Northwestern University Law Review Vol. 111

AMORAL MACHINES, OR: HOW ROBOTICISTS CAN LEARN TO STOP WORRYING AND LOVE THE LAW

Bryan Casey

ABSTRACT—The media and academic dialogue surrounding high-stakes decisionmaking by robotics applications has been dominated by a focus on morality. But the tendency to do so while overlooking the role that legal incentives play in shaping the behavior of profit-maximizing firms risks marginalizing the field of robotics and rendering many of the deepest challenges facing today's engineers utterly intractable. This Essay attempts to both halt this trend and offer a course correction. Invoking Justice Oliver Wendell Holmes's canonical analogy of the "bad man . . . who cares nothing for . . . ethical rules," it demonstrates why philosophical abstractions like the trolley problem—in their classic framing—provide a poor means of understanding the real-world constraints robotics engineers face. Using insights gleaned from the economic analysis of law, it argues that profit-maximizing firms designing autonomous decisionmaking systems will be less concerned with esoteric questions of right and wrong than with concrete questions of predictive legal liability. Until such time as the conversation surrounding so-called "moral machines" is revised to reflect this fundamental distinction between morality and law, the thinking on this topic by philosophers, engineers, and policymakers alike will remain hopelessly mired. Step aside, roboticists—lawyers have this one.

AUTHOR— Bryan Casey, CodeX Fellow, The Stanford Center for Legal Informatics. J.D. Candidate, Stanford Law School, Class of 2018. The author particularly thanks A. Mitchell Polinsky, the Josephine Scott Crocker Professor of Law and Economics at Stanford Law School, for his generous support.

You can see very plainly that a bad man has as much reason as a good one for wishing to avoid an encounter with the public force, and therefore you can see the practical importance of the distinction between morality and law. A man who cares nothing for an ethical rule which is believed and practised by his neighbors is likely nevertheless to care a good deal to avoid being made to pay money, and will want to keep out of jail if he can.

—Oliver Wendell Holmes, Jr.[1]

INTRODUCTION

It is the year 2031, and an autonomous vehicle faces a grave choice: it must either collide with a child who unexpectedly runs into its path or swerve so violently as to overturn the car along with the passenger inside. Given the vehicle's speed, either choice spells serious injury—even death. Which should it choose: passenger or child?

In October 2016, Mercedes-Benz[2] made global headlines by becoming one of the first major automakers to proffer an answer to this question of philosophical proportions.[3] Speaking to *Car and Driver Magazine*,

[1] *The Path of the Law*, 10 HARV. L. REV. 457, 459 (1897).

[2] The type of futuristic scenario outlined above may arrive much sooner than 2031. In January 2017, Uber's then-CEO Travis Kalanick announced that Daimler would "introduce and operate their own self-driving cars on Uber's ridesharing network" in "the coming years." Travis Kalanick, *Uber and Daimler Join Forces on Self-Driving Cars*, UBER NEWSROOM (Jan. 31, 2017), https://newsroom.uber.com/uber-daimler-self-driving-cars/ [https://perma.cc/QRA8-VJ6X]. Uber's partnerships with pioneering manufacturers of self-driving vehicles do not end with Daimler, however. *See* Bryan Casey, *A Loophole Large Enough to Drive an Autonomous Vehicle Through*, 63 STAN. L. REV. ONLINE 73, 80 (2016) (discussing Uber's use of self-driving Volvo XC90 SUVs); Andrew Hawkins, *Meet Uber's First Self-Driving Car*, VERGE (May 19, 2016, 8:07 AM), http://www.theverge.com/2016/5/19/11711890/uber-first-image-self-driving-car-pittsburgh-ford-fusion [https://perma.cc/K95L-3SG9] (discussing Uber's use of self-driving Ford Fusions); *see also* Alex Davies, *Uber's Self-Driving Crash Proves We Need Self-Driving Cars*, WIRED (Mar. 25, 2017, 3:10 PM), https://www.wired.com/2017/03/uber-self-driving-crash-tempe-arizona/ [https://perma.cc/W4YJ-D6LF] (describing the benefits of self-driving Ubers and as well as the need for increased testing).

[3] David Z. Morris, *Mercedes-Benz's Self-Driving Cars Would Choose Passenger Lives Over Bystanders*, FORTUNE (Oct. 15, 2016), http://fortune.com/2016/10/15/mercedes-self-driving-car-ethics/ [https://perma.cc/F55P-YQVT]; *see also* Mike Brown, *Mercedes's Self-Driving Cars Will Kill Pedestrians Over Drivers*, INVERSE MAG. (Oct. 14, 2016), https://www.inverse.com/article/22204-mercedes-benz-self-driving-cars-ai-ethics [https://perma.cc/WKT4-6NZZ]; Lindsay Dodgson, *Why Mercedes Plants to Let Its Self-Driving Cars Kill Pedestrians in Dicey Situations*, BUS. INSIDER (Oct.

Mercedes executive Christoph von Hugo implied that, in such a situation, the company's "future autonomous [vehicles] will save the car's driver and passengers, even if that means sacrificing the lives of pedestrians."[4] According to von Hugo, the rationale was simple: "If all you know for sure is that one death can be prevented, then that's your first priority."[5]

Within days, however, von Hugo's words raised a media fracas large enough to elicit an official clarification from the auto giant.[6] In response to criticism over the executive's statement, Mercedes-Benz claimed von Hugo had been misquoted, insisting that the company "continue[d] to adhere to the principle of providing the highest possible level of safety for all road users."[7]

But, contentious as von Hugo's particular rationale proved to be, virtually no experts contest the powerful assumption underlying the rationale.[8] For better or worse, a world where millions of robots will be entrusted with making life-or-death decisions is no longer the exclusive province of science fiction.[9] It is a fast-approaching reality—one that, many leading scholars argue, will require "bring[ing] artificial agents into the domain of ethics."[10]

These scholars belong to a rapidly advancing field known as "machine ethics," which seeks to "frame [the] discussion" surrounding morally consequential robots "in a way that constructively guides the engineering task of designing" them.[11] Those leading the charge argue that engineers "need to embed in [robots] some ethical guidelines,"[12] so they will "do the

12, 2016, 10:49 AM), http://www.businessinsider.com/mercedes-benz-self-driving-cars-programmed-save-driver-2016-10 [https://perma.cc/GJ5L-DNB8]; Natalie Walters, *Mercedes-Benz Will Let Self-Driving Cars Hit Pedestrians to Protect Passengers*, STREET (Oct. 13, 2016, 12:25 PM), https://www.thestreet.com/story/13852444/1/mercedes-benz-will-let-self-driving-cars-hit-pedestrians-to-protect-passengers.html [https://perma.cc/7K2L-MP2N].

[4] *See* Morris, *supra* note 3.

[5] Michael Taylor, *Self-Driving Mercedes-Benzes Will Prioritize Occupant over Pedestrians*, CAR & DRIVER MAG. (Oct. 7, 2016, 5:27 PM), http://blog.caranddriver.com/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/ [https://perma.cc/45UE-7JMQ].

[6] *See* Morris, *supra* note 3.

[7] *See Daimler Clarifies: Neither Programmers Nor Automated Systems Are Entitled to Weigh the Value of Human Lives*, DAIMLER (Oct. 18, 2016), http://media.daimler.com/marsMediaSite/en/instance/ko/Daimler-clarifies-Neither-programmers-nor-automated-systems-.xhtml?oid=14131869 [https://perma.cc/X22B-UW5G].

[8] *See infra* notes 51–63 and accompanying text.

[9] *See supra* note 3 and accompanying text.

[10] WENDELL WALLACH & COLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG 16 (2009) [hereinafter MORAL MACHINES].

[11] *Id.* at 6, 34.

[12] Joshua Greene et al., *Embedding Ethical Principles in Collective Decision Support Systems*, 30 PROC. AAAI CONF. ON ARTIFICIAL INTELLIGENCE 4147 (2016).

right thing"[13] in morally loaded situations like those anticipated by von Hugo.[14] This, they insist, "is fundamentally an ethics problem"[15]—one that "will require looking at the origins of human morality"[16] to light the way for engineers designing "moral machines"[17] with "values [that are] clear and consistent."[18]

But, compelling as this lofty vision of robotics engineering may appear at first blush, there is just one tiny detail standing in its way: the entire legal system. After all, liability for injury is governed not by moral codes, but by legal codes. Properly understood, the "practical importance of th[is] distinction between morality and law"[19]—to appropriate the canonical words of Justice Oliver Wendell Holmes—holds profoundly different implications for the role of ethics in robotics engineering.[20]

Economic analysis of law teaches that profit-maximizing firms will design their robots to behave not as good moral philosophers, but as Holmesian bad men—concerned less with "ethical rule[s]" than with the legal rules that dictate whether they will be "made to pay money" and can "keep out of jail."[21] Far from following a "clear and consistent"[22] moral code, optimized systems will instead follow an amoral code that reflects the messy economic realities of society's imperfect legal regimes. These robots will not maximize morality, but minimize liability. And if the goal of machine ethics "is to frame discussion in a way that constructively guides the engineering task of designing" these so-called "moral machines,"[23] it must begin by recognizing that high-minded ethics will surely take a backseat to amoral economics.

Yet "dismal"[24] as this description may initially seem, its implications are surprisingly sanguine. Contrary to the current consensus within the field

---

[13] Kris Hammond, *Ethics and Artificial Intelligence: The Moral Compass of a Machine*, RECODE (Apr. 13, 2016, 2:22 PM), http://www.recode.net/2016/4/13/11644890/ethics-and-artificial-intelligence-the-moral-compass-of-a-machine [https://perma.cc/X85D-2FLD].

[14] *See infra* notes 51–63 and accompanying text.

[15] Patrick Lin, *Why Ethics Matters for Autonomous Cars*, *in* AUTONOMOUS DRIVING: TECHNICAL, LEGAL AND SOCIAL ASPECTS 69, 73 (Markus Maurer et al. eds, 2015).

[16] MORAL MACHINES, *supra* note 10, at 8.

[17] *Id.*

[18] Joshua Greene, *Our Driverless Dilemma*, 352 SCIENCE 1514, 1515 (2016); *see also infra* notes 51–63 and accompanying text.

[19] *The Path of the Law*, *supra* note 1, at 459.

[20] *See infra* notes 64–67 and accompanying text.

[21] *The Path of the Law*, *supra* note 1, at 459.

[22] Greene, *supra* note 18, at 1515.

[23] MORAL MACHINES, *supra* note 10, at 6.

[24] *See* ROBERT DIXON, THE ORIGIN OF THE TERM "DISMAL SCIENCE" TO DESCRIBE ECONOMICS 1 (1999) (discussing Thomas Carlyle's famous phrase).

of machine ethics, the true designers of machine morality will not be the cloistered engineering teams of tech giants like Google, Tesla, or Mercedes, but ordinary citizens. As democratic stakeholders, they alone will possess the power to narrow the gap "between morality and law."[25] It will be their collective engineering task to design a legal system that ensures "a bad [robot] has as much reason as a good one" to behave ethically—ultimately, rendering meaningless the "practical importance of the distinction" between amoral machines and moral machines.[26]

I. A CRASH COURSE IN MACHINE ETHICS

Our machines need not hesitate when they see the Trolley coming. They will act in accord with whatever moral or ethical code we provide them and the value determinations that we set.

—Kris Hammond[27]

Modern robotics[28] begins with Isaac Asimov.[29] More than half a century ago, the beloved science fiction writer inspired a generation of rising engineers to begin "thinking about how minds might work"[30] by envisioning a future inhabited by "mechanical men" whose intellectual and physical abilities rivaled those of their biological counterparts.[31] Mindful of the profound implications posed by these agents, however, Asimov's

---

[25] *See The Path of the Law*, *supra* note 1, at 459.

[26] *Id.*

[27] Hammond, *supra* note 13.

[28] "Few complex technologies have a single, stable, uncontested definition. Robots are no exception." Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 529 (2015). For stylistic purposes, this Essay uses "robot" and "robotics" interchangeably with "artificial intelligence" ("AI"). All three terms lack a universally accepted definition but, in this Essay, refer broadly to any "computerized system that exhibits behavior that is commonly thought of as requiring intelligence." EXEC. OFFICE OF THE PRESIDENT NAT'L SCI. & TECH. COUNCIL COMM. ON TECH., PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE 6 (Oct. 2016). Further, this Essay's scope is limited to "Narrow AI," which refers to intelligent computer systems capable of "address[ing] specific application areas." *Id.* at 7. It does not extend to "General AI," which "refers to a notional future AI system that exhibits apparently intelligent behavior at least as advanced as a person across the full range of cognitive tasks." *Id*. The "current consensus of the private-sector expert community . . . is that General AI will not be achieved for at least decades." *Id.*

[29] NILS J. NILSSON, THE QUEST FOR ARTIFICIAL INTELLIGENCE: A HISTORY OF IDEAS AND ACHIEVEMENTS 25 (2010), http://ai.stanford.edu/~nilsson/QAI/qai.pdf [https://perma.cc/CEY9-EVY7] (noting that the "quest for artificial intelligence, quixotic or not, begins with dreams like [Asimov's]").

[30] John Markoff, *Technology; A Celebration of Isaac Asimov*, N.Y. TIMES (Apr. 12, 1992), http://www.nytimes.com/1992/04/12/business/technology-a-celebration-of-isaac-asimov.html?pagewanted=all [https://perma.cc/JZ8H-8NE8] (quoting the computer science pioneer and founder of MIT's Artificial Intelligence Laboratory, Marvin Minsky, who wrote: "After [Asimov's story] 'Runaround' appeared in . . . March 1942 . . . I never stopped thinking about how minds might work.").

[31] ISAAC ASIMOV, *Runaround*, *in* I, ROBOT (Bantam Dell 2004) (1950).

fictional engineers programmed the "robots" with directives meant to ensure their unwavering alignment with human values:

> A robot may not injure a human being or, through inaction, allow a human being to come to harm.
>
> A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
>
> A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. [32]

With these three precepts in place, Asimov's engineers believed they could rest easy—having instilled in their robots a set of laws sufficient to steer them safely through any conceivable dilemma.[33] If only life were so simple.[34]

Among the many shortcomings of Asimov's speciously watertight set of laws is its incompatibility with another set of slightly greater notoriety—Newton's. An object in motion tends to stay in motion, after all. And if a contemporary robot, such as a driverless car, were to unexpectedly encounter a jaywalker while travelling at full speed, sheer momentum could force the vehicle into a tragic choice between a collision with the pedestrian or an avoidance maneuver so violent as to endanger the passenger.[35] Violating Asimov's First Law, under such circumstances,

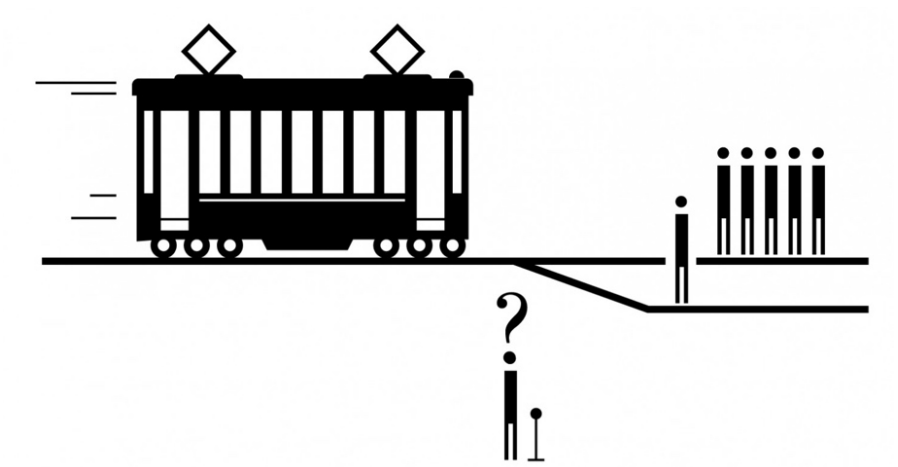
---

[32] *Id.* at 37.

[33] *Id.*

[34] Asimov was not naïve to the fallibility of his laws. *See* Keith Abney, *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 35, 43 (Patrick Lin et al. eds., 2012) ("[I]n story after story, Asimov demonstrated that [his] three simple, hierarchically arranged rules could lead to deadlocks when, for example, the robot received conflicting instructions from two people, or when protecting one person might cause harm to others."). Asimov remarked that "[t]here was just enough ambiguity in the Three Laws to provide the conflicts and uncertainties required for new stories, and, to my great relief, it seemed always to be possible to think up a new angle out of the sixty-one words of the Three Laws." ISAAC ASIMOV, THE REST OF THE ROBOTS 43 (1964).

[35] This claim has been convincingly argued by experts too numerous to list exhaustively. *See, e.g.*, Lin, *supra* note 15, at 75 (lamenting the "daunting number of factors" to account for in an inevitable collision between a driverless car and another person or object); Thierry Fraichard & Hajime Asama, *Inevitable Collision States: A Step Towards Safer Robots?*, 18 ADVANCED ROBOTICS 1001, 1001 (2004) (describing the "inevitable collision state" of a robotic system, which occurs when the future trajectory of the system will inevitably lead to a collision); Noah Goodall, *Ethical Decision Making During Automated Vehicle Crashes*, TRANSP. RES. REC. J. TRANSP. RES. BOARD 1, 3 (2014) (noting that "[w]hile any engineering system can fail, it is important to distinguish that, for automated vehicles, even a perfectly-functioning system cannot avoid every collision"); Jeffrey K. Gurney, *Crashing into the Unknown: An Examination of Crash-Optimization Algorithms Through the Two Lanes of Ethics and Law*, 79 ALB. L. REV. 183, 195–99 (2016) (discussing the implications behind an autonomous vehicle having to hit one of two motorcycles or cars with drivers present).



would be unavoidable. The question would thus become one of triage: who should the robot imperil?

This type of lesser-of-evils dilemma, where injury is both inevitable *and* variable, is known as a "trolley problem"—a term coined by the philosopher Judith Thomson in a now-classic thought experiment dating back to 1976.[36] In its most popular form, the experiment posits an observer who is witness to a runaway trolley car barreling toward five unwitting workers on the tracks ahead.[37] The observer, however, is standing at a switch. If pulled, the switch will divert the trolley onto another track where only one unlucky worker awaits.[*]

Tragedy of some kind is foreordained, but the observer holds the proverbial power to steer fate: "turn the trolley, killing the one," or "refrain from turning the trolley, killing the five?"[38]

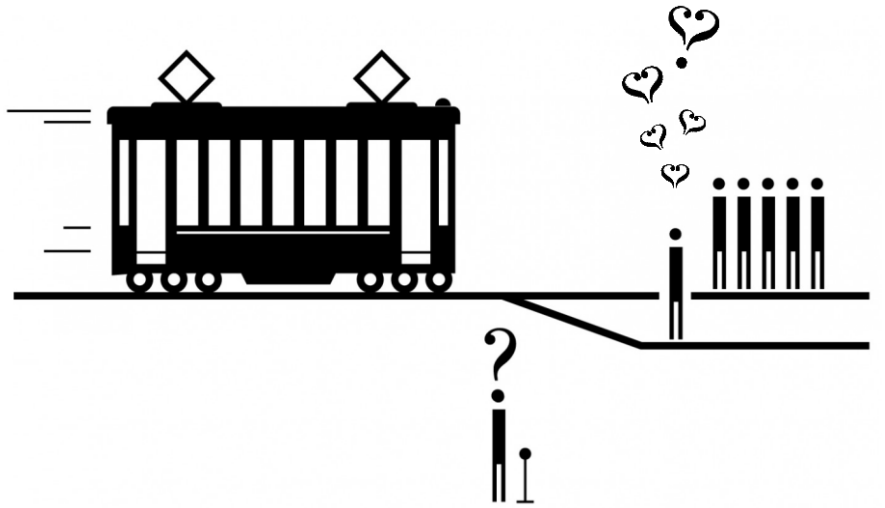
[36] Judith Jarvis Thomson, *Killing, Letting Die, and the Trolley Problem*, 59 MONIST 204, 206 (1976). Although Thomson coined the term "trolley problem," the first articulation of the thought experiment originated with the philosopher Philippa Foot. *See* Philippa Foot, *The Problem of Abortion and the Doctrine of Double Effect*, 5 OXFORD REV. 1, 3 (1967).

[37] Thomson's experiment asked subjects to imagine themselves as the trolley driver rather than as an outside observer at a switch.

[*] Illustrations by Samuel Granados of *The Washington Post*; reused with permission. Some images have been slightly altered for pedagogical purposes.

[38] *See* Thomson, *supra* note 36, at 206.

When surveyed, the vast majority of respondents—usually around 90%—choose to divert the trolley away from the five.[39] But anyone tempted to interpret these results as a sign of moral consensus should first consult the battery of similar surveys showing that even the slightest of modifications to the original experiment is apt to elicit a dramatically different response. Ask, for example, whether respondents would save five workers by steering the trolley toward a relative or loved one, and the percentage that pulls the switch takes a significant downward turn.[40]



The consequences, in their most abstract sense, remain the same: sacrifice one to spare five. But for many, the moral intuition changes—often without a consistent rationale.[41]

Even a formal philosophical education offers no antidote to this apparent moral confusion. According to a 2009 survey conducted by David Bourget and David Chalmers, widespread disagreement over the appropriate response to Thomson's original experiment exists even among professional philosophers.[42] Of those polled, only 68% reported they would

[39] John Cloud, *Would You Kill One Person to Save Five? New Research on a Classic Debate*, TIME (Dec. 5, 2011), http://healthland.time.com/2011/12/05/would-you-kill-one-person-to-save-five-new-research-on-a-classic-debate/ [https://perma.cc/U8K3-GQET].

[40] April Bleske-Rechek et al., *Evolution and the Trolley Problem: People Save Five Over One Unless the One is Young, Genetically Related, or a Romantic Partner*, 4 J. SOC., EVOLUTIONARY, & CULTURAL PSYCHOL. 115, 119–21, 124 (2010), http://psycnet.apa.org/journals/ebs/4/3/115.pdf [https://perma.cc/Q9Z5-EEBX] (finding that only 24.8% of study participants would have pulled the lever on the lone target when the target was a romantic partner).

[41] *Id.* at 126.

[42] David Bourget & David Chalmers, *What Do Philosophers Believe?*, 170 PHIL. STUD. 465, 477 (2014).

pull the switch—with a further 8% electing not to intervene, and the remainder indicating indecision of some sort.[43] Although it is easy to dismiss these scattershot opinions as yet further evidence of the perennial stereotype of the "contrarian philosopher," the fact remains that those who have had most occasion to contemplate the perplexities of the problem are those most perplexed by it.

After spending decades as an object of little more than academic fascination, however, the trolley problem recently crashed into the cultural mainstream thanks, in part, to the rapid advances made in robotic vehicle technology. Nowadays, killer robot spinoffs of Thomson's classic thought experiment—pitting humanity's fuzzy moral intuitions against the steely logic of consumer-ready driverless cars—are as apt to appear in popular media outlets as in scholarly journals. This time in the limelight, however, has brought the problem no nearer to a universal resolution. In fact, the opposite has occurred. The problem has come to stand as a synecdoche for the dizzying complexity of humanity's deepest-seated moral differences.

But while society as a whole may remain agnostic to the trolley problem, engineers at the cutting edge of robotics are afforded no such luxury. For them, trolley-like problems are not mere philosophical curiosities. They are real-world contingencies that require prospective programming.[44] Those designing the decisionmaking systems behind autonomous vehicles, weaponized drones, and countless other emerging robotics applications cannot simply shrug their shoulders. Rather, they must decide in advance how their systems will respond when life and limb are on the line. Resolution of some sort is simply unavoidable. As the scholars Sven Nyholm and Jilles Smids note, even choosing *not* to program a response to morally loaded situations "amounts to knowingly relinquishing the important responsibility we have to try to control" such systems.[45]

Though a world where robots hold human lives in the balance may still read like a description better suited for Asimov's science fiction, it is a modern engineering reality. These machines are not merely coming—many have already arrived. And it is an unflinching recognition of this fact that has spurred many leading scholars to call for a moral reckoning in the field

---

[43] *Id.*

[44] *See, e.g.*, Sven Nyholm & Jilles Smids, *The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?*, 19 J. ETHICAL THEORY & PRACT. 1276, 1278 (2016) (showing why a prospective engineering approach to even edge cases is necessary)*; see also* Goodall, *supra* note 35, at 3 (explaining inevitable collision states).

[45] *See* Nyholm & Smids, *supra* note 44, at 1279.

of robotics—with, perhaps, no work more influential than *Moral Machines: Teaching Robots Right from Wrong*.[46]

II. OF MACHINES AND (BAD) MEN

> *In the future, moral philosophy will be a key industry sector . . . . You would want . . . [robots] preloaded with a pretty good set of values . . . . So presumably the robot companies will get their values loaded into the robot from a values company.*
>
> —Stuart Russell[47]

The year is 2009, and visionary tech companies like Google are only beginning to take their first secretive steps toward developing autonomous vehicles.[48] DeepMind's earth-shaking "Go" victory over the grandmaster Lee Sedol—the robotics community's own "Kasparov moment"[49]—will not come for another half-decade.[50] To have surveyed expert opinion for the most ethically exacting emerging technologies then would have likely produced a shortlist dominated by the fields of nanotechnology, genetic engineering, or perhaps stem cell research. But Yale University's Wendell Wallach and co-author Colin Allen saw the writing on the wall—and it was written in computer code.

In their 2009 book, *Moral Machines: Teaching Robots Right from Wrong*, the two laid out an urgent—and strikingly prescient—vision of a near future replete with "autonomous [robotics] systems . . . increasingly in charge of a variety of decisions that have ethical ramifications."[51] And

---

[46] *See infra* notes 48–56 and accompanying text.

[47] Queena Sook Kim, *Stuart Russell on Why Moral Philosophy Will Be Big Business in Tech*, KQED NEWS (Oct. 27, 2015), https://ww2.kqed.org/news/2015/10/27/stuart-russell-on-a-i-and-how-moral-philosophy-will-be-big-business/ [https://perma.cc/NGU5-8HHF] (quoting Stuart Russell, co-author of *Artificial Intelligence: A Modern Approach*). *Artificial Intelligence: A Modern Approach* is considered the standard textbook in the field, used by over 1,334 universities in 118 countries. *1334 Schools Worldwide That Have Adopted AIMA*, U.C. BERKELEY (Feb. 20, 2017), http://aima.cs.berkeley.edu/adoptions.html [https://perma.cc/5GVN-AZZ5].

[48] Adam Fisher, *Google's Self-Driving Cars: A Quest for Acceptance*, POPULAR SCI. (Sept. 18, 2013), http://www.popsci.com/cars/article/2013-09/google-self-driving-car [https://perma.cc/V5PK-Z2VT].

[49] The phrase refers to IBM Deep Blue's defeat of the world chess champion Garry Kasparov in 1997. *See* Matt McFarland, *Google Just Mastered a Game That Vexed Scientists—and Their Machines—For Decades*, WASH. POST (Jan. 27, 2016), https://www.washingtonpost.com/news/innovations/wp/2016/01/27/google-just-mastered-a-game-thats-vexed-scientists-for-decades/ [https://perma.cc/DGR7-LADJ] (discussing the phrase "Kasparov moment").

[50] Cade Metz, *In a Huge Breakthrough, Google's AI Beats a Top Player at the Game of Go*, WIRED MAG. (Jan. 27, 2016), https://www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/ [https://perma.cc/NU9X-3FGZ]. Go is an ancient Eastern strategy game that is comparable to chess, though far more computationally complex. *Id.*

[51] MORAL MACHINES, *supra* note 10, at 15.

while acknowledging that "a concern for safety and societal benefits ha[d] always been at the forefront of engineering," the scholars cautioned that preventing this sci-fi-like future from taking a dystopian turn demanded a qualitatively different approach to designing complex robotics—one that went far "beyond traditional product safety."[52]

Their solution? An engineering undertaking of Asimovian ambition: "to bring artificial agents into the domain of ethics" by programming them in accordance with explicit "moral standards."[53] But rising to meet "the engineering challenge entailed in going from Aristotle to Asimov and beyond," they stressed, would be no easy feat.[54] Rather, it would "require looking at the origins of human morality as viewed in the fields of evolution, learning and development, neuropsychology, and philosophy" to apply "[t]he values and concerns expressed in the world's religious and philosophical traditions . . . to machines."[55] Thus, they called for scholars of diverse backgrounds to coalesce around a formal "discipline of artificial morality" which, they hoped, would "frame discussion in a way that constructively guide[d] the engineering task" of "building explicit ethical reasoning into [robotics] system[s]."[56]

Since *Moral Machines*'s call to ethical arms in 2009, the number of scholars to have joined the ranks of this emerging discipline—now widely known as "machine ethics"—has grown too large to catalog, except by representative sample. Like Wallach and Allen, these scholars assert that robotics engineers will "need to embed in [machines] some ethical guidelines, so they can act in their environment following values that are aligned to the human ones."[57] Doing so, they insist, will require first "figur[ing] out how to make our values clear and consistent."[58] This, they argue, "is fundamentally an ethics problem"[59]—one that may mean big business for "moral philosophy" in the future.[60] Indeed, with society "about to endow millions of vehicles with autonomy," they say, "serious consideration of algorithmic morality has never been more urgent."[61] Yet, whether these algorithms should ultimately be "deontological,"

---

[52] *Id.* at 4, 17.

[53] *Id.* at 16, 78–81.

[54] *Id.* at 8.

[55] *Id.*

[56] *Id.* at 6, 30–32.

[57] Greene et al., *supra* note 12, at 4147.

[58] Greene, *supra* note 18, at 1515.

[59] Lin, *supra* note 15, at 73.

[60] *See* Kim, *supra* note 47.

[61] Jean-François Bonnefon, Azim Shariff & Iyad Rahwan, *The Social Dilemma of Autonomous Vehicles*, 352 SCIENCE 1573, 1576 (2016).

"consequentialist," or of another philosophic school entirely, is still a matter of debate.[62] "For 21st-century moral philosophers," they predict, "this may be where the rubber meets the road."[63]

But compelling as this ethics-based vision of robotics engineering may appear at first blush, it essentially ignores the fact that the law already occupies the field. After all, liability for injury is governed by legal codes, not moral codes. And in a world where firms can influence their threat of public sanction by acting in accordance with legal standards, economics teaches that man's laws—not moral laws—can be expected to play the dominant role in shaping profit-maximizing behavior.[64]

This insight, known as liability minimization, is "foundational" to economic analysis of law.[65] Its premise is straightforward. For profit-maximizing firms, an "ex post damages system translates into ex ante [changes to behavior] because the prospect of civil liability encourages organizations . . . to minimize the justiciable harm they cause."[66] The crucial adjective, of course, is "justiciable." Rational, forward-looking firms are not incentivized to minimize harm per se, but only those harms that the governing liability regime will force them to pay for.[67]

To illustrate the allocative role that the legal system plays in shaping liability-minimizing behavior, consider the following hypothetical. A group of teenagers is playing a ball game known as Jackpot. The game involves a designated thrower, plus five others vying to catch the ball when it is lobbed in their direction. In this instance, one such lob goes awry—sending the ball bouncing into the street. Eager to be the first to retrieve it, all five teenagers dash across the road—directly into the path of an oncoming autonomous vehicle whose view of them, until that moment, had been obscured by parked cars lining the sidewalk. Even with its superhuman reaction time, the vehicle's system cannot safely avoid a collision. Instead,

---

[62] J. Christian Gerdes & Sarah Thornton, *Implemental Ethics for Autonomous Vehicles*, *in* AUTONOMOUS DRIVING: TECHNICAL, LEGAL AND SOCIAL ASPECTS, *supra* note 15, at 90–94.

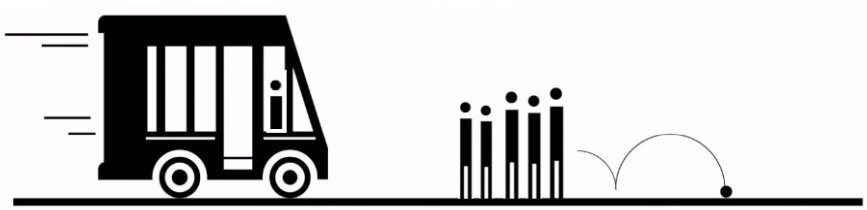
[63] *See* Greene, *supra* note 18, at 1515.

[64] *See infra* notes 72–75 and accompanying text.

[65] Margo Schlanger, *Operationalizing Deterrence: Claims Management*, 2 J. TORT L. 1 (2008).

[66] As Schlanger notes, "Liability minimization is, of course, weighted against accident-prevention costs." *Id.* at 1 n.1. *See also* GUIDO CALABRESI, THE COSTS OF ACCIDENTS: A LEGAL AND ECONOMIC ANALYSIS 26 (1970) ("I take it as axiomatic that the principal function of accident law is to reduce the sum of the costs of accidents and the costs of avoiding accidents."); Richard Posner, *A Theory of Negligence*, 1 J. LEGAL STUD. 29, 33 (1972) ("If . . . the benefits in accident avoidance exceed the costs of prevention, society is better off if those costs are incurred and the accident averted, and so in this case the enterprise is made liable, in the expectation that self-interest will lead it to adopt the precautions in order to avoid a greater cost in tort judgments.").

[67] *See* Louis Kaplow & Steven Shavell, *Economic Analysis of Law*, *in* HANDBOOK OF PUBLIC ECONOMICS 1661, 1667–82 (A.J. Auerbach and M. Feldstein eds., 2002).



it faces a choice. It must either strike the five teenagers or swerve so suddenly as to overturn the vehicle and the passenger inside. Either option is likely to result in serious injury, or even death.

But rather than stopping the description short, let us bring the hypothetical a step closer to reality by adding a final consideration that the trolley problem so conspicuously lacks: a legal liability regime.

Imagine, further, that the vehicle is operating within a jurisdiction that holds firms "strictly liable" for any damages their autonomous vehicles cause—meaning, in this instance, that the firm must pay for any harm done to the teenagers regardless of whether or not they are at fault for the accident.[68] With this liability regime in view, the "optimal" choice for a profit-maximizing firm suddenly becomes straightforward. Given that the firm must pay for all resultant injuries, the decision boils down to simple arithmetic. A compensatory payout to one victim is cheaper than similarly expensive payouts to five. Thus, a profit-maximizing firm's autonomous vehicle should swerve.[69]

[68] *Id.* at 1667.

[69] This is, of course, a necessary oversimplification. An essay of far greater length than this one could be written on the warped incentive signals conceivably sent by transaction costs, first- and third-party insurance intermediaries, administrative costs, technical limitations, agency costs, information costs, human error and incompetence, consumer psychology, potential media backlash, and judicial and regulatory uncertainty. But as Schlanger herself emphatically states, "[A]ll these caveats operate only at the edges of the main point, which is that . . . damage actions function to price and internalize to risk-creating organizations many harms . . . that would otherwise remain externalities." *See* Schlanger, *supra* note 65, at 4. Robot behavior that accounts for these "caveats" will be no more ethical than in the simplified hypothetical above. Rather, the robots' profit-maximizing behavior will simply be more nuanced.

JACKPOT HYPOTHETICAL: STRICT LIABILITY

Now imagine an identical hypothetical, but this time in a jurisdiction that recognizes the legal defense of "contributory negligence"—meaning, in this instance, that the firm will not be held liable for injury caused by its autonomous vehicle if the victims' own negligence contributed to the accident.[70] Under this liability regime, the optimal economic choice is the opposite. Because the teenagers' negligent behavior would be said to have "contributed" to the collision, the firm would not be held liable for their injuries. Thus, the decision would again be reduced to arithmetic. A compensatory payout to one victim—i.e., the passenger—is more expensive than no payout at all. A profit-maximizing firm's vehicle should not swerve.

[70] Kaplow & Shavell, *supra* note 67, at 1669.

JACKPOT HYPOTHETICAL: CONTRIBUTORY NEGLIGENCE

Note that the two hypotheticals are morally indistinguishable. But legally, they are altogether distinct. It is this very distinction—what the canonical jurist Oliver Wendell Holmes referred to as the "distinction between morality and law"—that many leading lights within the field of machine ethics fail to take into account.[71]

In an 1897 essay, today regarded as among the "most important . . . ever written by an American on the law,"[72] Holmes set out to dispel what he regarded as a commonplace "confusion" between morality and legality.[73] Invoking the analogy of a "bad man" devoid of moral scruples, Holmes described the legal system's distinct ability to give even a bad actor "as much reason as a good one" to behave morally.[74] The key, according to Holmes, was to pull not on the actors' heartstrings but on their purse strings. For even "[a] man who cares nothing for . . . ethical rule[s]," he observed, "is likely nevertheless to care a good deal to avoid being made to pay money, and will want to keep out of jail if he can."[75]

To this end, Holmes viewed the legal system as an "instrument of . . . prophesy"—enabling society to "predict" the behavior of rational, self-interested actors by "look[ing] at [the law] as a bad man" would.[76] And

[71] *The Path of the Law*, *supra* note 1, at 459.

[72] *See* Albert Alschuler, *The Descending Trail: Holmes' Path of the Law One Hundred Years Later*, 49 FLA. L. REV. 353, 354 (1997).

[73] *The Path of the Law*, *supra* note 1, at 459.

[74] *Id.*

[75] *Id.*

[76] *Id.* at 458–59.

though Holmes couched his analogy in explicitly ethical terms, he could as easily have used the phrase "blank man"—or the more familiar "blank slate"—to communicate the same point. With the right legal incentives, Holmes stressed, even amoral actors could be made to behave indistinguishably from moral ones.[77] But for Holmes, the tale was ultimately cautionary. Take away those same incentives, and amoral actors will find no countervailing reasons to behave morally within "the vaguer sanctions of [their] conscience."[78]

For robotics firms duty-bound to maximize profits—not morality—Holmes's analogy is no less apposite. Indeed, it illustrates precisely why the diagnosis made by many luminaries in the field of machine ethics is dead-on, but their prognosis is off.[79] Robotics systems of the future will undoubtedly make decisions of immense ethical import. But their decisionmaking will be guided less by the vagaries of "conscience" than by the "prophesy" of profit.[80] These robots will view the world not as good moral philosophers, but as bad men—concerned less with idealized "ethical rule[s]" than with the legal rules that dictate whether their firms will face public sanction.[81] And those that are instead engineered to follow "a clear and consistent" moral code will behave irrationally under a legal code lacking both such qualities.[82]

Nearly a decade ago, Wallach and Allen warned in *Moral Machines* that "to avoid the consequences of bad autonomous artificial agents, people must be prepared to think hard about what it will take to make such agents good."[83] But vital as this admonition may be, mere hard thinking is not enough. To think profitably about this "engineering challenge"[84] requires first understanding it as a challenge fundamentally about profit. And if the goal of machine ethics is "to frame discussion in a way that constructively guides" this effort, machine ethics must begin by acknowledging that firms engineering so-called moral machines will be concerned foremost with economics, not ethics.[85]

---

[77] *Id.* at 459.

[78] *Id.*

[79] *See supra* notes 57–64 and accompanying text.

[80] *The Path of the Law*, *supra* note 1, at 458–61.

[81] *Id.*

[82] *See* Greene, *supra* note 18, at 1515.

[83] MORAL MACHINES, *supra* note 10, at 7.

[84] *Id.* at 8.

[85] *Id.* at 6.

III. "BAD MAN" ALGORITHMS AND THE PATH OF ROBOTICS LAW

> *It's not possible to make a moral judgement of the worth of one individual person vers[us] another—convict versus nun . . . . When we think about the [trolley] problem, we try to cast it in a frame that we can actually do something with.*
>
> —Chris Urmson[86]

Call Holmes's outlook dim—even "dismal"[87]—but the evidence of its validity abounds. Happen to activate the "autopilot" function of a Tesla while driving down the highway?[88] It should come as no surprise that the system is designed to closely correlate the car's speed with the posted legal limit.[89] At this seemingly trivial engineering decision, few so much as bat an eye. But make no mistake, the moral calculus underlying it is as urgent as in the trolley problem.

In theory, Tesla's engineers could have opted for a different speed than that set forth by the law. And each increment or decrement so chosen would, in turn, translate to a corresponding variation in risk to the vehicle's passenger, as well as to third parties. Multiply even the slightest change in risk by the millions of individuals foreseeably impacted by it, and one is left with an ethical dilemma eerily reminiscent of an edge case like the trolley problem. Except in this instance, actual lives are on the line.

Yet, far from expecting Tesla to commission a grand philosophical treatise on the precise "moral" speed for each U.S. roadway, most simply take it for granted that Tesla's engineers looked to law and designed accordingly. In fact, the failure to see this decision for what it is—as an act of profit prevailing over ethics—seems to have less to do with the decision's undeniable moral implications than with the clearly defined standards set forth by the law.

Further, in situations where the law's mandate is less clear, our ethical intuitions may change, but the economic incentives driving the engineers do not. Confronted with a moral dilemma wherein the legal system offers

---

[86] Matt McFarland, *Google's Chief of Self-Driving Cars Downplays 'The Trolley Problem'*, WASH. POST (Dec. 1, 2015), https://www.washingtonpost.com/news/innovations/wp/2015/12/01/googles-leader-on-self-driving-cars-downplays-the-trolley-problem/?tid=a_inl [https://perma.cc/Q42K-FJJJ] (quoting Chris Urmson, Google's Chief Autonomous Vehicles Engineer). Urmson replaced former Chief Autonomous Vehicles Engineer, John Markoff, in August 2016. *Latest to Quit Google's Self-Driving Car Unit: Top Roboticist*, N.Y. TIMES (Aug. 5, 2016), http://www.nytimes.com/2016/08/06/technology/alphabet-google-autonomous-car-chris-urmson.html?r=0 [https://perma.cc/Y9RK-WDCS].

[87] S*ee* DIXON, *supra* note 24, at 1.

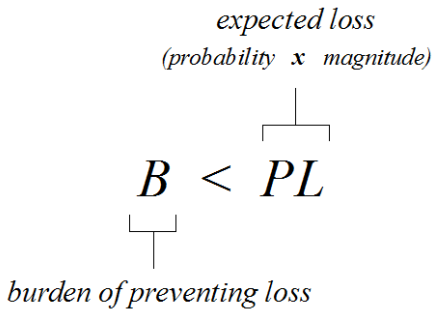
[88] Jordan Golson, *Tesla's New Update Restricts Autopilot to the Speed Limit on Undivided Roads*, VERGE (Dec. 22, 2016 1:09 PM), http://www.theverge.com/2016/12/22/14057634/tesla-autopilot-speed-limit-restriction-update [https://perma.cc/HB2F-7KU9].

[89] *Id.*

less clear-cut guidance, today's most sophisticated robots reputedly shift to a more probabilistic approach.[90] According to the robotics scholar Noah Goodall:

> Each potential outcome is assigned a likelihood as well as a positive or negative magnitude (either a benefit or a cost). Each event's magnitude is multiplied by its likelihood, and the resulting values [are] summed. If the benefits outweigh the costs by a reasonable margin, the [robot] . . . execute[s] the action . . . ."[91]

Should this description sound familiar to readers with a legal background, it is for good reason. The calculus is strikingly similar to a theory of negligence first formulated by yet another American jurist of canonical stature, Learned Hand, who algebraically expressed the legal "duty of care" as "a function of three variables:"[92] (1) the probability, *P*, that an accident will occur; (2) the magnitude of the resulting loss, *L*; and (3) the burden, *B*, of taking adequate care to prevent the accident.[93]



But while the cost–benefit calculations of Hand's formula are meant to reflect a decision's actual costs to society—separate from any costs imposed by the legal system—calculations performed by robots will be just the opposite. Systems optimized for profit will not fret over negative externalities, but only those costs the firm can expect to incur.[94]

---

[90] Noah Goodall, *Can You Program Ethics into a Self-Driving Car?*, IEEE SPECTRUM (May 31, 2016), http://spectrum.ieee.org/transportation/self-driving/can-you-program-ethics-into-a-selfdriving-car [https://perma.cc/7UZW-QYMG] [hereinafter *Can You Program Ethics?*]; *see also* Noah Goodall, *Away from the Trolley Problem and Toward Risk Management*, 30 APPLIED ARTIFICIAL INTELLIGENCE 810, 815 (2016) (discussing this calculus is greater detail).

[91] *Can You Program Ethics?*, *supra* note 90.

[92] *See* United States v. Carroll Towing Co., 159 F.2d 169, 171–73 (2d Cir. 1947).

[93] *Id.* If the inequality holds, the injurer is negligent.

[94] *Cf.* Kaplow & Shavell, *supra* note 67.

Accordingly, the negative event magnitude associated with each executable action will largely be a function of the action's foreseeable legal sanctions.[95]

Or, to borrow a more poetic phrase, these systems will operate as "instrument[s] of . . . prophesy"—looking "at the [law] as a bad man" to determine the economically optimal course of action.[96] Their cost–benefit calculations will not maximize morality, but minimize liability. And though the complexity of these systems remains constrained by the limits of present-day technology, as the feasibility of designing legally sophisticated "bad man" algorithms increases with time, so too will the economic incentives to implement them.

Meanwhile, if the "21st-century moral philosophers"[97]—or anyone else for that matter—find the decisions executed by these systems unethical, the solution will not entail sermonizing outside the engineering headquarters of Google, Tesla, or Mercedes-Benz. Rather, it will require realigning existing legal incentives through the piecemeal work of democratic change.

But though the means by which society ultimately shapes robot behavior may be less grandiose than some currently imagine, rest assured that the end will be no less grand. We, the people, will be the true engineers of machine morality. As democratic stakeholders, it will be our collective "engineering task" to ensure that even the worst of our robots are incentivized to behave as the best of our philosophers. Ironically, success on this front will require overturning Holmes's century-old insight. For it will entail designing a legal system that truly closes the gap "between morality and law"—thereby rendering meaningless "the practical importance of the distinction between" a moral machine and amoral machines.[98]

CONCLUSION

Although the notion that a nineteenth-century jurist has much to teach about twenty-first century robotics may sound far-fetched, the burgeoning fields of robotics law and machine ethics would do well to take stock of Holmes's timeless insight. While his "distinction between morality and law"[99] may seem a hair-splitting academic maundering, its implications for

---

[95] *But see supra* note 69 (clarifying the additional costs that may factor into this calculus under certain conditions).

[96] *The Path of the Law*, *supra* note 1, at 458–61.

[97] *See* Greene, *supra* note 18.

[98] *The Path of the Law*, *supra* note 1, at 459.

[99] *Id.*

engineering are anything but. Great challenges undoubtedly lie ahead for societies poised to embed millions of robots with high-stakes decisionmaking capabilities. But rising to meet these challenges requires first understanding them. For twenty-first century democracies—composed of engineers, lawyers, and moral philosophers alike—"this may be where the rubber meets the road."[100]

[100] *See* Greene, *supra* note 18.