REVENGE AGAINST ROBOTS

Christina Mulligan*

I. INTRODUCTION

When a robot hurts a human, how should the legal system respond? Our first instinct might be to ask who should pay for the harm caused, perhaps deciding to rest legal liability with the robot's hardware manufacturer or its programmers. But besides considering tort or criminal actions against corporate and human persons, legal actors might also target the most immediate source of the harm—the robot itself.

The notion of holding a robot accountable for its actions initially evokes absurd and amusing mental images—a prosecutor pointing to a smart toaster shouting, "And what do you have to say in your defense? Jury, note that the toaster says nothing. It says nothing because it is guilty." And it is easy to laugh at this scenario and brush the idea aside. But there are more rational ways to hold robots accountable for their actions and reasons why law and policy makers would want to do so.

---

* Associate Professor of Law, Brooklyn Law School. Thank you to Ryan Calo, Kate Darling, Woodrow Hartzog, Alex Lemann, and the participants in the 2017 Nebrooklyn Junior Technology Law Scholars Workshop and the 2017 Junior Faculty Workshop series at Brooklyn Law School.

This Essay proceeds by first exploring the benefits to human persons, namely the creation of psychological satisfaction, that could be advanced by punishing robots for their actions (Part II) and goes on to address the concern that punishing robots would psychologically harm humans (Part III). It then shifts focus to robots themselves, arguing that it is justifiable for humans to blame robots for their actions because, like animals, autonomous robots are best understood as the causes of their own actions (as "agents") (Part IV). The Essay then explores whether a robot's moral culpability is relevant to the issue of robot punishment (Part V) and, finally, considers how revenge against robots could be implemented (Part VI).

II. SEEKING SATISFACTION THROUGH REVENGE

When asked what the goal of tort law is, many say that its goal is "to make victims whole."[1] "Making a victim whole" is usually understood to mean returning a victim to the position they were in before a harm occurred or placing a victim in the position they would have been in had they not been

---

1. *See* 1 MARILYN MINZER ET AL., DAMAGES IN TORT ACTIONS § 3.01, at 3-4 (2002) ("The general purpose of compensatory damages in tort actions is to give the injured person a sum of money which will restore him, as nearly as possible, to the position he would have been in if the wrong had not been committed; in other words, to make the plaintiff whole"); W. PROSSER, HANDBOOK OF THE LAW OF TORTS § 2, at 7 (4th ed. 1971) (explaining that the purpose of a tort action is to compensate victim for damage he has suffered); Heidi Li Feldman, *Harm and Money*, 75 TEX. L. REV. 1567, 1577-80 (1997) (discussing various courts' approaches to making tort victims whole); Adam J. Kolber, *The Experiential Future of the Law*, 60 EMORY L.J. 585, 608 (2011) ("No matter one's theory of tort law, the dominant view of tort compensation is that damages are supposed to return individual plaintiffs to the status quo ante . . . ."); Mary Alice McLarty, *Medical Malpractice Remedies*, 49-MAY TRIAL 6 (2013) (in which the then-president-elect of the Dallas Trial Layers stated, "As trial lawyers, we know the purpose of tort law is to make injured people whole."); Pam Mueller, *Victimhood and Agency: How Taking Charge Takes Its Toll*, 44 PEPPERDINE L. REV. 691, 693-97 (2017) (citing cases that articulate the goal of making victims whole); Ellen S. Pryor, *Rehabilitating Tort Compensation*, 91 GEO. L.J. 659, 660-61 (2003) (stating that "a dominant theme" among legal academics and practitioners is that tort law is designed to return plaintiffs to the status quo ante); Benjamin C. Zipursky, *Civil Recourse, Not Corrective Justice*, 91 GEO. L.J. 695, 749 (2003) ("Our tort system clearly puts great emphasis on damages, and particularly on compensatory damages. In so doing, it obviously makes use of the concept of making whole, and of a principle that the plaintiff is entitled to be made whole"); *cf.* Price Waterhouse v. Hopkins, 490 U.S. 228, 264 65 (1989) (O'Connor, J., concurring) ("Like the common law of torts, the statutory employment 'tort' created by Title VII has two basic purposes. The first is to deter conduct which has been identified as contrary to public policy and harmful to society as a whole. . . . The second . . . is to make persons whole for injuries suffered . . . ." (internal quotations omitted)).

injured.[2] In practice, this means awarding tort victims financially.[3] When a pedestrian gets hit by a car and physically injured, they can sue the car's driver and may receive as compensation the cost of their medical bills, lost wages from the injury, and even an approximation of the value of their physical pain and suffering.[4] But consider for a moment an alternate society where, instead of suing the driver, an injured pedestrian appeals to a government body or a private charity to cover the cost of the accident. In this society, governments or charities are very generous, and injured individuals receive in compensation the same as or more than they would have received in our more litigious society. Plus, the procedure is fast and efficient and does not require long, contentious legal battles. Your instinct may be that this alternate society sounds preferable because injured parties are made whole quicker. But you might also imagine yourself in that situation and feel a nagging frustration—the person who caused your injuries does not appear in the story at all. The driver hit you with a car and then just walked away from the situation. Their car insurance payments did not even increase. Maybe they never thought about you again, knowing that existing institutions would make sure you were taken care of.

One might suppose that this frustration comes from a desire for revenge—the desire to see the car driver suffer in some way because of the suffering they inflicted.[5] Here, it is useful to distinguish between retribution and revenge. Retributivists might support the law taking punitive actions because

---

2. Restatement (Second) of Torts § 901 cmt. a (AM. LAW INST. 1977) ("[T]he law of torts attempts primarily to put an injured person in a position as nearly as possible equivalent to his position prior to the tort."); 25 C.J.S. DAMAGES § 2 (1996) (stating that compensation should "put the injured party in the position in which he was before he was injured"). Mueller, *supra* note 1, at 695 (citing cases articulating these goals).

3. See RICHARD A. POSNER, ECONOMIC ANALYSIS OF THE LAW 147 (3d ed. 1986) (stating that "wrongs that subject the wrongdoer to a suit for damages by the victim . . . are called torts"); STEVEN SHAVELL, ECONOMIC ANALYSIS OF ACCIDENT LAW 127 (1987) (concluding that liability translates to damages owed).

4. Feldman, *supra* note 1, at 1575 ("The traditional tort system requires a tortfeasor to pay damages for pain and suffering as well as for lost wages and medical expenses.").

5. *See* Scott Hershovitz, *Tort As A Substitute For Revenge*, *in* PHILOSOPHICAL FOUNDATIONS OF THE LAW OF TORTS 87–88 (John Oberdiek ed., 2014) ("The idea that tort suits are a substitute for revenge is still with us today . . . . On this picture, tort damages are a payoff to prevent private violence . . . . [Some] suggest that a tort suit provides a plaintiff an opportunity to take revenge by inflicting harm on the person who harmed her, even when only compensatory damages are on offer."); Alexander B. Lemann, *Stronger Than the Storm: Disaster Law in a Defiant Age*, 78 LA. L. REV. 437, 481-82 (2017) ("[R]eceiving compensation that does not come from a tortfeasor . . . may make the victim 'whole' by undoing a certain allocation of costs related to an injury, but it does nothing to offer 'satisfaction' to assert a right to a certain standard of treatment at the hands of others and thus help the victim get even.").

the driver deserves to suffer for their wrongdoing. But while retribution focuses on what a wrongdoer deserves, the notion of revenge additionally involves a personal desire in the wrongdoer's victim.[6]

To the extent that we focus on the effect of revenge on the victim rather than the wrongdoer, one might be concerned that indulging a desire for revenge would cultivate socially undesirable behaviors and attitudes in the victim. One might additionally resist the idea of vengeance, as distinct from retributive punishment, under a theory that revenge merely harms a wrongdoer while doing little to benefit society. And depending on the form it takes, revenge does not necessarily contribute to making the original victim whole in any concrete sense; rather, as Mahatma Gandhi purportedly said, "An eye for an eye leaves the whole world blind."[7]

Yet this understanding of revenge omits a consideration of the psychological benefit that revenge might create in harmed individuals—*satisfaction*[8]—which is more defensible as an outcome the law and society should support. The purpose of revenge may initially seem to be to inflict harm on a wrongdoer, but revenge (as contrasted with retribution) is more precisely conceived of as the means and not the end. The *end* of revenge is satisfaction—the re-establishment of the social standing and power balance between the wrongdoer and victim so that, in the eyes of any of the wrongdoer, the victim, or third parties, the noneconomic injury committed against the victim has been set aright.[9]

Psychology research indicates that satisfaction is a complicated state—it is not simply created by inflicting harm on a wrongdoer.[10] However, while

---

6. *See* ROBERT NOZICK, PHILOSOPHICAL EXPLANATIONS 367 (1981) (distinguishing revenge from retribution).

7. See YALE BOOK OF QUOTATIONS 269–70 (Fred R. Shapiro ed., 2006).

8. *Satisfaction*, Oxford English Dictionary (2d ed. 1989) (stating, among other definitions, that satisfaction is "the atoning for . . . an injury, offence, or fault by reparation, compensation, or the endurance of punishment" and "to be avenged on an offender").

9. *See* Hershovitz, *supra* note 5, at 92–95; *id.* at 98 ("Revenge is rarely just a private performance."); JEFFRIE G. MURPHY & JEAN HAMPTON, FORGIVENESS AND MERCY 25 (1988) (stating "our self-respect is *social*"); Pamela Hieronymi, *Articulating an Uncompromising Forgiveness*, 62 PHIL. & PHENOMENOLOGICAL RES. 529, 546 (2001) ("I suggest that a past wrong against you, standing in your history without apology, atonement, retribution, punishment, restitution, condemnation, or anything else that might recognize as a wrong, makes a claim. It says, in effect, that you can be treated this way, and that such treatment is acceptable.").

10. *See e.g.*, Kevin M. Carlsmith et al., *The Paradoxical Consequences of Revenge*, 95 J. PERSONALITY & SOC. PSYCHOL. 1316, 1324 (2008) ("[R]evenge has hedonic consequences that are the opposite of what people expect. Revenge can prolong peoples' hedonic reactions to a

harming wrongdoers alone may not be satisfying to parties seeking revenge, one psychological study has demonstrated that vengeance *is* satisfying when the objects of revenge are not only harmed but also recognize that their earlier actions resulted in their later suffering.[11]

Psychological satisfaction need not only be achieved by a combination of acknowledgement and harm. A third party's judgment could generate the effect as well, such as when courts, the press, or one's social circle state and agree that an alleged wrongdoer is at fault. Scholars have argued that one role of the court process is to create a shared narrative of what "the truth" of a situation is.[12] Psychological studies show that victims who act as agents and tell their stories in court are more likely to feel positively about the outcome of their cases, even though they are less likely to be as highly economically compensated as someone who takes less of an active role in telling their story or moving their case forward.[13] Other work indicates that court judgments affirm a plaintiff's worth to their community or society.[14] Even the availability of "nominal damages" exemplifies how the public nature of the court system can provide satisfaction in cases where what a victim really wants is an acknowledgement by an authority figure or society that they were wronged.[15]

---

transgression because punishing others can cause people to continue to think about (rather than to forget) those whom they have punished.").

11. Mario Gollwitzer et al., *What Gives Victims Satisfaction When They Seek Revenge?*, 41 EUR. J. SOC. PSYCHOL. 364, 364–74 (2011); *see also* Eric Jaffe, *The Complicated Psychology of Revenge*, ASSOCIATION FOR PSYCHOLOGICAL SCIENCE: OBSERVER (Oct. 4, 2011), https://www.psychologicalscience.org/observer/the-complicated-psychology-of-revenge.

12. *See* Paul Schiff Berman, *Rats, Pigs and Statues on Trial*, 69 N.Y.U. L. REV. 288, 316–21 (1994); James B. White, *Law as Language: Reading Law and Reading Literature*, 60 TEX. L. REV. 415, 444 (1982).

13. *See* Mueller, *supra* note 1, at 697–700 and accompanying footnotes.

14. *See* Kenworthey Bilz, *The Puzzle of Delegated Revenge*, 87 B.U. L. REV. 1059, 1062 (2007); Emily Sherwin, *Comments on Stephen Smith's Duties, Liabilities, and Damages*, 125 HARV. L. REV. F. 164, 169 (2012) (stating that a "wrong can be viewed as a denigration of the victim's moral worth" and that "providing the victim with a retaliatory remedy is a way to recognize, and allow the victim to reassert, moral equality"); Jason M. Solomon, *Equal Accountability Through Tort Law*, 103 NW. U. L. REV. 1765, 1795 (2009); Sarah Swan, *Triangulating Rape*, 37 N.Y.U. REV. L. & SOC. CHANGE 403, 429 (2013).

15. *See* 25 C.J.S. DAMAGES § 17 (2017) ("Nominal damages are not compensation for loss or injury but rather recognition of a violation of rights; they are a symbolic recognition of harm that may be awarded without proof of actual harm and have only declaratory effect."); *see also* Saul Litvinoff & Ronald J. Scalise Jr., 6 LA. CIV. L. TREATISE, LAW OF OBLIGATIONS IN THE LOUISIANA JURISPRUDENCE § 7.21 (2d ed.) ("[T]he symbolized conclusion [in granting nominal damages] is that the court regards the defendant's conduct as reprehensible even if it has not caused actual, or measurable, loss to the plaintiff, and that such a defendant should not be let go without at least a reprimand or a slap on the wrist. In those cases where a mere finding by the court may be the best satisfaction for the aggrieved party . . . the symbolic function of a

Indeed, some litigants claim that what they really want from the objects of their lawsuits is a sincere apology.[16] In private life, we are also likely familiar with situations where two individuals have a personal conflict and appear to try to convince their mutual acquaintances of "who was right" in the absence of one party's conceding that they were wrong. These sorts of actions—apology, making a public statement, and authoritative or public judgment—all serve to provide satisfaction to a victim or party to a conflict by indicating that the victim or party was in the right.

Satisfaction is ultimately about restoring perceived and real power and social standing, and a variety of behaviors besides those described above can reset the power dynamics between two parties. As a historic example, duelists often evaded physical harm,[17] and no one was necessarily established as the wrongdoer following a duel.[18] The ritualization of the duel served as a mechanism by which social standing between two feuding parties could be restored.[19] Although duels served to restore one's standing in the eyes of one's community, they also served to change the psychological disposition of the

---

trifling award is evident . . . .") (citing CHARLES T. MCCORMICK, A HANDBOOK ON THE LAW OF DAMAGES 95 (1935)); *see e.g.*, Keith Coffman & Jann Tracey, *Taylor Swift wins groping trial against DJ, awarded symbolic $1*, REUTERS, Aug. 14, 2017, https://www.reuters.com/article/us-people-taylorswift/taylor-swift-wins-groping-trial-against-dj-awarded-symbolic-1-idUSKCN1AU108.

16. JENNIFER K. ROBBENNOLT & VALERIE P. HANS, THE PSYCHOLOGY OF TORT LAW 20 (2016) ("Many claimants want an apology. Many say they would not have filed a lawsuit if the other person had apologized; settlement negotiations may stall in the absence of an apology; and many claimants express disappointment when they do not receive an apology.").

17. In 1817, a British commentator estimated that a duelist had a one-in-four chance of being killed or wounded. Alison L. LaCroix, To Gain *the Whole World and Lose His Own Soul: Nineteenth-Century American Dueling as Public Law and Private Code*, 33 HOFSTRA L. REV. 501, 517 (2004) (citing Antony E. Simpson, *Dandelions on the Field of Honor: Dueling, the Middle Classes, and the Law in Nineteenth-Century England*, 9 CRIM. JUST. HIST. 99, 112 (1988)); *cf.* LIN-MANUEL MIRANDA ET AL., *Ten Duel Commandments*, *on* HAMILTON (ORIGINAL BROADWAY CAST RECORDING) (Atlantic Records 2015) ("Most disputes die, and no one shoots.").

18. *See* LaCroix, *supra* note 17, at 521 ("After the first round of fire, the seconds met to determine whether the injured party's honor had been satisfied, which it often was after a single round with no injuries.").

19. C.A. Harwell Wells, *The End of the Affair? Anti-Dueling Laws and Social Norms in Antebellum America*, 54 VAND. L. REV. 1805, 1823 (2001) ("As one anthropologist explains it, in a society with such a view of honor 'the being and truth about a person are identical with the being and truth that others acknowledge in him.' Thus the need for Southern men to participate in the 'affair of honor,' even when morally opposed to dueling. The point of a duel was not to reaffirm one's self-worth, but to demonstrate that worth to others.") (citing EDWARD L. AYERS, VENGEANCE AND JUSTICE: CRIME AND PUNISHMENT IN THE 19TH-CENTURY AMERICAN SOUTH 13 (1984)).

parties with respect to each other. Duelists found themselves in a position of both deadly power over and vulnerability to their opponent, and in circumstances where no deadly shots were attempted, a duelist found himself in both a superior and grateful position: superior because he spared his opponent when he did not have to, and grateful because his opponent spared him. The gravity of the moment could be understood as not simply satisfying arguably-barbaric social expectations, but as genuinely changing the psychological disposition that the duelists had towards each other.[20]

III. DIRECTING EMOTIONS AT THE ROBOT

So instead of our original example of the driver and pedestrian, let us instead consider an autonomous, self-driving car that injures a pedestrian. When considering how to approach the injured party, we now may be inclined to ask at least two questions: who should make the pedestrian economically whole, and how can the pedestrian achieve satisfaction from the aftereffects of the accident. Depending on the circumstances, we can imagine being attracted to several options for holding human persons and corporations accountable for the accident. Of the parties, we could target the hardware manufacturer or repairer, the software programmers, or even the parties who chose to install the software in the car. In choosing to privilege valuing safety or encouraging innovation, we could imagine courts' holding these parties to different standards of liability.[21] To encourage innovation, various parties could be granted immunity for various kinds of accidents, or damages could be capped. Alternatively, to encourage exacting safety precautions, some parties could be strictly liable even if they were not at fault.

While our inclinations first may be to ask what legal standards will encourage innovation and safety, we can also ask what legal frameworks will provide satisfaction to victims of robot-related accidents. Depending on what kind of robot we are dealing with and the kind of harm that occurred, the target of one's need for satisfaction might be quite different. In the case of a hardware or physical failure, one might direct one's ire to the manufacturers. In the case of a non-autonomous smart device, where the cause of a problem

---

20. As an anecdotal example of someone intuitively refusing to put themselves in a duelist's position and re-establish the social standing between disputing parties, I am reminded of an apocryphal story of a groom in Texas whose fiancée cheated on him with his best man. The best man, apologetic upon being discovered, suggested that the groom remedy the situation by punching the best man in the face. Preferring instead to be done with the friendship rather than to re-establish their standing with respect to each other, the groom declined the invitation because his former friend "wasn't worth it."

21. *See generally* M. Ryan Calo, *Open Robotics*, 70 MD. L. REV. 571 (2011).

within the computer's code is clear, one might reasonably be inclined to criticize the software developer or the party who chose to run the software in the device. In the case of any computer, if *many* similar devices are making the same error, even if we cannot understand why, one might again focus on the software developer or manufacturer, with an instinct towards the doctrine of *res ipsa loquitur*—even if we cannot tell exactly what happened, we might be inclined to think there was negligent design.

But one can also imagine human users' blaming and experiencing anger with the offending object itself instead of the programmer or manufacturer. Both current events and fiction suggest this insight is accurate. News reports about the "drone slayer," who shot down an aerial drone over his house, illustrate the instinct to target offending objects instead of or in addition to the human actors responsible for them.[22] Similarly, one of the most famous scenes in the movie *Office Space* consists of several characters carrying a frustrating office printer into an empty field and destroying it with a baseball bat.[23]

This phenomenon is likely even more pronounced in the case of social and autonomous robots. "Social robots" are defined by robot ethics researcher Kate Darling as "a physically embodied, autonomous agent that communicates and interacts with humans on a social level."[24] Some social robots only act in predetermined ways, but others are autonomous. "Autonomous" robots have the ability to "make (limited) decisions about what behaviors to execute based on perceptions and internal states, rather than following a pre-determined action sequence based on pre-programmed commands."[25] Many studies and anecdotes indicate that humans feel more empathy towards robots the more life-like they seem—as they appear more social and autonomous.[26] This empathy can manifest as a powerful emotional

---

22. *See* Cyrus Farivar, *Judge Rules in Favor of "Drone Slayer," Dismisses Lawsuit Filed by Pilot*, ARS TECHNICA (Mar. 24, 2017), https://arstechnica.com/tech-policy/2017/03/judge-rules-in-favor-of-drone-slayer-dismisses-lawsuit-filed-by-pilot/; James Vincent, *Judge Rules Kentucky Man Had the Right to Shoot Down His Neighbor's Drone*, VERGE (Oct. 28, 2015), https://www.theverge.com/2015/10/28/9625468/drone-slayer-kentucky-cleared-charges.

23. *See* Office Space Movie Clip, https://www.youtube.com/watch?v=_KinUMIS3Yc (last visited Mar. 2, 2018); OFFICE SPACE (Twentieth Century Fox 1999).

24. Kate Darling, *Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects*, *in* ROBOT LAW 213, 215 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016).

25. *Id.* at 215 n.5 (quoting Matthias Scheutz & Charles Crowell, *The Burden of Embodied Autonomy: Some Reflections on the Social and Ethical Implications of Autonomous Robots*, WORKSHOP ON ROBOETHICS INT'L CONF. ON ROBOTICS & AUTOMATION 1 (2007)).

26. *See* Sherry Turkle, *In Good Company? On the Threshold of Robotic Companions*, *in* CLOSE ENGAGEMENTS WITH ARTIFICIAL COMPANIONS: KEY SOCIAL, PSYCHOLOGICAL,

aversion to causing robots pain and suffering even though humans consciously know that the robots are not alive and cannot feel anything.[27] One could analogously imagine humans being equally inclined to feel negative or vengeful emotions towards autonomous, social robots, even if they "know" that robots are merely objects that run code.

Kate Darling has made a Kantian-style argument[28] that even though robots do not experience suffering, governments should pass laws that protect robots from cruelty for some of the same reasons that laws against animal cruelty exist—to guide the psychological state of humans who might act

---

ETHICAL AND DESIGN ISSUES 3, 3–10 (Yorick Wilks ed., 2010); Matthias Scheutz, *The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 205, 205–22 (Patrick Lin, Keith Abney & George A. Bekey eds., 2012); Ja-Young Sung, Lan Guo, Rebecca Grinter & Henrik Christensen, *My Roomba is Rambo: Intimate Home Appliances*, 9TH INT'L CONF. ON UBIQUITOUS COMPUTING 145, 145–62 (2007); *see also* Darling, *supra* note 24, at 217–18 ("[R]obots employed within military teams . . . evoke fondness and loyalty in their human teammates, who identify with the robots enough to name them, award them battlefield promotions and 'purple hearts,' introduce them to their families, and become very upset when they 'die.'") (citing Julie Carpenter, The Quiet Professional: An Investigation of U.S. Military Explosive Ordinance Disposal Personnel Interactions with Everyday Field Robots (2013) (unpublished Ph.D. dissertation, University of Washington) (on file with University of Washington Libraries)); Michael Kolb, Soldier and Robot Interactions in Combat Environments (2012) (unpublished Ph.D. dissertation, University of Oklahoma).

27. *See, e.g.*, Kate Darling et al., *Empathic Concern and the Effect of Stories in Human-Robot Interaction,* 2015 PROC. IEEE INT'L WORKSHOP ON ROBOT & HUM. COMM., https://ssrn.com/abstract=2639689; Mel Slater et al., *A Virtual Reprise of the Stanley Milgram Obedience Experiments*, 1 PLOS ONE, no. 1, e39, at 1–10 (2006), http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000039; Christoph Bartneck et al., *Daisy, Daisy, Give Me Your Answer Do! Switching Off a Robot*, PROC. 2ND ACM/IEEE INT'L CONF. ON HUM.-ROBOT INTERACTION 217, 217–22 (2007); Christoph Bartneck et al., *To Kill a ~~Mockingbird~~ Robot*, PROC. OF THE 2ND ACM/IEEE INT'L CONF. ON HUMAN-ROBOT INTERACTION 81, 81–87 (2007). One article describes a landmine-diffusing robot that would lose a leg when it stepped on a mine and continue walking on its remaining legs. The colonel in command of the operation called off the exercise because he "just could not stand the pathos of watching the burned, scarred and crippled machine drag itself forward on its last leg. This test, he charged, was inhumane." Joel Garreau, *Bots on the Ground*, WASH. POST (May 6, 2007), http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html.

28. "[I]f a man has his dog shot, because it can no longer earn a living for him, he is by no means in breach of any duty to the dog, since the latter is incapable of judgement, but he thereby damages the kindly and humane qualities in himself, which he ought to exercise in virtue of his duties to mankind. Lest he extinguish such qualities, he must already practice a similar kindliness towards animals; for a person who already displays such cruelty to animals is also no less hardened towards men." IMMANUEL KANT, LECTURES ON ETHICS 212 (Peter Heath & J.B. Schneewind eds., Cambridge Univ. Press 1997).

against them and to reflect moral judgments about how humans should behave rather than to protect the animals (or robots) themselves.[29]

But, even if torturing innocent robots is psychologically harmful to humans, enacting revenge on a robot that has caused harm, as part of a formal process, may not be. Indeed, it may be that vengeance conducted in particular, formal, sanctioned situations and cultural contexts may cultivate positive characteristics in vengeful actors, such as repaired self-confidence and restored feelings of personal autonomy. Although counterintuitive, we can argue that violence against misbehaving robots is justifiable using reasoning similar to Darling's: our actions against robots should be chosen based on what is psychologically beneficial to humans rather than on what the robots "deserve." If it turns out that punishing robots provides the right kind of psychological benefit to humans following an injury, we should punish robots.

A practical concern about both this claim and Darling's argument stems from the fact that humans know robots do not have phenomenal experiences.[30] This knowledge seems as though it would undermine the psychological effects of acting on the robots. Robots don't have feelings, and so while some people might be psychologically harmed by torturing innocent robots or psychologically healed by enacting justified revenge on one, others might have minimal or nonexistent reactions because they have internalized that robots are "not real."[31] In fact, one might think that we *should* fight the instinct to feel emotions for robots because those emotions are just a manifestation of our brains making an error that we should train ourselves to correct.[32] Darling

---

29. Darling, *supra* note 24, at 226–29; *cf.* Hunter Walk, *Amazon Echo Is Magical. It's Also Turning My Kid into an Asshole*, HUNTER WALK (Apr. 6, 2016), https://hunterwalk.com/2016/04/06/amazon-echo-is-magical-its-also-turning-my-kid-into-an-asshole/ ("Cognitively I'm not sure a kid gets why you can boss Alexa around but not a person. At the very least, it creates patterns and reinforcement that so long as your diction is good, you can get what you want without niceties.").

30. "Phenomenal experiences" as in conscious experiences, not super awesome experiences. *See* David Woodruff Smith, *Phenomenology*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., Winter ed. 2016), https://plato.stanford.edu/archives/win2016/entries/phenomenology/.

31. Despite talking about "autonomous" robots, this Paper explicitly does not consider the question of when a robot stops being a "thing" and becomes intelligent or autonomous enough to become a rights-bearing "person." The author presumes that all currently-existing and near-future robots will not be complex enough to raise the question, but that Commander Data is rights-bearing. *See Star Trek: The Next Generation: The Measure of a Man* (CBS television broadcast Feb. 13, 1989) (depicting a trial adjudicating the question of whether a self-aware android has rights to self-determination). For more on this topic, see Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231 (1992).

32. Indeed, a major function of cognitive behavioral therapy is often to teach patients to adjust their emotional reactions to more reasonably reflect the real stakes of a situation. *See*

notes that, in many cases, empathy for robots is counterproductive, recounting several military operations where soldiers were inclined to make choices to prevent their robotic tools from getting hurt.[33] In light of these real concerns, Darling concludes that the ultimate answer to how we treat most robots supervenes on how that treatment affects humans.[34] If a behavior is helpful, we should do it. If it is not, we shouldn't.

IV. ROBOTS AS AGENTS

In the future, more robots will be in a position to run what are called "black box algorithms"—algorithms that are so complex that even their designers and programmers cannot explain what the computers running the algorithms end up doing.[35] While we might choose to "make people whole" economically by having the manufacturer, programmer, seller, or insurance company pay for any harm done by robots that are running black box algorithms, the actions of rogue robots cannot always neatly be said to be "caused" by any of these actors.

Suppose someone develops and markets, for instance, a gardening robot capable of learning new behaviors, and one person's robot starts breaking ground-floor windows in an apartment building instead of pruning bushes. It may be that some unusual inputs caused this robot alone to start acting unexpectedly. The fact that other gardener robots do not start breaking windows might indicate that this effect was not reasonably foreseeable. It might also be true that the individuals who programmed the robot would not be able to explain to you why the robot started breaking windows. (In other

---

SAMUEL T. GLADDING, COUNSELING: A COMPREHENSIVE PROFESSION (7th ed., 2007); *see also, e.g.*, Alice Boyes, *Cognitive Restructuring*, PSYCHOL. TODAY (Jan. 21, 2013), https://www.psychologytoday.com/blog/in-practice/201301/cognitive-restructuring.

33. Darling, *supra* note 24, at 217–18 ("[R]obots employed within military teams . . . evoke fondness and loyalty in their human teammates, who identify with the robots enough to name them, award them battlefield promotions and 'purple hearts,' introduce them to their families, and become very upset when they 'die.'") (citing Carpenter, *supra* note 26; Kolb, *supra* note 26); Kate Darling, *Who's Johnny?: Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy*, *in* ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE 173, 173–92 (Patrick Lin et al. eds., 2017).

34. *See* Darling, *supra* note 33.

35. *See* W. Nicholson Price II, *Big Data, Patents, and the Future of Medicine*, 37 CARDOZO L. REV. 1401, 1404 (2016) (describing black box algorithms that analyze health information as "'black-box' precisely because the relationships at [their] heart are opaque—not because their developers deliberately hide them, but because either they are too complex to understand, or they are the product of non-transparent algorithms that never tell scientists, 'this is what we found.' Opacity is not desirable but is rather a necessary byproduct of the development process.").

cases, it might be that the harm caused is also too small to justify a true "autopsy" of what the robot's algorithm was doing.)

So, in the case of the autonomous, black-box-algorithm-running gardener, what should we say caused the broken windows? We might talk about the actions of the programmers and manufacturers being necessary conditions for the windows to break, a few but-for causes among thousands of causes. But, the nature of learning algorithms prevents these figures' actions from seeming like proximate causes of the accident. Given the rarity of the gardener's behavior and the lack of explanation (even in hindsight) for why it occurred, the window breakage could easily be understood as neither proximately caused nor reasonably foreseeable by the robot's manufacturers and developers. Indeed, for the accident to have occurred, there had to be a series of intervening causes; the robot had to have had a set of experiences that set it on a path to break windows instead of prune leaves. And, as a result, in cases of robots that are running black-box algorithms, the best answer to the question, "What proximately caused this action?" is "The robot." Any other answer tortures the meaning of 'proximate cause.' And although the law can choose to find economic liability in parties that did not proximately cause the accident, human victims cannot necessarily choose to derive satisfaction from arbitrarily different sources than the apparent cause of their suffering. In this sense, autonomous robots are much like animals.[36] Although other parties and circumstances, including training, can be said to influence them, both autonomous robots and animals are most reasonably understood as the cause of their own actions.

Thus, in the case of learning robots, the psychological instinct to blame the robot does not rest on a fiction. The robot actually did cause the injury, and not because it was being used by or under the direction of some other being. As far as humans can understand what happened, the robot is the agent. And so in this case, the instinct to make the robot the focus of a reaction against the wrong is not based on a human brain's misunderstanding of what a robot is. It is, instead, as correct and rational as being upset with a dog for biting the mailman or a rabbit eating vegetables in one's garden.

---

36. *See* Enrique Schaerer, Richard Kelley & Monica Nicolescu, *Robots as Animals: A Framework for Liability and Responsibility in Human-Robot Interactions*, 18TH IEEE INT'L SYMP. ON ROBOT & HUM. INTERACTIVE COMM. 72, 72–77; David J. Calverley, *Android Science and Animal Rights, Does an Analogy Exist?*, 18 CONNECTION SCI. 403, 403–17 (2006).

V. THE QUESTION OF MORAL CULPABILITY

Although we might be able to say that an autonomous robot caused a harm and that no one and nothing else proximately caused the harm, it is worth considering if proximate cause is enough to override the critique that trying to exact satisfaction from robots (and animals) remains a misguided endeavor because these creatures cannot be morally blameworthy.[37] Many presume that moral responsibility attaches because humans with capacity have a choice over how they behave—because when they choose to do wrong they could have chosen to do otherwise.[38] Does the sense that blameworthiness supervenes on the existence of "free will" change whether robots should be punished for their actions to satisfy their victims? Does it change whether the victim *feels* that the punishment is morally justified?

The answer to both of these questions ought to be "no." Without weighing in too deeply to the dense philosophical literature discussing free will, we can recognize that the questions of whether humans have "free will" and what "free will" even means are among the great intractable problems at the nexus of philosophy, theology, and physics.[39] On the one hand, the physical world seems determined: if I throw a ball at the wall at a certain angle, it always bounces off the same way. If our brains and bodies simply consist of millions of small objects moving around, that environment appears to be a more complex instance of the ball and the wall. On the other hand, there are various avenues to attack the apparent determinism of the world. We know that the physical world behaves very strangely when objects are very small.[40]

---

37. *See* Peter M. Asaro, *A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 169, 176 (Patrick Lin, Keith Abney, George A. Bekey eds., 2012) ("It has been recognized that robots might be treated very much like domesticated animals, in that they clearly have some capacity for autonomous action, yet we are not inclined to ascribe to them moral responsibility, or mental culpability, or the rights that we grant to a human person.").

38. "Most philosophers suppose that the concept of free will is very closely connected to the concept of moral responsibility. Acting with free will, on such views, is just to satisfy the metaphysical requirement on being responsible for one's action." Timothy O'Connor, *Free Will*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., Spring ed. 2002), https://plato.stanford.edu/archives/spr2002/entries/freewill/.

39. Indeed, David Hume described the question of what "free will" even is as "the most contentious question of metaphysics." David Hume, *An Enquiry Concerning Human Understanding*, *in* 37 THE HARVARD CLASSICS (P.F. Collier & Son Co. 1910) (1748), at 23, *available at* http://www.bartleby.com/37/3/11.html.

40. *See generally* Scott Aaronson, *The Ghost in the Quantum Turing Machine*, *in* THE ONCE AND FUTURE TURING (S. Barry Cooper & Andrew Hodges eds., 2016), http://www.scottaaronson.com/papers/giqtm3.pdf; Jennifer Chu, *Closing the 'Free Will'*

Theologians have long argued that there is more to the story than the physical scenario I described, and philosophers have similarly tried to reconcile notions of free will with conceptions of a determined or indetermined world.[41] The question of what free will is and whether anyone has it is far from settled. As a result, resting policy or legal consequences on the notion that "one could have done otherwise" borders on incoherence absent a more robust description of what "could have done otherwise" means. We could rest legal distinctions on a weaker, more articulable characterization of free will—that nothing external to the party in question was forcing its actions and that it did what it "intended" to do. But if this is what free will means, the concept applies almost equally to humans, many animals, and autonomous robots. From this perspective, a robot is just as blameworthy as a human for its actions and just as deserving of consequences for its actions. The case for robot vengeance may even be stronger than for a human or an animal; to the extent that the harm caused by punishment or social judgment is something to be avoided as a wrong-in-itself, there is no reason to spare the rod because a robot will experience neither pain nor shame.

In fairness, if we attach moral blameworthiness to *understanding* one's actions, blameworthiness might seem to attach to some humans in a way that it does not to animals or robots.[42] However, it may again be difficult to separate humans from algorithms in this context. A robot might be able to recall and articulate *what* it did, just as a human. It may be tempting to claim that a robot cannot be blameworthy because it cannot explain *why* it acts, but often humans cannot satisfactorily explain why they act either. And, if by "understanding," we mean that something must consciously experience the connection between cause and effect, we again are thrust into a longstanding, unsolved philosophical questions: what is consciousness and what consequences flow from that state?

Questions surrounding what qualities render an actor blameworthy are vexing. But we need not resolve these questions to decide that revenge against robots can be appropriate. Even if blameworthiness does not supplement our account of why vengeance may be justified, it does not diminish other

---

Loophole, MIT NEWS (Feb. 20, 2014), https://news.mit.edu/2014/closing-the-free-will-loophole-0220.

41. *See generally* Michael McKenna & D. Justin Coates, *Compatibilism*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., Winter ed. 2016), https://plato.stanford.edu/archives/win2016/entries/compatibilism/.

42. *Cf.* ST. THOMAS AQUINAS, SUMMA THEOLOGICA 205 (Blackfriars 1975) (arguing that because animals were devoid of understanding, they could not commit a wrong) (cited in Berman, *supra* note 12, at 306).

justifications. Either a robot is as morally blameworthy and as deserving of penalty or other legal action as a human, or the robot is like a rock and is neither deserving nor undeserving of any sort of treatment. In both situations, the robot's moral status does not supply a reason to avoid taking action against it, given the presence of other reasons to do so.

VI. DESIGNING ROBOT PUNISHMENT

Media studies professor Peter Asaro wonders whether it is even "possible to punish a robot."[43] Although robots "have bodies to kick, . . . it is not clear that kicking them would achieve the traditional goals of punishment . . . retribution, reform, or deterrence."[44] But robot punishment—or more precisely, revenge against robots—primarily advances a different goal: the creation of psychological satisfaction in robots' victims.

What could robot punishment look like? Vengeful behavior seems more likely to give rise to satisfaction when revenge is accompanied by recognition that the wrongful behavior led to the vengeful action.[45] Because a robot may not be able to recognize its wrongs in a way that humans appreciate,[46] such procedures ought to be formally sanctioned or publicly recognized by an authority figure or by members of the public.

We also want to know what the effect of such actions would be on innocent third-party robot owners. The latter question is partially informed by the overwhelming criticism of civil asset forfeiture that has developed in recent years.[47] The criticism, at its core, is that it is unjust for the state to take property that was used in the commission of a crime from innocent owners because doing so primarily has the effect of punishing the owner who is innocent of any wrongdoing.[48] Perhaps the strongest argument in favor of civil

---

43. Asaro, *supra* note 37, at 181.

44. *Id.* at 182.

45. *See supra* notes 10-16 and accompanying text.

46. *See supra* note 42 and accompanying text.

47. *See, e.g.*, Shaila Dewan, *Police Use Department Wish List When Deciding Which Assets to Seize*, N.Y. TIMES (Nov. 9, 2014), https://www.nytimes.com/2014/11/10/us/police-use-department-wish-list-when-deciding-which-assets-to-seize.html; Conor Friedersdorf, *The Injustice of Civil-Asset Forfeiture*, ATLANTIC (May 12, 2015), https://www.theatlantic.com/politics/archive/2015/05/the-glaring-injustice-of-civil-asset-forfeiture/392999/; Sarah Stillman, *Taken*, NEW YORKER (Aug. 12, 2013), http://www.newyorker.com/magazine/2013/08/12/taken; *Asset Forfeiture Abuse*, ACLU, https://www.aclu.org/issues/criminal-law-reform/reforming-police-practices/asset-forfeiture-abuse (last visited Feb. 18, 2018); *Policing for Profit*, INSTITUTE FOR JUSTICE, http://ij.org/report/policing-for-profit/ (last visited Feb. 24, 2018).

48. *See* Dewan, *supra* note 47; Friedersdorf, *supra* note 47. Asset forfeiture is also often disproportionately targeted at disadvantaged groups. *See* Rebecca Vallas et al., *Forfeiting the*

asset forfeiture is that, with a nod to products liability law, it creates incentives for individuals to be particularly vigilant about how their property is used, but there seems to be little evidence that any societal benefit from owner vigilance outweighs the tremendous harm owners can face when their property is taken from them due to the actions of another.

But civil asset forfeiture is a different case than robot forfeiture because civil forfeiture more often involves objects (money, cars, goods) rather than agents (animals and robots). A robot is the cause of harm in a way that a (human-driven) car cannot be. The car is a mere tool of a human who intends or negligently does harm, but an autonomous robot is not merely a tool of another but is itself an agent. Just as the law sometimes sanctions euthanizing dangerous dogs over the objections of their owners,[49] the law might reasonably say that robots who have caused certain kinds of harm are also forfeit from their owners. Knowing this is possible, robot owners will have increased incentives to take care in training and controlling their robots (even though owners cannot know exactly what will work and will not necessarily have complete control over whether they are successful). The possibility of having to forfeit the robot might also provide incentives for the creation of insurance against a robot going rogue.[50] Alternatively, though perhaps less satisfyingly, the law could follow some later incarnations of the medieval deodand system and allow owners to pay the value of their robots without giving up the actual objects themselves.[51]

One could also imagine situations where the notion of separating a rogue robot from its owner would create a disproportionate burden on the owner, for example if a robot was unique, unusually expensive relative to the harm caused, or difficult to replace. In this situation, we could countenance more modest actions that might still provide a degree of satisfaction to harmed

---

American Dream: How Civil Asset Forfeiture Exacerbates Hardship for Low-Income Communities and Communities of Color, CENTER FOR AMERICAN PROGRESS (Apr. 2016), https://cdn.americanprogress.org/wp-content/uploads/2016/04/01060039/CivilAssetForfeiture-reportv2.pdf.

49. *See, e.g.*, CAL. FOOD & AGRIC. CODE § 31645 (West 2001) ("A dog determined to be a vicious dog may be destroyed by the animal control department when it is found . . . that the release of the dog would create a significant threat to the public health, safety, and welfare.").

50. *See* Calo, *supra* note 21, at 609–11.

51. Anna Pervukhin, *Deodands: A Study in the Creation of Common Law Rules*, 47 AM. J. LEGAL HIST. 237, 237 (2005) ("Under [the law of deodands], a chattel . . . was deemed to be a deodand [and was forfeited to the English crown] whenever a coroner's jury decided that it had caused the death of a human being . . . . In practice, deodands were rarely taken away from their owners. In most cases, the jury that adjudged the deodand also appraised its value; owners were then expected to pay a fine equal to the value of the deodand.").

parties, such as requiring someone to evaluate the robot's code and determine if there was a way to avoid the future harm, either by adding to the robot's program or erasing some or all of its memory and forcing it to relearn how to behave.

But at the end of the day, the most satisfying outcome for a person wronged by a robot might be the early Middle Age practice of "noxal surrender," wherein "animals or objects causing serious damage or death, called banes, were handed over directly to a victim or to his family."[52] The opportunity to take control of a robot for one's own purposes or to destroy it, when combined with the social signal of being given the device by law, could together best serve to provide satisfaction to victims. In which case, a wronged party may indeed be quite justified in dragging a robot out into an empty field and walloping it with a baseball bat.

VII. CONCLUSION

This Essay makes an outlandish argument, and yet robots are beginning to make our present environment just as bizarre. Human interactions are complex and subtle; we are constantly sending each other signals that create and alter our relationships with one another. Inserting autonomous and social robots into humans' experiences will alter and confuse those experiences in ways that are difficult to anticipate. While our collective response will likely not be to return to a literally medieval system of law, the task of understanding how the addition of robots into our lives will affect us psychologically and emotionally remains a critical component of how we address the technologically-changing world.

---

52. *Id.* at 241 (citing Jacob J. Finkelstein, *The Goring Ox: Some Historical Perspectives on Deodands, Forfeitures, Wrongful Death and the Western Notion of Sovereignty*, 46 TEMP. L.Q. 169, 181 (1973)); THE LAWS OF THE EARLIEST ENGLISH KINGS 71 (F. L. Attenborough ed., Univ. Press 1922). For hundreds of years, legal actions were taken against animals and objects for the harm they caused to humans. *See* E.P. EVANS, THE CRIMINAL PROSECUTION AND PUNISHMENT OF ANIMALS (1906); Walter Woodburn Hyde, *The Prosecution and Punishment of Animals and Lifeless Things in the Middle Ages and Modern Times*, 64 U. PA. L. REV. 696, 706 (1916). In Europe and other locations, from around the ninth to nineteenth century, "[i]ndividual animals were tried—usually for killing human beings—in secular courts according to common law precedents . . . [and] many animals were tried in groups as public nuisances before ecclesiastical tribunals." Berman, *supra* note 12, at 289.