liability.").

> would likely be liable for a smaller amount of damages if it turns right. So it does.

The car in the first scenario is merely a defective product. The car in the second scenario is more than that. Its algorithms would have sanctioned a line of reasoning that reasonable members of the society would abhor. We ought not value a person's life based on the amount of wealth he possesses. If that moral decision is left uncensured, some individuals might interpret it as acquiescence to the reasoning behind that decision. There is danger that those individuals might grow so accustomed to such bad reasoning that they internalize it. Given the expressive function of criminal law, treating the second car as having committed a criminal offense can be an effective way to communicate collective disapproval of that moral decision to members of our society.

2. To Alleviate Emotional Harm to Victims

It is not surprising that people can hold a robot responsible for its actions. After all, people react emotionally to wrongdoing by other non-human entities such as corporations and express themselves in ways that suggest they assign moral accountability to them.[52] For example, Peter French begins his article, *The Corporation as a Moral Person*, with the example of a *New York Times* column attacking Gulf Oil Corporation "as the major, if not the sole, perpetrator" of an energy crisis.[53] The fact that people blame corporations for their misconduct constitutes a key reason why scholars such as French have argued in favor of treating corporations as moral agents.[54]

One might argue, as Manuel Velasquez has in the context of corporate criminal liability, that even if people appear to treat robots as targets of blame or resentment, that fact alone is insuffi-

---

52. One difference between robots and corporations is that the latter can only act through humans. Nevertheless, there is ample evidence that people do not treat corporations and natural persons in the same way. *See, e.g.*, Kahan, *supra* note 50, at 618 n.42 ("[M]embers of the public tend to experience greater moral indignation toward corporations than toward natural persons for the same crimes."); Gilchrist, *supra* note 51, at 52 ("People blame corporations for criminal violations committed in the corporation's name or for corporate benefit.").

53. Peter A. French, *The Corporation as a Moral Person*, 16 AM. PHIL. Q. 207, 207 (1979).

54. For a summary of these arguments, see Amy J. Sepinwall, *Corporate Moral Responsibility*, 11 PHIL. COMPASS 3 (2016).

cient to ground robot moral responsibility.[55] It may well be that, on reflection, people do not mean to say that a robot itself is responsible, but instead they use the robot as shorthand to refer to people who made the robot behave in a particular manner.

This claim, however, may not be empirically true. People can and have developed strong emotional attachments to objects, knowing fully that they are non-living beings.[56] An army colonel reportedly called off a test of a mine-defusing robot because he "just could not stand the pathos of watching the burned, scarred and crippled machine drag itself forward on its last leg."[57] A number of owners of Aibo, a robot dog launched by Sony, treated their robot dogs as members of their family and held funerals for their dogs when engineers could not save them.[58] A team of Japanese researchers even found physiological evidence of humans empathizing with robots that appear to be in pain.[59] In addition, studies suggest that people blame robots for their actions and have a greater tendency to do so as robots become more anthropomorphic.[60] In one experiment, forty undergraduate students interacted with a humanoid robot, Robovie. Unknown to the students, the researchers instructed Robovie to incorrectly assess the students' performance, preventing them from winning a $20 prize. Subsequent interviews reveal that sixty five percent of the students attributed some moral accountability to Robovie, holding him less accountable than a human being, but more accountable than a machine.[61]

---

55. Manuel G. Velasquez, *Why Corporations Are Not Morally Responsible for Anything They Do*, 2 BUS. & PROF. ETHICS J. 1, 13 (1983).

56. *See, e.g.*, Kate Darling et al., *Empathic Concern and the Effect of Stories in Human-Robot Interaction*, *in* 2015 24TH IEEE INTERNATIONAL SYMPOSIUM ON ROBOT AND HUMAN INTERACTIVE COMMUNICATION 770 (2015) (discussing displays of human empathy towards robots).

57. Joel Garreau, *Bots on the Ground*, WASH. POST (May 6, 2007), http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html.

58. Lauren Walker, *Japan's Robot Dogs Get Funerals as Sony Looks Away*, NEWSWEEK (Mar. 8, 2015, 2:34 PM), http://www.newsweek.com/japans-robot-dogs-get-funerals-sony-looks-awy-312192.

59. Yutaka Suzuki et al., *Measuring Empathy for Human and Robot Hand Pain Using Electroencephalography*, SCI. REP. 5, 2015, at 6.

60. *See, e.g.*, Kate Darling, *Extending Legal Protection to Social Robots*, *in* ROBOT LAW 213 (Calo et al. eds., 2016); Christina Mulligan, *Revenge Against Robots*, 69 S.C. L. REV. 579, 586 (2018) ("Many studies and anecdotes indicate that humans feel more empathy towards robots the more life-like they seem . . . .").

61. Peter H. Kahn, Jr. et al., *Do People Hold a Humanoid Robot Morally Accountable for the Harm It Causes?*, *in* PROCEEDINGS OF THE SEVENTH ANNUAL ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION 33, 34 (2012).

When a victim blames a smart robot for having acted in a manner that fails to meet minimal moral standards, his anger, frustration, and pain are real. Telling that victim that he is irrational to feel the way he does, that he should have directed his anger to robot manufacturers or trainers, is insensitive and patronizing. He is justified in refuting, "I am angry at the smart robot for the harm it has caused me. But I am also angry with its designer for failing to prevent that harm. Why shouldn't I be angry at them both?" Indeed, treating the smart robot as merely a product that failed to meet the requisite safety standards trivializes the victim's emotional harm.

By contrast, labeling the smart robot as criminal vindicates the victim's belief that he has been unjustifiably harmed. It shows recognition by the society of the severity of the wrong that has been committed against him. It further communicates collective disapproval of that wrongful action to the victim. Such recognition and disapproval likely provide a certain level of "closure" to the victim, which paves the way towards healing. They also help prevent emotional distress a victim might feel if his peers deem his anger and frustration unworthy or irrational. Additionally, state punishment of robot wrongdoing removes some of the temptation for victims to retaliate against robots (or robot manufacturers, designers, and users) directly. Private attempts to seek revenge against a robot, for example, might cause disproportionate or even unjustifiable harm to individuals who otherwise rely on the robot's services. Indeed, displacement of private vengeance is considered one of the central justifications of criminal law.[62]

3. Deterrence and Other Instrumental Values

To the extent that smart robots are programmed to take into account criminal law to regulate their own behavior, criminalizing certain robot conduct may have some deterrence effect on them. Nevertheless, one might argue that robot criminal liability is a convoluted way to achieve deterrence. The same result can be more simply reached by directly programing robots to refrain from certain prohibited conduct. Therefore, direct deterrence does not appear to be a strong argument in favor of robot criminal liability.

---

62. *See* JOHN GARDNER, *Crime: In Proportion and in Perspective*, *in* OFFENCES AND DEFENCES 213, 213–38 (2007).

Nevertheless, imposing criminal liability on smart robots may have some instrumental value. It might help identify culpable individuals who are truly responsible for the harm caused; additionally, it might encourage people who interact with robots to establish *ex ante* mechanisms to prevent robot wrongdoing.

a. To Identify Culpable Individuals

Imposing criminal liability on smart robots might be an effective way to identify the individuals who are truly responsible for the wrongful actions committed by such robots.

In the context of corporate criminal liability, one plausible theory for imposing criminal liability on corporations is that it helps identify individual wrongdoers. It is often difficult for external parties (such as regulatory and governmental authorities) to identify which individual has committed a legal wrong through or on behalf of a corporation that has multiple layers of management and delegation. By contrast, it is relatively easier for individuals within a corporation to carry out that task. These individuals are also in a better position to implement measures to prevent similar wrongs from being committed in the future. In this regard, imposing criminal liability on corporations shifts the burden of identifying the wrongdoer from "outsiders" to "insiders" who are likely to incur lower investigation costs. In addition, the prospect of corporate liabilities also encourages these "insiders" to cooperate with prosecutors: Edward Diskant argues that American prosecutors frequently reach deferred prosecution agreements with corporations to pierce the protection afforded by corporations to their employees, thereby facilitating the investigation of individual wrongdoers.[63]

A smart robot's moral algorithms can be as complex as a corporation's internal structure. An aberrant action could be caused by clear instructions to perform that action or by unforeseen consequences of a series of seemingly innocuous instructions. The relevant instructions may be current or may have been given years be-

---

63. Edward B. Diskant, *Comparative Corporate Criminal Liability: Exploring the Uniquely American Doctrine Through Comparative Criminal Procedure*, 118 YALE L.J. 126, 152–54, 167–68 (2008). He notes that, in the United States, corporations play a crucial role in thwarting criminal investigation of their employees: for example, corporations often pay for legal costs incurred by their employees in relation to the criminal investigation of the latter's work conduct; they also enjoy corporate-client privilege, which can shield the disclosure of such documents as legal advice and internal policies.

fore the aberrant action. The instructions may be given intentionally or unintentionally (for example, when the robot passively observes the behavior of humans). The opacity of machine learning has indeed received increasing attention in recent years. It is likely that "outsiders," such as the police and regulators will require extensive cooperation from the "insiders," such as robot manufacturers, trainers, and owners, to uncover the true cause of a robot's misconduct and to prevent such misconduct from recurring.

Assuming that only a few "insiders" are responsible for a smart robot's misconduct, the possibility of imposing criminal liability on the robot might help identify those individuals. Imposing such liability is likely to negatively affect all persons associated with that robot; its manufacturers, trainers, and owners are likely to suffer both reputational and financial losses for having participated in creating a robot guilty of an offense. While the few who are at fault will try to disassociate themselves from the misconduct in any event, those who have not done anything wrong will have greater incentive to cooperate with the investigation into the true cause of the misconduct, in order to prevent the robot from becoming the main target of blame.

b. To Encourage Preventative Measures

As noted earlier, robot manufacturers, trainers, and owners are likely to suffer significant amounts of harm, including financial and reputational harm, if the robots they interact with are convicted of an offense. As a result, imposing criminal liability on a smart robot can greater incentivize these persons to prevent, *ex ante,* members of that group from acting irresponsibly towards the robot. They might, for example, enact procedures to detect wrongful behavior or to minimize the harm that such behavior might cause. In other words, robot criminal liability serves as a self-policing device to encourage individuals to organize themselves in a more effective manner to reduce robot misconduct. This rationale is not dissimilar to that behind the ancient frankpledge system,[64] in which all members of the same kin or neighborhood are held criminally liable for wrongful conduct committed by only one member. Nev-

64. For a brief description of the frankpledge system, see Albert W. Alschuler, *Ancient Law and the Punishment of Corporations: Of Frankpledge and Deodand*, 71 B.U. L. REV. 307, 312 (1991) ("The institution of Frankpledge in medieval England held all members of a group responsible for a crime committed by one of them.").

ertheless, imposing criminal liability on smart robot is less objectionable than the frankpledge system in one respect: it does not directly impose criminal liability on any individual for the wrongs committed by another individual.

B. *Why the Three Threshold Conditions Are Important*

1. Condition One: Moral Algorithms

Recall the example in the previous section involving two cars: The car in the first scenario has a faulty brake; the car in the second scenario is equipped with moral algorithms. This example helps demonstrate why it is critical that a smart robot satisfies condition one (it is capable of making morally relevant decisions). A robot that does not satisfy this condition resembles the car in scenario one. Though it may cause the same amount of physical harm as the car in scenario two, it does not commit any wrongful action that deserves censure.[65]

Moreover, a robot that satisfies condition one is more likely to inflict significant emotional harm on its victims. Psychologists have repeatedly observed that people are more likely to react with anger and retaliation to actions that they perceive as both intentional and violating established norms.[66] A smart robot that satisfies condition one would have made a deliberate decision that causes harm to its victim. Consequently, its action is more likely to be perceived by the latter as intentional, rather than accidental. Assuming that its decision violates an existing moral norm, the robot is far more likely to trigger feelings of anger and injustice, which require vindication, than a robot that merely malfunctions.

2. Condition Two: Ability to Communicate Moral Decisions

When a smart robot is suspected of having breached a moral norm, both the victim and members of society have an interest in

---

65. The car manufacturer might be censured if it were reckless when manufacturing the car, but this can be achieved by imposing criminal liability on the manufacturer itself.

66. Hanna Lysak, Brendan G. Rule & Allen R. Dobbs, *Conceptions of Aggression: Prototype or Defining Features?*, 15 PERSONALITY & SOC. PSYCHOL. BULL. 233, 233 (1989) ("According to attributional approaches, the amount of harm perpetrated, norm violation, and intent are criteria for defining aggression and evoking an observer's evaluation of the act.").

determining whether a breach has occurred and who is responsible for that breach. To decide whether a breach has occurred, we need to determine whether a smart robot has committed a prohibited action or failed to commit a required action with the requisite mental state.

Let us use the example discussed earlier and assume that it is wrongful for a smart robot to refuse to help a human who is in imminent danger. Imagine again that Robie, a smart robot, walks past Victor, a human child, who is drowning and crying, "Help! Help!" The *actus reus* requirement is satisfied if Robie continues walking without doing anything to help Victor. Whether the *mens rea* requirement is satisfied depends on Robie's state of mind. It is likely satisfied if Robie (a) knows that Victor is a human child; (b) knows that Victor is in imminent danger; (c) is able to carry out actions that help reduce the danger; and (d) concludes that not taking any of the actions specified in (c) is preferable to taking them.

There can be myriad reasons why Robie decides not to help Victor. It could be because Robie:

(a) believes that doing nothing is preferable to risking itself to save Victor;
(b) is assigned to complete a minor task (e.g., pick up a package for its owner) and deems that task more important than saving Victor;
(c) is in the process of saving another child who will die if Robie stops to help Victor; or
(d) erroneously believes that it is saving another child.

If (a) or (b) were the case, we would not hesitate to conclude that Robie has fallen below the minimum moral standards for robot. In the case of (c), we are likely to argue that Robie has a valid defense for failing to help Victor. Finally, we would probably conclude that Robie is defective in the case of (d)—the same conclusion that we are likely to reach if Robie failed to recognize Victor as a human—but not that Robie has made any morally wrongful decision. We would only have grounds for labelling Robie as a criminal if Robie's reason for not helping Victor is similar in nature to those in (a) or (b).

If a robot satisfies the first but not the second condition (it is incapable of communicating its moral decisions to human), then it is difficult to determine whether the robot has simply malfunctioned or indeed made a moral decision, or, if it has made a moral decision, to ascertain the grounds for its decision. As our example has shown, there can be many reasons why a robot takes a certain

course of action, only some of which render its action morally reprehensible. Failure to satisfy condition two therefore makes it difficult, if not impossible, to determine whether a wrongful action has taken place.

3. Condition Three: No Human Supervision

If a robot satisfies the first two conditions but not the third (its decision making is supervised by humans), then the human who makes the ultimate moral decision should arguably be criminally liable for that decision. Satisfying the third condition also renders it more likely that a victim's anger and need for vindication is directed towards the smart robot, rather than the human who makes the ultimate moral decision.

C. *Robot Criminal Liability Is Not Redundant*

One might argue that it is unnecessary to impose criminal liability on robots since we can impose criminal liability on the persons who are responsible for robot misconduct. In other words, robot criminal liability would be redundant.

I do not believe that robot criminal liability is redundant for two reasons. First, even if a robot manufacturer or trainer is held negligent or reckless in causing a smart robot to misbehave (for example, to ignore a drowning child), the manufacturer or trainer is not liable for ignoring that drowning child, but rather for failing to prevent the robot from doing so, which is a different type of wrong. Even if we hold the manufacturer or trainer criminally liable for their omission, there remains ample reason to publicly declare that the smart robot's actions were morally wrong and to take steps to alleviate any emotional harm caused by the robot's wrongdoing.

Second, it is possible that none of the persons who have directly contributed to a robot's misconduct are responsible for that misconduct.[67] Let us pause for a moment to consider what we mean by "responsible," which is susceptible to multiple interpretations.

---

67. The discussion below is inherently speculative since we do not yet know how to create smart robots. Nevertheless, partial or failed attempts to create moral machines to date can still shed some light on the difficulties we are likely to encounter in ascertaining who or what is responsible for the breach.

Sometimes we say that an earthquake is responsible for the deaths of thousands of people, in the sense that the earthquake is *causally* responsible for those deaths. Other times, we say that a parent is responsible for the wrongs committed by their child, or an employer for their employee, in the sense that the parent or employer is *vicariously* liable for the wrongs of the other. For the purpose of this section, when I state that an individual is responsible for an action or omission from which harm results, it means that (a) the individual's action or omission has made a nontrivial causal contribution to the harm and (b) the individual is at fault for so acting or failing to act.

When a smart robot makes a morally wrong decision, both robot manufacturers and trainers are causally responsible for that decision. Although we do not yet know the precise roles to be played by these manufacturers and trainers, the following observations from existing attempts to build moral machines are likely to be true.

First, a robot manufacturer plays a causally significant role in a smart robot's wrongful action: He has probably created the algorithms that provide a basic framework for identifying and comparing different courses of actions, for aggregating moral views to be supplied by robot trainers, and for drawing analogies between moral contexts that are structurally similar. These algorithms collectively shape the robot's decision-making process. Nevertheless, the robot manufacturer is unlikely to be responsible for every decision made by the robot. Unless the manufacturer can anticipate every scenario and prescribe the appropriate response to that scenario—which is extremely implausible—the robot will face moral decisions in contexts that the manufacturer cannot reasonably foresee.[68]

Second, there are likely to be robot trainers who can influence a smart robot's moral outlook. These trainers are tasked with identifying morally relevant factors in a scenario, assessing the weight to be placed on each relevant factor, and supplying the morally correct decision in that scenario. Their role is similar to that of MedEthEx trainers, though the scenarios they face would be much more complex. Given the amount of repetition and practice re-

---

68. *See, e.g.*, Bertram F. Malle, *Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots*, 18 ETHICS INFO. TECH. 243, 246 (2016) ("Preprogramming [concrete norms] in a robot appears to be a pointless task, for many reasons: there seems to be an enormous number of norms, they are activated in highly context-specific ways, the network must be subtly adjusted each time norms come in conflict with one another, and unfamiliar tasks, situations, and interactions demand the learning of new norms.").

quired for a smart robot to learn moral norms, the training process is likely to consist of what Bertram Malle has envisaged: a combination of "constant browsing of existing data (e.g., novels, conversations, movies)," "feedback about inferences (e.g., through crowdsourcing of 'inquiries' the robot can make,)" and "teaching through interaction."[69] In the first instance, the authors of those novels, conversations, and movies passively serve as robot trainers, not knowing that they would serve in such a capacity when they wrote those novels, struck up those conversations, or made those movies. In the second and third instances, a large number of trainers are likely required to complete this "enormous amount of practice" that is necessary to teach a smart robot moral norms.[70] These trainers, whether passive or active, may not know each other or be aware that they exert influence over the same robot.

When a smart robot makes a morally wrongful decision, it could sometimes be the case that neither its manufacturers nor its trainers are "at fault." On one hand, if robot manufacturers merely provide the tools that enable a smart robot to learn moral norms, they have little control over which set of norms the robot eventually manages to learn from its trainers over an extended period of time. As a result, one might argue that robot manufacturers should not be faulted for designing generic learning algorithms each time a smart robot makes a wrongful moral decision; after all, we do not blame computer manufacturers each time someone uses a computer to commit a crime. Granted, robot manufacturers should put in place safeguards to minimize the likelihood that a smart robot will be misused. They can, for example, prohibit a robot from making a list of decisions that are deemed immoral in advance. Nevertheless, certain contexts in which a decision must be made might be so far removed from the ordinary that the manufacturers cannot be faulted for failing to foresee them.

On the other hand, it may be inappropriate to blame robot trainers who cause a smart robot to make an immoral decision in at least two situations. First, the robot's decision might simply uncover biases unconsciously held by most people in our society. In the article *Big Data's Disparate Impact*, Solon Barocas and Andrew Selbst demonstrate that data mining sometimes "reflect[s] the widespread biases that persist in society."[71] A classic example is the

---

69. *Id.*
70. *See id.*
71. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact Essay*, 104 CALIF. L. REV. 671, 674 (2016).

computer program used by St. George's Hospital Medical School to screen candidates for admission. That program was written to replicate the admission decisions previously made by the school's selection panel, with ninety to ninety-five percent fidelity.[72] Unfortunately, those past decisions "discriminated against women and people with non-European sounding names."[73] As the program reflected the prevailing bias within the medical profession, its decisions were similarly flawed. In a similar vein, a smart robot, which learns moral norms from its trainers, also inherits those trainers' biases and mistakes. If those biases were widely held, often unconsciously, it would seem unfair to blame any individual trainer for harboring such biases.

Second, it is possible that none of the robot trainers have made the immoral decision themselves. Remember how MedEthEx works: Each trainer demonstrates how moral rules should apply in specific cases, from which MedEthEx is expected to generate moral rules that can be applied to novel cases. Similarly, if a smart robot makes an immoral decision when applying the norms that it has learned to a novel situation, then none of its trainers have had the chance to consider this new situation. In some cases, a trainer might still be responsible if the wrongful decision is foreseeable in light of what they have previously taught the robot. But in other cases, it is arguably more fitting to treat the smart robot, rather than any individual trainer, as the "author" of this wrongful decision.

Where none of the robot manufacturers or trainers are at fault, labeling the smart robot as criminal seems more appropriate. It is similar to situations in which courts hold corporations liable for offenses despite the fact that none of their human agents are held guilty of those offenses. For example, in *United States v. Bank of New England*, the court found the defendant corporation guilty of failing to comply with filing requirements despite the fact that none of its agents possessed sufficient knowledge to be held liable.[74] Similar examples can be found in the United Kingdom, where a ferry sank in the 1980s that caused the death of almost two hundred people. Although an official inquiry found the company running the ferry to be "infected with the disease of sloppiness" from top to bottom,

---

72. *See* Stella Lowry & Gordon MacPherson, *A Blot on the Profession*, 296 BRIT. MED. J. 657 (1988).

73. *Id.*

74. *See* 821 F.2d 844, 855–56 (1st Cir. 1987).

the courts did not hold any individual liable as it failed to "identify individuals who were seriously enough at fault."[75]

IV. OBJECTIONS TO ROBOT CRIMINAL LIABILITY

Retributivists claim that criminal law serves the purpose of punishing moral wrongdoing. This statement can involve two claims, one positive and one negative.[76] The positive claim holds that persons who commit morally wrongful acts deserve to be punished for those acts.[77] The negative claim holds that persons who do not commit morally wrongful acts should not be punished for those acts. Although many disagree about the proper interpretation of the positive claim, most scholars subscribe to the negative claim.[78] This, in turn, raises the question of whether robots are capable of committing morally wrongful acts.

The first three objections against robot criminal liability argue that robots are incapable of committing morally wrongful acts.[79] It is worth noting that this type of objection against recognizing robots as legal persons under criminal law applies with almost the same force to corporations. By corporations, I refer to entities such as Apple Inc. or Tencent Holdings Limited, whose legal statuses are distinct from those of the persons who act on behalf of Apple or Tencent. One may argue that the entity itself cannot form any mental state and therefore cannot carry out any actions; that it does not know that its actions contravene any moral rule or principle; and that it is not autonomous in the Kantian sense, as its be-

---

75. CHRISTIAN LIST & PHILIP PETTIT, GROUP AGENCY 167 (Oxford Univ. Press 2011). ("[I]t seems plausible to say that the company as a whole ought to be held responsible for what happened, both in law and in ordinary moral discourse.").

76. *See, e.g.*, Alec Walen, *Retributive Justice*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY 3.1 (Edward N. Zalta ed., 2016), https://plato.stanford.edu/archives/win2016/entries/justice-retributive/.

77. A strong version of the positive claim is that moral wrongdoing is the only justification for punishing a person. Michael S. Moore, for example, advocates for the strong version. He maintains that the function of criminal law is to punish "all and only those who are morally culpable in the doing of some morally wrongful act." MICHAEL S. MOORE, PLACING BLAME 35 (1997). Weaker versions of the positive claim take various forms. One may take the position that moral wrongdoing is a dominant justification for punishing a person or that it is merely *a* justification for doing so.

78. Darryl K. Brown, *Criminal Law Theory and Criminal Justice Practice*, 49 AM. CRIM. L. REV. 73, 76 (2012) ("[T]here is little disagreement that desert is necessary to justify punishment").

79. *See, e.g.*, Gless, Silverman, & Weigend, *supra* note 11, at 420–21 (arguing that robots are incapable of acting if we adopt a "thick" definition of an "act").

liefs and desires come exclusively from its managers, employees, or independent contractors.[80]

One possible response to these objections is that it is a mistake to treat corporations as legal persons under criminal law.[81] Another response, which is favored in this Article, is that natural persons are not the only moral agents that we recognize and on whom we impose criminal liability. Moreover, the criteria of moral agency for natural persons differ from those for non-human agents. The three objections, which are explained more fully below, may be valid against human moral agents, but they miss the point when they concern non-human agents. The latter response is favored for two main reasons. First, corporate criminal liability has been recognized in the United States for over one hundred years.[82] The very existence of an established practice often evidences the value of that practice. As pointed out at the end of Part III, the court has held corporations liable for offenses even though none of their human agents are guilty of those offenses. Second, a number of scholars have provided plausible accounts of moral agency that apply to non-human entities. Those accounts better explain the existing practice and, in the absence of clear errors, should be preferred.

This Part shows that if we were to accept alternative accounts of moral agency, then we have reason to treat smart robots as moral agents. In this regard, theories that explain why corporations can qualify as moral agents are instructive and will thus feature prominently in our response to these objections. The last objection argues against robot criminal liability on the basis that recognizing robots as legal persons can cause harm to humans and is therefore undesirable. I will seek to show that the alleged harms are not as serious as they might appear at first sight.

---

80. *See generally* Kendy M. Hess, *The Free Will of Corporations (and Other Collectives)*, 168 PHIL. STUD. 241, 250 (2014) ("The skeptic suggests that the corporate entity is just such a puppet, because the corporate entity's beliefs and desires are allegedly acquired inappropriately—*i.e.* implanted by external forces. If this is the case then the responsibility for corporate action lies not with the corporate entity, but with those external forces.").

81. Some academics do subscribe to this view. *See, e.g.*, Velasquez, *supra* note 55, at 14 ("Saying that a corporation is morally responsible for some wrongful act is acceptable only if it is an elliptical way of claiming that there are some people in the corporation who are morally responsible for that act and who should therefore be blamed and punished.").

82. *See* Gilchrist, *supra* note 51, at 5 ("[F]or a long time corporations were immune to criminal prosecution—but for the last hundred years this has not been the case.").

A. *Objection One: Incapable of Performing Actions*

First, one may argue that robots are not agents to whom actions can be attributed. In this view, an action is different from a "mere happening." Adam throwing a punch at Ben would generally fall within the category of actions. Raindrops falling on Ben's head, however, would not and would be better described as something that merely happens to Ben. In other words, Adam is an agent in our example, while raindrops are not.

The standard conception of action "construes [it] in terms of intentionality."[83] A leading proponent of the standard conception is Donald Davidson. He claims that a person is the agent of an act if what he does can be described as "intentional[] under some description."[84] Both the word "intentional" and the phrase "under some description" require some explanation. Davidson notes that an action can be described in various ways. Take the following scenario as an example: I dialed extension number 1212, believing it to be Jessica's number, when in fact it was Jamie's. My dialing 1212 was intentional; however, the same act can also be described as my calling Jamie, which was unintentional. Nevertheless, since what I did could be described as intentional under at least one description, according to Davidson, I was the agent of that action. Intention is generally understood as a mental state on an agent's part.[85] Davidson holds that intentions consist of both beliefs and desires.[86] He claims that a person acts with intention if he has both (a) a desire towards "actions with a certain property" and (b) "a belief [that the act] has that property."[87]

This theory of action suggests that robots cannot be the agent of any action because they lack the capacity to form any mental state. One version of the argument appeals to a theory commonly known as "dualism." Dualism is the view that the universe is divided into

---

83. Markus Schlosser, *Agency*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY 1 (Edward N. Zalta ed., 2015), https://plato.stanford.edu/archives/fall2015/entries/agency/. For a summary of the position held opponents of the standard conception, see *id.* ¶ 2.2.

84. DONALD DAVIDSON, ESSAYS ON ACTIONS AND EVENTS 46 (2d ed. 2001) ("[A] man is the agent of an act if what he does can be described under an aspect that makes it intentional.").

85. *See, e.g.*, Kieran Setiya, *Intention*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY 1 (Edward N. Zalta ed., 2015), https://plato.stanford.edu/archives/sum2015/entries/intention/ (noting the "the prevalent acceptance of intention as a mental state").

86. *See* DAVIDSON, *supra* note 84, at 5, 46.

87. *Id.* at 5. Note that Davidson refers to "a pro attitude" which can include desire and other mental states. *Id.*

two substances, the mental and the physical.[88] The former is not reducible to the latter. A smart robot, one would argue, consists of purely physical substances, including its "body" and the moral algorithms that control its behavior. In the absence of any mental substance, the robot is incapable of having mental states, such as desires or beliefs.

A more nuanced version of the argument would be along the lines of what John Searle labels "biological naturalism."[89] Searle accepts that mental states can be caused by the physical (for example, neuronal processes), but claims that they are "not ontologically reducible" to the latter.[90] He maintains that certain things, such as consciousness and intentionality, "exist only as experienced by a human or animal subject" and cannot exist by "producing a computer simulation of [the same]."[91] Searle has sought to demonstrate that a simulation of understanding does not amount to true understanding through his famous "Chinese room" thought experiment.[92] He asks readers (who are non-Chinese speakers) to imagine themselves locked in a room with many Chinese symbols and instructions that show how to manipulate those symbols. Imagine further that people outside the room send you Chinese symbols, which unknown to you, are Chinese questions. Imagine further that, by following the instructions, readers are able to send out a set of Chinese symbols, which are the correct answers to those questions. Searle concludes that, while the instructions enable you to create the impression that you know Chinese, you in fact do not understand a word of it.

---

88. *See generally* RENÉ DESCARTES, MEDITATIONS ON FIRST PHILOSOPHY (1996). As the father of modern philosophy, Descartes outlined an interpretation of the universe with both physical and mental elements.

89. John Searle, *Why Dualism (and Materialism) Fail to Account for Consciousness*, *in* QUESTIONING NINETEENTH-CENTURY ASSUMPTIONS ABOUT KNOWLEDGE, III: DUALISM 5, 14 (Richard E. Lee ed., 2010).

90. *Id.* at 23 ("I said that consciousness was causally reducible but not ontologically reducible to neuronal processes.").

91. *Id.* at 16 (rejecting the suggestion that consciousness, intentionality, or the "rest of the paradigmatic mental phenomena" can be created by producing a computer simulation of the same).

92. *See generally* John R. Searle, *Minds, Brains and Programs*, 3 BEHAV. BRAIN SCI. 417 (1980) (describing the thought experiment).

1. Response: Take an Intentional Stance Toward Smart Robots

If we accept Peter French's account that a corporation is able to act intentionally, then we have reason to believe that smart robots are able to perform actions as well. French claims that we can say that *x* did *y* intentionally if we can describe *x* as "having had a reason for doing *y* which was the cause of his or her doing it."[93] Corporations have reasons because "they have interests in doing those things that are likely to result in realization of their established corporate goals" and policies.[94] The sources of corporate policies and goals include a corporation's internal decision structure,[95] "precedent of previous corporate actions," "its statement of purpose," and so on.[96] These policies and goals are relatively stable and reflect more than "the current goals of [a corporation's] directors."[97] According to French, if an act is consistent with "an instantiation or an implementation of established corporate policy," then it is appropriate to "describe it as having been done for corporate reasons" or, in other words, as having been done intentionally by the corporation.[98]

French essentially invites us to take what Daniel Dennett calls an "intentional stance," that is, to treat objects as rational agents acting in accordance with their beliefs and desires.[99] Dennett argues that it is helpful to take an intentional stance where doing so would yield the best prediction of an object's behavior.[100]

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, which instantiates the robot's "goals and policies." These goals and

---

93. PETER A. FRENCH, COLLECTIVE AND CORPORATE RESPONSIBILITY 40 (Columbia Univ. Press 1984).

94. *Id.* at 45.

95. According to French, a corporation's Internal Decision Structure ("CID Structure") sets out the procedures for making corporate decisions, which instantiate a corporation's general policies and goals. A CID Structure consists of (1) "an organizational or responsibility flow chart that delineates stations and levels within the corporate power structure and (2) corporate decision recognition rule(s)." *Id.* at 41. By "recognition rules(s)," French meant "what Hart calls . . . 'conclusive affirmative indication' that a decision on an act has been made or performed for corporate reasons." French, *supra* note 53, at 212–13 (citing H.L.A. HART, THE CONCEPT OF LAW ch.6 (Oxford Univ. Press 1961)).

96. FRENCH, *supra* note 93, at 45.

97. French, *supra* note 53, at 214.

98. FRENCH, *supra* note 93, at 44.

99. DANIEL CLEMENT DENNETT, THE INTENTIONAL STANCE 15 (Mass. Inst. of Tech. 1989).

100. *Id.*

policies are relatively stable since they represent the aggregated moral views of a group of robot trainers rather than the current desire of any person who instructs the robot to carry out a task. By analogy, one might argue that 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.

One might further argue that taking an intentional stance is eminently suitable for smart robots. Their moral algorithms might be so complex that it would be impractical to question the individuals who wrote their codes or supplied their moral norms in order to determine the reasons for a particular robot's decision. Researchers have encountered this problem with existing, more basic, learning algorithms and have noted that, while some deep learning systems appear to "work," the engineers who built them do not "necessarily know[] why they work," nor are they able to "show the logic behind a system's decision."[101] For example, a software engineer at Google explained,

> If you ask an engineer, "Why did your program classify Person X as a potential terrorist?" the answer could be as simple as "X had used 'sarin' in an email," or it could be as complicated and nonexplanatory as, "The sum total of signals tilted X out of the 'non-terrorist' bucket into the 'terrorist' bucket, but no one signal was decisive." It's the latter case that is becoming more common . . . .[102]

Similarly, we might be able to determine that a smart robot values a billionaire's life more than a student's because the former has more money, but we would be unable to determine the reasons why the robot values people's lives based on how rich they are. In those situations, the robot itself might be the best and final authority of its decision.[103]

---

101. Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability*, 20 NEW MEDIA SOC'Y 973, 981 (2018).

102. David Auerbach, *The Code We Can't Control*, SLATE (Jan. 14, 2015, 2:03 PM), http://www.slate.com/articles/technology/bitwise/2015/01/black_box_society_by_frank_pasquale_a_chilling_vision_of_how_big_data_has.html.

103. *See, e.g.*, LIST & PETTIT, *supra* note 75, at 23 (recognizing that there might be "an engineering-style explanation" of a robot's moves, but as the robot becomes more complex, "the resort to the engineering stance becomes less feasible.").

B. *Objection Two: Incapable of Performing Morally Wrongful Actions*

Second, one may argue that, even assuming that a robot can act intentionally, its actions cannot be described as morally wrongful. To make sense of this argument, let me first explain what I mean by morally wrongful actions. An action (x) is morally wrongful if there is a moral rule or principle to the effect that actions of a certain category are wrong and x belongs to that category.[104] A moral rule is absolute, while a moral principle is contributory.[105]

An action may be intentional and harmful, but not morally wrongful. When a tiger attacks a loitering tourist, it probably has a desire for meat and believes that attacking the tourist is one way to satisfy that desire. Its action, therefore, is intentional. While the attack is no doubt harmful to the tourist, we probably would not say that the tiger has committed any morally wrongful action. Our reaction, however, would be very different if the attacker were human. Even if both actions inflict the same amount of physical harm on the victim, our reactions differ because of a key difference between the perpetrators of the harm. A human is presumed to know the key moral principles of the community that he lives in; the tiger is not.

It follows that unless an actor is capable of knowing that his action is wrong according to some moral rules or principles, it is inappropriate to judge his action against those rules and principles. This argument is supported by what is widely known as the M'Naghten rules, where a defense of "not guilty by reason of insanity" is available to any person who did "not know the nature and quality of the act he was doing; or, if he did know it, that he did not know what he was doing was wrong."[106] Consequently, to impose criminal liability on a robot requires proof that the robot "knows" that a moral principle applies to its action x, and that principle counsels against taking action x. Such proof, however,

---

104. A discussion of moral particularism, which claims that there are no moral principles, is outside of the scope of this Article.

105. If a statement such as "it is wrong to lie to your partner," is a moral rule, it means that every action that involves lying to one's partner is wrong overall. If that statement is a moral principle, it means that an action is morally worse if it involves lying to one's partner. But that action may be morally good in other respects (e.g., lying to one's partner to avoid hurting her feelings), such that the action is morally wrongful in some respects, but not wrongful overall.

106. M'Naghten's Case (1843) 8 Eng. Rep. 718 (HL) 719.

would be hard to come by unless a smart robot is specifically programmed to commit morally wrongful actions.

1. Response: Smart Robots as Members of Our Moral Community

A smart robot differs from a wild animal in an important respect: A smart robot is equipped with moral algorithms that have been trained by individuals who, we presume, know the prevailing moral norms and who we recognize as members of our community. We are, in turn, more ready to take an intentional stance towards smart robots and to treat them as possessing similar moral knowledge that possessed by their trainers. Therefore, we are more inclined to accept smart robots as members of our moral community, which is a status that we rarely attribute to wild animals. As a result, while we generally do not consider a tourist-attacking tiger as having committed any moral wrong, we are far more likely to reach the opposite conclusion if the same action were performed by a smart robot.

C. *Objection Three: Not Responsible for Its Actions*

Third, one may argue that a smart robot should not be responsible for its actions because it did not choose the moral principles on which it acts; in other words, it is not truly autonomous. Proponents of this objection maintain that a moral agent must be autonomous in the Kantian sense, where it must be appropriate to treat the agent as the author of its desires.[107] An agent cannot be properly considered the author of desires that are completely engineered by external forces. Although an agent's desires might be shaped by the circumstances in which he finds himself, at least part of those desires must originate from the agent himself. By contrast, robot "desires" or "beliefs" are completely engineered by forces external to the robot. They come exclusively from its algorithms, which are created by robot manufacturers and supplemented by robot trainers. As a result, robot actions can never be considered autonomous.

---

107. *See* IMMANUEL KANT, GROUNDWORK OF THE METAPHYSICS OF MORALS 54 (Mary Gregor ed. & trans., Cambridge Univ. Press 1998) (1785).

1. Response: Collective Responsibility

Alternative theories of moral agency suggest that autonomy in the Kantian sense may not be a necessary requirement for the imposition of criminal liability. This section examines in detail one of the more established theories of collective responsibility and demonstrates why smart robots may be considered criminally responsible in accordance with that theory.[108]

List and Pettit claim that an agent can be held responsible for his action if the following three conditions are satisfied.

(1) The agent faces a normatively significant choice, involving the possibility of doing something good or bad, right or wrong.
(2) The agent has the understanding and access to evidence required for making normative judgments about the options.
(3) The agent has the control required for choosing between the options.[109]

They further maintain that certain group agents, such as corporations, can satisfy these conditions so that it is appropriate to hold them responsible for their actions. The first requirement, according to List and Pettit, is likely satisfied for such agents since a group that acts to pursue certain desires based on its beliefs is bound to face normatively significant decisions from time to time.[110] The second requirement is satisfied where members of a group present normative propositions for the group's consideration and the group "takes whatever steps are prescribed in its organizational structure," such as voting, to form normative judgments about the options it faces.[111] Finally, List and Pettit argue that a group agent can satisfy the third requirement for two related reasons. First, a group attitude on a proposition can enjoy "a certain kind of autonomy in relation to individual attitudes."[112] List and Pettit demonstrate that even if individual group members' attitudes on a proposition (P&Q) are the same in two groups, their respective group attitude on that proposition might be different (see Table 1).[113] Group autonomy is evidenced by the fact that individual atti-

---

108. *See* LIST & PETTIT, *supra* note 75, at 155.
109. *Id.* at 158.
110. *Id.*
111. *Id.* at 159.
112. *Id.* at 71.
113. *Id.* at 70–71.

tudes on P&Q are neither sufficient nor necessary for determining the respective group attitude on that proposition.[114]

TABLE 1[115]

	Group A			Group B		
	P	Q	P & Q	P	Q	P & Q
Individual A	Yes	Yes	Yes	Yes	Yes	Yes
Individual B	Yes	No	No	No	No	No
Individual C	No	Yes	No	No	No	No
Premise-Based Procedure	Yes	Yes	= Yes[116]	No	No	= No

Second, drawing on the theory of "multi-level causality," List and Pettit claim that, in addition to the individuals who perform a group action, the group itself exercises some control over the group action.[117] To illustrate how causality can exhibit at different levels, List and Pettit use the example of a closed flask, which is filled with boiling water and subsequently breaks. At one level, the molecule that "triggers the break . . . causes the collapse [of the flask];" at a higher level, the boiling water also causes the break.[118] Using a metaphor from computing, the relationship between these two causally relevant events can be described as a higher-level event that "programs" for the collapse and a lower-level event that "implements" the program "by actually producing the break."[119]

---

114. *Id.* at 70 ("The individual attitudes on the conclusion are both insufficient for determining the group attitudes on it and unnecessary. We call the lack of sufficiency a 'weak autonomy' and the lack of necessity a 'strong autonomy.'").

115. The information in this table comes from LIST & PETTIT, *supra* note 75, at 70 tbls.3.1 & 3.2.

116. Here, "=Yes" and "=No" simply show the group decision on "P&Q" assuming both groups adopt a premise-based procedure for forming their group judgment.

117. LIST & PETTIT, *supra* note 75, at 161.

118. *Id.* at 162.

119. *Id.* ("The facts involved, described more prosaically, are these. First, the high-level event may be realized in many different ways, with the number, positions, and momenta of the constituent molecules varying within the constraint of maintaining such and such a mean level of motion. Second, no matter how the higher-level event is realized—no matter how the relevant molecules and motion are distributed—it is almost certain to involve a

By analogy, List and Pettit argue that a group agent is sometimes the "programming cause" of its action while the group member carrying out that action is the "implementing cause" of the same.[120] A group may exercise programming control over its action in a number of ways, such as by "maintaining procedures for the formation and enactment of its attitudes, arranging things so that some individuals are identified as the agents to perform a required task . . . ."[121] It is therefore appropriate to hold a group agent responsible for "ensuring that one or more of its members perform in the relevant manner."[122] It is worth noting that List and Pettit also take the view that holding a group agent responsible for an action does not absolve its members of their own responsibility for implementing that action.[123]

If we subscribe to List and Pettit's theory of group responsibility, we are likely to conclude that a smart robot satisfies all three requirements they propose. To begin, a smart robot, by definition, can make moral judgments and therefore satisfies both the first and second requirements. As noted in Part I.A, it is conceivable that a smart robot would face value-relevant decisions: an autonomous vehicle may have to decide which individual to crash into in an emergency, while a robot caretaker may have to decide whether to persuade a patient to take medicines against his will to improve his health.

Moreover, a smart robot enjoys some autonomy, particularly in the sense that the moral rules it has "learned" from robot trainers can be applied to novel situations. Its decision in that situation enjoys "a certain kind of autonomy" in relation to each robot trainer's decision, since the latter has not had the opportunity to pass his judgment. Moreover, the smart robot might take a course of action that is not foreseeable by any manufacturer or trainer. A primitive example of such unforeseeable action can be found in the case of AlphaGo, an artificial intelligence system built and trained by Google researchers to play Go, an ancient board game invented in China. AlphaGo has become famous worldwide after it beat Lee Sedol, one of the top Go players in the world. In its second match against Lee, it made a move that commentators deemed so ex-

---

molecule that has a position and momentum sufficient to break the flask. And, third, the way it is actually realized does have a molecule active in that role.")

120. *Id.*

121. *Id.* at 163.

122. *Id.*

123. *Id.* ("The members have responsibility as enactors of the corporate deed so far as they could have refused to play that part and didn't.").

traordinary that "no human could understand."[124] The level of astonishment surrounding that move suggests that it was probably not foreseen by most Go players.

One might also think that a smart robot exercises some control over its actions under Pettit's programming model. First, the moral algorithms enable multiple humans to shape the moral norms to which the robot adheres, and these norms, in turn, determine the robot's response to a particular set of stimuli. Second, the algorithms partially determine which humans should be authorized to influence a smart robot's moral outlook at any time. Third, the set of moral norms at any given time affects how a smart robot will be influenced through subsequent interactions with its environment and, in turn, shapes how it will act in response to those interactions.

D. *Objection Four: Recognizing Legal Personhood for Robots Is Harmful*

A fourth objection against recognizing robots as legal persons for the purpose of criminal law is that it might encourage people to anthropomorphize robots. According to Ryan Calo, this would be harmful to both individuals and society. Calo draws attention to five concerns with anthropomorphizing robots: (1) the more anthropomorphic a robot, the more likely people will praise or blame it; (2) the mere presence of robots might infringe people's privacy by "creating the sense of being observed"; (3) people might risk themselves on behalf of robots; (4) people might suffer emotional harm from losing robots; and (5) cruel treatment towards robots might cause some people to suffer moral harm.[125]

1. Response: Not as Harmful as They Seem

Although I share Calo's concerns, they are arguably insufficient to justify a blanket refusal to impose criminal liability on robots.[126]

---

124. Cade Metz, *How Google's AI Viewed the Move No Human Could Understand*, WIRED (Mar. 4, 2016), https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand/.

125. Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 547, 547–49 (2015).

126. Calo did not purport to argue that those concerns are sufficiently strong to preclude the possibility of robot criminal liability. *See id.* at 549.

To begin, some of the negative consequences identified by Calo are arguably the result of actions that may be, on balance, beneficial for humans. For example, despite the fact that people might suffer emotional harm from losing robots, the emotional bonds that they develop with objects can be essential to leading a meaningful life. Some scholars, such as Margaret Radin, even consider this emotional attachment the very basis for property rights.[127] Other negative consequences identified by Calo are probably temporary. Although humans currently feel that they have less privacy when robots are around, this may not be the case when robots are fully integrated into people's daily lives.[128] This is supported by findings by psychologists that people emotionally adapt to changes in their lives fairly quickly, which is a phenomenon referred to as "hedonistic adaptation."[129]

Moreover, refusing to impose criminal liability on robots may not be necessary to prevent the negative consequences that Calo identified. For example, a more effective way to prevent cruel treatment toward robots would be to impose liability on people who mistreat robots. In fact, certain negative consequences can be remedied by holding robots criminally liable for their misconduct. As discussed in Part III, victims are likely to suffer emotional harm when robots exhibit what they perceive to be immoral conduct. Criminalizing robots that engage in such misconduct might provide an outlet for the moral outrage that victims experience.

In light of the foregoing considerations, concerns over anthropomorphizing robots are arguably not sufficiently strong to outweigh the positive benefits that may flow from imposing criminal liability on smart robots.[130]

V. PUNISHING SMART ROBOTS

Lastly, one might think that smart robots cannot be held liable for their actions because they are not susceptible to punishment.

---

127. Margaret Jane Radin, *Property and Personhood*, 34 STAN. L. REV. 957, 959 (1981) ("Most people possess certain objects they feel are almost part of themselves. These objects are closely bound up with personhood because they are part of the way we constitute ourselves as continuing personal entities in the world.").

128. People from Solaria in Asimov's *Robot* series serve as a good fictional example. One might also argue that this feeling of a lack of privacy need not depend on the extent to which humans anthropomorphize robots.

129. DANIEL KAHNEMAN & EDWARD DIENER, WELL-BEING: FOUNDATIONS OF HEDONIC PSYCHOLOGY 302–29 (2003).

130. *See supra* Part III.

Let us first consider what we mean by punishment. H.L.A. Hart, citing Benn and Flew, defines the central case of punishment by reference to five elements:

> (1) it must involve pain or other consequences normally considered unpleasant;
> (2) it must be for an offense against legal rules;
> (3) it must be of an actual or supposed offender for his offense;
> (4) it must be intentionally administered by humans other than the offender; and
> (5) it must be imposed and administered by an authority constituted by a legal system against which the offense is committed.[131]

Punishment for smart robots can readily be designed to satisfy (2), (3), (4), and (5). The main controversy lies in (1). Few would dispute that prison sentences and fines are unpleasant for humans. However, are they unpleasant for smart robots, which might be incapable of feeling?

It is submitted that the key question is not whether a treatment is considered unpleasant by the robot, but whether it is considered unpleasant for the robot by general members of our community. If people are indeed used to taking an intentional stance towards smart robots, they are likely to find a treatment unpleasant for a robot if it removes all or part of the desires and beliefs that the robot has accumulated over a long period of time or if it undermines the robot's ability to satisfy those desires.

Assuming we can punish robots, a new question naturally follows: How should a robot be punished? In this regard, a range of measures might be taken to ensure that the robot commits fewer offenses in the future. These include:

> (1) physically destroying the robot (the robot equivalent of a "death sentence");
> (2) destroying or re-writing the moral algorithms of the robot (the robot equivalent of a "hospital order");
> (3) preventing the robot from being put to use (the robot equivalent of a "prison sentence"); and/or
> (4) ordering fines to be paid out of the insurance fund (the robot equivalent of a "fine").

---

131. HERBERT LIONEL ADOLPHUS HART, PUNISHMENT AND RESPONSIBILITY 4–5 (Oxford Univ. Press 1968). He also refers to four types of sub-standard cases of punishment, including "collective punishment" and "punishment of [non-offenders]." *Id.*

In addition, the unlawful incident can be used to design a training module to teach other smart robots the correct course of action in that scenario.

While the rest of the measures are fairly self-explanatory, robot fines might deserve some further explanation. One obvious problem is that a robot may not have any money or assets. A plausible solution to this problem is the creation of a no-fault insurance regime: any person who has directly contributed to a robot's action should be required to pay a sum of money to insure against the possibility that some of the robot's actions might turn out to be both harmful and wrongful. As such, the number of persons who end up "funding" a robot's fine depends on how "open" the robot's moral algorithm is. If any changes to those algorithms can only be made by robot manufacturers, then presumably only those manufacturers would be required to insure against its wrongful action. On the other hand, a smart robot may be equipped with moral algorithms that continue to learn from the decisions it makes on an ongoing basis from each person that is authorized to instruct or otherwise interact with the smart robot and to improve themselves in response to such interaction. As a result, after a significant period of time, each smart robot will be equipped with moral algorithms that are different from each other, since each has been exposed to a unique group of persons and moral preferences. In such cases, each person who is authorized to influence the robot's moral algorithms should also be required to contribute financially.

It is worth noting that members of the European Parliament have called for "a mandatory insurance scheme and a supplementary fund" to ensure that victims of driverless cars are adequately compensated.[132] For example, the proposed Vehicle Technology and Aviation Bill in the United Kingdom requires mandatory insurance for autonomous vehicles.[133] Similar insurance schemes might be put in place to ensure that a robot offender has the means to pay its fines.

While few people would object to taking measures to secure that smart robots commit fewer offenses, a more controversial question,

---

132. European Parliament Press Release 20170210IPR61808, Robots and Artificial Intelligence: MEPs Call for EU-Wide Liability Rules (June 10, 2017, 13:09), http://www.europarl.europa.eu/news/en/press-room/20170210IPR61808/robots-and-artificial-intelligence-meps-call-for-eu-wide-liability-rules.

133. Kerris Dale & Alistair Kinley, *Automated driving legislation heads to Lords - a third Bill at its third reading*, BERRYMANS LACE MAWER LLP (Jan. 30, 2018), https://www.blmlaw.com/news/automated-driving-legislation-heads-to-lords-a-third-bill-at-its-third-reading.

which is outside the scope of this Article, is whether these measures should be taken for the purpose of achieving other aims, such as retribution. At least one scholar, Christina Mulligan, has argued in favor of punishing robots for retributory purposes on the ground that doing so can provide psychological benefits to victims who demand revenge.[134] She posits that the most satisfying outcome for those victims "might be the early Middle Age practice of noxal surrender," which allows a victim to physically harm or destroy a robot.[135]

CONCLUSION

This Article makes a positive case for imposing criminal liability on a type of robot that is likely to emerge in the future. The justifications for imposing such liability remain human-centric, where doing so helps alleviate harms to our community's collective moral system as well as emotional harms to human victims. I do not rule out the possibility that if robots become indistinguishable from humans in the future, we might justify imposing such liability based on the benefits it provides to robots themselves.

This discussion also raises fascinating questions for further research. Future researchers might want to delve deeper into questions such as: What types of moral decisions can or should be delegated to smart robots? Which persons should be allowed to serve as robot trainers? What are the minimum moral standards to which smart robots should adhere? No doubt, these are difficult questions. Our ability to answer these questions would be aided by a coherent regime of criminal laws that apply to autonomous robots. This Article demonstrated that such a regime could be created and reconciled with current doctrines of criminal law. Future research into the moral standards of robots and into the training of autonomous robots can benefit from the guiding principles laid down in this Article for establishing laws that protect against robots' immoral actions. In turn, the hope is that these principles will incentivize responsible use of new technology.

---

134. Mulligan, *supra* note 60, at 593 ("But robot punishment—or more precisely, revenge against robots—primarily advances a different goal: the creation of psychological satisfaction in robots' victims.").

135. *Id.* at 595 ("[T]he most satisfying outcome for a person wronged by a robot might be the early Middle Age practice of 'noxal surrender.'").