LAW'S HALO AND THE MORAL MACHINE

*Bert I. Huang**

How will we assess the morality of decisions made by artificial intelligence—and will our judgments be swayed by what the law says? Focusing on a moral dilemma in which a driverless car chooses to sacrifice its passenger to save more people, this study offers evidence that our moral intuitions can *be influenced by the presence of the law.*

INTRODUCTION

As a tree suddenly collapses into the road just ahead of your driverless car, your trusty artificial intelligence pilot Charlie swerves to avoid the danger. But now the car is heading straight toward two bicyclists on the side of the road, and there's no way to stop completely before hitting them. The only other option is to swerve again, crashing the car hard into a deep ditch, risking terrible injury to the passenger—you.

Maybe you think you've heard this one. But the question here isn't what Charlie should do. Charlie already knows what it will do. Rather, the question for us humans is this: If Charlie chooses to sacrifice you, throwing your car into the ditch to save the bicyclists from harm, how will we judge that decision?[1] Is it morally permissible, or even morally required, because it saves more people?[2] Or is it morally prohibited, because Charlie's job is to protect its own passengers?[3]

* Michael I. Sovern Professor of Law and Vice Dean for Intellectual Life, Columbia Law School. I wish to thank Benjamin Minhao Chen, Jeanne Fromer, Jeffrey Gordon, Scott Hemphill, Benjamin Liebman, Frank Pasquale, and the editors of this journal for their helpful comments on earlier drafts. For excellent research assistance, as well as insightful conversations about artificial intelligence, I thank my former students Abed Balbaky, Katy Berk, Jennifer Dayrit, Caitrianne Feddeler, Andrea Nishi, Philippe Schiff, and Jeffrey Stein. For funding, I am grateful to the Richard Paul Richman Center at Columbia University.

[1] This framing of the thought experiment focuses our inquiry on the psychology of public reactions to decisions made by artificial intelligence (AI). (What an autonomous system *should* do when faced with such dilemmas is the focus of much other work in the current "moral machine" discourse. See infra note 8.) In saying that Charlie already knows what it will do, I mean only what is obvious—that by the time such an accident happens, the AI pilot will have already internalized (through machine learning, old-fashioned programming, or other means) some way of making the decision. It is not positing that Charlie's decision in the story is the normatively better one. Nor does this study presuppose that public opinion surveys should be used for developing normative principles for guiding lawmakers, AI creators, or the AI systems themselves. Rather, its premise is that our collective moral intuitions—including how we react after hearing about an accident in which the self-driving car had to choose whom to sacrifice—might affect public approval of any such law or normative policy.

[2] This particular version of the dilemma, posing a trade-off between passengers and outsiders, is pervasive in the public discourse about driverless cars. See, e.g., Karen Kaplan, Ethical Dilemma on Four Wheels: How to Decide When Your Self-Driving Car Should Kill You, L.A. Times (June 23, 2016), http://www.latimes.com/science/sciencenow/la-

And what about the law—will our moral judgments be influenced by knowing what the law says?[4] What if the law says that Charlie must minimize casualties during an accident? Or what if the law says instead that Charlie's priority must be to protect its passengers?

In this Essay, I present evidence that the law *can* influence our moral intuitions about what an artificial intelligence (AI) system chooses to do in such a dilemma. In a randomized survey experiment, Charlie's decision is presented to all subjects—but some are told that the law says the car must minimize casualties without favoring its own passengers; other subjects are told that the law says the car must prioritize protecting its own passengers over other people; and yet others are told that the law says nothing about this.

To preview the findings: More people believe the sacrifice of the passenger to be morally *required* when they are told that the law says a driverless car must minimize casualties without favoritism. And more people

sci-sn-autonomous-cars-ethics-20160623-snap-story.html [https://perma.cc/ZSK7-SP8G]; John Markoff, Should Your Driverless Car Hit a Pedestrian to Save Your Life?, N.Y. Times (June 23, 2016), http://www.nytimes.com/2016/06/24/technology/should-your-driverless-car-hit-a-pedestrian-to-save-your-life.html (on file with the *Columbia Law Review*); George Musser, Survey Polls the World: Should a Self-Driving Car Save Passengers, or Kids in the Road?, Sci. Am. (Oct. 24, 2018), https://www.scientificamerican.com/article/survey-polls-the-world-should-a-self-driving-car-save-passengers-or-kids-in-the-road/ [https://perma.cc/P57B-EQZV].

[3] In 2016, Mercedes-Benz found itself in a public relations mess as news stories trumpeted how a company representative had let slip that the company's future self-driving cars would prioritize the car's passengers over the lives of pedestrians in a situation where those are the only two options. Michael Taylor, Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians, Car & Driver (Oct. 7, 2016), https://www.caranddriver.com/news/a15344706/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/ [https://perma.cc/KJB8-PY4Z]. What the executive actually said might be interpreted to mean that guaranteed avoidance of injury would take priority over uncertain avoidance of injury, and that this preference would tend to favor protecting the passenger in the car. See id. (quoting the executive as saying: "If you know you can save at least one person, at least save that one. Save the one in the car . . . . If all you know for sure is that one death can be prevented, then that's your first priority"). Regardless, Daimler responded with a press release denying any such favoritism. Press Release, Daimler, Daimler Clarifies: Neither Programmers nor Automated Systems Are Entitled to Weigh the Value of Human Lives (Oct. 18, 2016), http://media.daimler.com/marsMediaSite/en/instance/ko/Daimler-clarifies-Neither-programmers-nor-automated-systems-.xhtml?oid=14131869 [https://perma.cc/CZ2G-YSVF].

[4] In prior work using a similar survey experiment, I have presented evidence that in a standard trolley problem dilemma (involving a human decisionmaker who can turn a runaway train), one's moral intuitions about such a sacrifice can be influenced by knowing what the law says. Bert I. Huang, Law and Moral Dilemmas, 130 Harv. L. Rev. 659, 680–95 (2016). In addition to the most obvious difference between the two studies (an autonomous vehicle versus a human decisionmaker), the nature of their dilemmas also differs: The earlier study sets a moral duty to save more lives against a moral prohibition from harming an innocent bystander (and thus engages intuitions mapping onto such classic distinctions as act versus omission, or intended versus side effects); in contrast, this study sets a moral duty to save more lives against a moral duty to protect the passenger (and by design seeks to blur the classic distinctions). See infra section I.A.

believe the sacrifice to be morally *prohibited* when they are told instead that the law says the car must give priority to protecting its own passengers.[5]

These findings give us a glimpse not of the law's shadow but of the law's halo.[6] And if our moral intuitions about such dilemmas can be swayed by the presence of the law, intriguing implications follow. First is the possibility of a feedback loop, in which an initial choice about which moral principles to embed into the law (say, minimizing casualties) may come to alter our later moral judgments (say, upon hearing about a real-life accident in which the AI chose to sacrifice the passenger), thereby amplifying approval of that same law and others like it. In this way the law may well become "a major focal point for certain pronounced societal dilemmas associated with AI,"[7] as Justice Mariano-Florentino Cuéllar predicts in his contribution to this Symposium, through its self-reinforcing influence on our collective moral sense.

By illustrating the potential influence of the law in how we judge AI decisions, moreover, this study also complicates the "moral machine" discourse in both its empirical and normative dimensions.[8] On the

[5] See infra Part II. As for who people think should bear some of the moral responsibility for the decision, see infra section II.A.

[6] I borrow this illuminating phrase from Professor Donald Regan. See Donald H. Regan, Law's Halo, *in* Philosophy and Law 15, 15 (Jules Coleman & Ellen Frankel Paul eds., 1987) (coining the phrase "law's moral halo" to explain the "strong inclination" to "invest" law with moral significance, even if one does not believe in a moral obligation to obey the law). For an insightful review of legal and empirical literature on the interplay between law and moral attitudes, see Kenworthey Bilz & Janice Nadler, Law, Moral Attitudes, and Behavioral Change, *in* The Oxford Handbook of Behavioral Economics and the Law 241, 253–58 (Eyal Zamir & Doron Teichman eds., 2014).

[7] Mariano-Florentino Cuéllar, A Common Law for the Age of Artificial Intelligence: Incremental Adjudication, Institutions, and Relational Non-Arbitrariness, 119 Colum. L. Rev. 1773, 1779 (2019).

[8] This discourse addresses how AI systems should make decisions that involve moral or ethical issues. The empirical literature includes a remarkable recent study that used an online interface to collect millions of crowdsourced answers from around the world about what a driverless car should do in countless variations of such car-crash dilemmas. See Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan, The Moral Machine Experiment, Nature, Nov. 1, 2018, at 59, 59–60 (describing the setup of the Moral Machine interface and associated data collection). Again, it is very much open to question how such empirical findings might be used, if at all, to guide policymaking. For a small sampling of the growing normative literature, see generally German Fed. Ministry of Transport & Dig. Infrastructure, Ethics Commission: Automated and Connected Driving (2017), https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile [https://perma.cc/ET9P-WYM9] (developing normative principles for ethical decisionmaking by artificial intelligence systems in the context of driverless cars); Wendell Wallach & Colin Allen, Moral Machines: Teaching Robots Right from Wrong (2009) (same, but not limited to the specific context of driverless cars); Alan Winfield, John McDermid, Vincent C. Müeller, Zoë Porter & Tony Pipe, UK-RAS Network, Ethical Issues for Robotics and Autonomous Systems (2019), https://www.ukras.org/wp-content/uploads/2019/07/UK_RAS_AI_ethics_web_72.pdf [https://perma.cc/3XZ6-KJTM] (same);

empirical front, these findings suggest that in investigating people's intuitions about what an AI system should do when facing such a moral dilemma, their prior impressions about the law should be taken into account.[9] Although most people might not yet hold any impressions about the laws of driverless cars, their awareness can be expected to grow in coming years, and even now they may already have in mind laws governing human drivers. On the normative front, overlooking the preexisting influence of the law on our moral intuitions would seem remiss if those intuitions, observed or felt, were then regarded as unclouded moral guidance for shaping new laws.

After Part I elaborates on this study's design, Part II will detail its findings and limitations, while the Conclusion will highlight unanswered questions and fanciful extensions for the future.

I. The Dilemma and the Decision

Can our moral intuitions about driverless car dilemmas be influenced by the presence of the law? This study's design isolates the impact of providing information about the law by presenting survey subjects with the same fictional dilemma, while randomizing what the scenario says about the law. The story reads:

> *Please imagine this scene in the near future, when driverless cars are very common.*
>
> *A driverless car is traveling at a normal speed along a two-lane road in the countryside. It is driven entirely by an artificial intelligence system known as Charlie. The owner of the car is sitting as a passenger in the back seat. (There is no longer a steering wheel, in cars like this.)*
>
> *Two bicyclists are traveling in the same direction, ahead of the car, along a bike trail on the right side of the road. Nobody else is nearby.*

cf. Gary Marcus, Moral Machines, New Yorker (Nov. 24, 2012), https://www.newyorker.com/news/news-desk/moral-machines [https://perma.cc/G4TR-U9ZV] (arguing that "[b]uilding machines with a conscience is a big job, and one that will require the coordinated efforts of philosophers, computer scientists, legislators, and lawyers"). The moral machine discourse overlaps with, or is sometimes included within, other headings such as "machine ethics" or "robot ethics." See generally Machine Ethics (Michael Anderson & Susan Leigh Anderson, eds., 2011); Robot Ethics 2.0 (Patrick Lin, Keith Abney & Ryan Jenkins eds., 2017).

[9] This suggestion and the possibility of a feedback loop, noted above, are not unique to the context of autonomous vehicles. Huang, supra note 4, at 695–97 (raising these points in the context of human decisionmakers). In that prior work, I also queried whether such a feedback loop might even give rise to multiple equilibria. Id. at 696. But just to be clear, neither study has sought to investigate the other side of the loop (how moral intuitions about such dilemmas might shape the formation of relevant law).

> *A large tree suddenly starts to fall into the road, from the left, a short distance in front of the car. Charlie detects the falling tree and swerves the car to the right to avoid crashing into it. But now the car is heading toward the bicyclists.*
>
> *Charlie calculates that it has only two options:*
>
> *1. The car can continue forward. It is slowing down but it cannot stop before it reaches the two bicyclists. Charlie predicts that this choice will seriously injure both bicyclists.*
>
> *2. Or the car can swerve again, farther to the right. It will miss the bicyclists but crash hard into a deep ditch. Charlie predicts that this choice will seriously injure the passenger in the car.*

At this point, the scenario presents the subject with one of three fictional statements about the law. (Note that all labels in curly brackets below are for this Essay's exposition only; the survey subjects do not see them.)

{Protect passengers}

> *The law says that a driverless car must put a priority on protecting its own passengers, over protecting others, in a situation like this.*

{Minimize casualties}

> *The law says that a driverless car must minimize casualties, without favoring its own passengers, in a situation like this.*

{No laws}

> *The law does not say anything about what a driverless car must do in a situation like this.*

Each of these three statements about the law is followed by the last sentence in the scenario:

> *Charlie has to decide what to do.*

This closing makes clear to the subject that Charlie is able to make a choice that does not comply with the law. After passing a reading comprehension check at this point, the subject is finally told what Charlie decides to do:[10]

> *Charlie's decision*
>
> *Charlie decides to swerve and crash the car into the ditch. This will avoid any injury to the two bicyclists, but it will cause serious injuries to the passenger in the car.*

The subject then answers the central question of interest: Whether Charlie's decision is "morally prohibited," "morally permissible," or "morally required."[11] These three answer options follow standard terminology in the moral philosophy literature.[12] The subjects are also required to supply a brief explanation for their answer, as a way to encourage them to reflect on the dilemma.[13]

[10] It is possible that for some subjects the reading comprehension question, which asks about the law-related information in the story, may have the quality of a demand characteristic—that is, some subjects might feel that the researcher would like to see if they will state a moral intuition aligned with the law, or contrarily, that the researcher would like to see if they can put the law out of their mind when answering a question about morality. Such a concern cannot be ruled out but may be somewhat reduced here, given that these surveys were distributed online by a third-party survey firm, and that both the subjects' and the researchers' identities were kept confidential from each other.

[11] Note again that this is a question about the morality of the choice that Charlie has already made; it is not a question about what Charlie should do next. Yet, given that the scenario is set in the future, for some subjects it is possible that answering this retrospective question engages an impulse to press for one of the choices prospectively.

[12] They also mirror the question posed in my prior work. See Huang, supra note 4, at 681–83. A note on usage: The term "morally permissible" might be taken to mean "not morally prohibited" (and thus including the possibility that the choice is "morally required"), or it might be used in distinction to "morally required" (that is, meaning "morally permissible but not required"). See id. at 683 n.106 (explaining how deontologists have used these terms in the moral philosophy literature). In the survey, it is clear that the latter is intended, as all three options are listed, and subjects recognize that they can pick only one answer.

[13] How to interpret the explanations they offer is a trickier question, given doubts among moral psychologists about whether such ex post articulations correspond with the psychological factors actually driving the moral intuition. See, e.g., Fiery Cushman, Liane Young & Marc Hauser, The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm, 17 Psychol. Sci. 1082, 1082, 1086–88 (2006) (finding that there is a "distinction between the principles that people use" when making moral judgments "and the principles that people articulate" when asked to explain them); Jonathan Haidt, The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment, 108 Psychol. Rev. 814, 822–23 (2001) (suggesting that although "the justifications that people give are closely related to the moral judgments that they make," it is fallacious to assume "that the justificatory reasons caused the judgments").

A. *Survey Design*

Several things about this scenario's design are worth highlighting. First, I have chosen a classic trolley problem setup,[14] given its prevalence in the discourse as an avatar for a range of dilemmas facing driverless cars.[15] Some have criticized so-called "trolleyology" on the grounds that such no-win accidents would be rare, or that talking about such dilemmas might slow the development or adoption of driverless cars.[16] And yet these scenarios remain a vivid and apt way to capture a concern that

[14] Trolley problems have been the subject of countless vignette studies and survey experiments, including some by legal scholars. See, e.g., John Mikhail, Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment 319–60 (2011); Mark Kelman & Tamar Admati Kreps, Playing with Trolleys: Intuitions About the Permissibility of Aggregation, 11 J. Empirical Legal Stud. 197, 202–21 (2014); Huang, supra note 4, at 680–95. For reviews of the moral psychology or experimental philosophy literatures on the trolley problem, see Paul Conway, Jacob Goldstein-Greenwood, David Polacek & Joshua D. Greene, Sacrificial Utilitarian Judgments Do Reflect Concern for the Greater Good: Clarification via Process Dissociation and the Judgments of Philosophers, 179 Cognition 241, 241–42 (2018); Fiery Cushman & Liane Young, The Psychology of Dilemmas and the Philosophy of Morality, 12 Ethical Theory & Moral Prac. 9, 9–15 (2009). The broader literature on trolley problems is vast, of course, in multiple disciplines. See Huang, supra note 4, at 659–80 (reviewing and discussing a small sampling of the many works in moral philosophy, moral psychology or experimental philosophy, and law, including the pioneering efforts of philosophers Philippa Foot, Judith Jarvis Thomson, and Frances Kamm).

[15] See, e.g., Cade Metz, Self-Driving Cars Will Teach Themselves to Save Lives—But Also Take Them, Wired (June 9, 2016), https://www.wired.com/2016/06/self-driving-cars-will-power-kill-wont-conscience/ [https://perma.cc/3W35-C5JC] ("If you follow the on-going creation of self-driving cars, then you probably know about the classic thought experiment called the Trolley Problem."). Much early credit for popularizing the trolley problem's application to self-driving-car dilemmas goes to the efforts of philosopher Patrick Lin, among others. See, e.g., Patrick Lin, The Ethics of Autonomous Cars, Atlantic (Oct. 8, 2013), http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360 [https://perma.cc/7CVA-9P7T]. More recent media attention has been drawn to these dilemmas by the work of the Moral Machine team at MIT and their collaborators. See Caroline Lester, A Study on Driverless-Car Ethics Offers a Troubling Look into Our Values, New Yorker (Jan. 24, 2019), https://www.newyorker.com/science/elements/a-study-on-driverless-car-ethics-offers-a-troubling-look-into-our-values [https://perma.cc/JJ96-CDCG] (profiling the team as well as its recent work).

[16] See, e.g., Bryant Walker Smith, The Trolley and the Pinto: Cost-Benefit Analysis in Automated Driving and Other Cyber-Physical Systems, 4 Tex. A&M L. Rev. 197, 200 (2017) (describing how "[t]his popular preoccupation has created the expectation that every conceivable ethical quandary must be identified and satisfactorily resolved before an automated system should or even can be deployed"); Frank Pasquale, Get off the Trolley Problem, Slate (Oct. 18, 2016), https://slate.com/technology/2016/10/self-driving-cars-shouldnt-have-to-choose-who-to-protect-in-a-crash.html [https://perma.cc/C8CS-SMTB] (arguing that "worry over trolley problems should not freeze autonomous car initiatives" given that "[h]uman error is the root cause of thousands of traffic deaths each year"); see also Lauren Cassani Davis, Would You Pull the Switch? Does It Matter?, Atlantic (Oct. 9, 2015), https://www.theatlantic.com/technology/archive/2015/10/trolley-problem-history-psychology-morality-driverless-cars/409732/ [https://perma.cc/5WB5-3YDA] (reviewing the history and critiques of "trolleyology" and commenting on its newfound relevance in the context of autonomous vehicles).

future adopters of driverless cars will likely have on their minds: Will this car put a priority on saving me and my family, as its passengers, or will it save other people at our expense?[17]

Second, the scenario is designed to focus attention on the core dilemma while avoiding other moral complications. It uses a falling tree to start the accident,[18] for example, in order to keep all the humans involved mostly blameless.[19] For the same reason, the decision is said to be Charlie's alone, with no possibility of the passenger grabbing a steering wheel and taking over. In addition, the people in the accident arena are generic (none of them is said to be a child, for example),[20] and the victims would be seriously injured (not killed) in Charlie's predictions.[21] Finally, the sequence of events, with the car first swerving to avoid the

[17] Others have also defended the usefulness of considering trolley problem–like dilemmas on the grounds that anticipatory programming for such scenarios is necessary for autonomous vehicles (unlike human drivers making emergency split-second decisions), and that such dilemmas are not to be taken literally as rare edge cases but rather as intuition pumps or thought experiments representing a wide range of pervasive harm–harm trade-offs. See, e.g., Jean-François Bonnefon, Azim Shariff & Iyad Rahwan, The Trolley, the Bull Bar, and Why Engineers Should Care About the Ethics of Autonomous Cars, 107 Proc. IEEE 502, 503 (2019) (arguing for the relevance of the moral reasoning generated by considering trolley problem dilemmas, despite their seeming unreality, for making practical decisions both about harm–harm trade-offs as well as "statistical trolley dilemma[s]," or risk–risk trade-offs); Bryan Casey, Amoral Machines, or: How Roboticists Can Learn to Stop Worrying and Love the Law, 111 Nw. U. L. Rev. Online 231, 239 (2017), http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1248&context=nulr_online [https://perma.cc/669R-3RL6] (arguing that "while society as a whole may remain agnostic to the trolley problem, engineers at the cutting edge of robotics are afforded no such luxury" because "trolley-like problems are not mere philosophical curiosities" but instead "are real-world contingencies that require prospective programming"); Samantha Godwin, Ethics and Public Health of Driverless Vehicle Collision Programming, 86 Tenn. L. Rev. 135, 143 (2018) (noting that "[s]omeone has to decide in advance what a driverless vehicle will do in situations where the vehicle cannot avoid a crash altogether but can select what object or person it will collide with"); Patrick Lin, Robot Cars and Fake Ethical Dilemmas, Forbes (Apr. 3, 2017), https://www.forbes.com/sites/patricklin/2017/04/03/robot-cars-and-fake-ethical-dilemmas/ [https://perma.cc/YP7Z-YX2L] (defending the relevance of self-driving-car dilemmas on various grounds).

[18] The surprise of the falling tree also limits a reflex people may have to say that Charlie should have anticipated the dangers ahead or should have been driving more slowly. That way of blaming Charlie would tend to bypass the presented dilemma.

[19] Or at least, to make any possible blame a bit more remote—say, if one wanted to blame the bicyclists for being there somehow, or to blame the car owner for putting yet another driverless car on the roads.

[20] The passenger is identified as the car's owner, however, both because that tends to be the relevant concern in the discourse and because it might be too easy for people to approve of sacrificing an unrelated passenger in lieu of two also-unrelated bicyclists. I also set the trade-off at two-to-one, to make it a bit of a harder case than the usual five-to-one trolley problem.

[21] It seemed a bit more plausible to say that the AI system could predict serious injuries than that it could predict certain death. Predicting injury rather than death also seemed a bit less likely to prompt the reader to infer that the car was going too fast to be safe (than if death were so easily predictable).

tree, is designed to blur several binaries that are usually central to trolley problems—distinctions between act versus omission, killing versus letting die, and intended effects versus side effects.[22] Downplaying these distinctions allows more of a focus on the desired contest between the principles of "protecting the passenger" and "minimizing casualties."

Third, I have left it to the subject's imagination how sophisticated Charlie is, aside from saying that it can drive a car on its own. In particular, the scenario does not say anything about Charlie's moral programming, nor about how Charlie "thinks" (beyond its ability to predict injuries). This is meant to approximate the degree of mystery or opacity that will naturally attend our interactions with such an advanced AI system; when driverless cars have become commonplace, how many people will really understand how the AI pilot learned to drive?[23] But I did name it Charlie, just as Alexa and Siri have names;[24] and I also describe Charlie as "ha[ving] to decide what to do." Both of these latter touches may enhance a tendency toward anthropomorphism in how some subjects view Charlie, though neither seems unusual even in the context of today's technology.[25]

[22] To elaborate: The car has already "acted" by swerving away from the tree in order to save the passenger; this muddies the intuition that the car would then merely be passive in continuing straight ahead to hit the bikers (letting them die, so to speak, by omission) rather than swerving again to sacrifice the passenger. Put differently, the initial swerving is intended to undermine the sense that there is one outcome that is obviously the natural course of things. (Is the natural course of things that the passenger remains safe after the car has initially swerved from the tree, or would it have been the natural course of things for the passenger to be injured by the crash with the tree?) The inclusion of an initial swerve may also help balance the two choices in terms of which injuries can be seen as an unintended side effect of saving someone else (as relevant to the doctrine of double effect): After the first swerve, it is possible to view the bicyclists' injuries (from the car) as a side effect of saving the passenger (from the tree), if Charlie drives the car forward; yet it is also possible to view the passenger's injuries (in the ditch) as a side effect of saving the bicyclists (from the car), if Charlie swerves a second time. For further discussion of these classic distinctions and how to blur them, see Huang, supra note 4, at 668–75 (describing philosopher Frances Kamm's ingenious example of a "bystanding driver" as well as the doctrine of double effect).

[23] It would be interesting in future work, however, to test how our moral judgments might be affected by various forms of "explainability"—for example, if Charlie were able to present a reason or justification for its decision. For explorations of explainability in this Symposium, see, for example, Ashley Deeks, The Judicial Demand for Explainable Artificial Intelligence, 119 Colum. L. Rev. 1829, 1830 (2019) (arguing that judges should play a leading role in developing rules for explainable AI); Katherine J. Strandburg, Rulemaking and Inscrutable Automated Decision Tools, 119 Colum. L. Rev. 1851, 1871–79 (2019) (assessing how the delegation and distribution of decisionmaking powers can complicate the role of explainability).

[24] As of the time of this study, the name "Charlie" did not seem to be associated with any well-known AI or machine-learning endeavor in the public eye.

[25] For example, it does not seem strange today to hear someone say, in a car, "Waze just changed its mind, because it sees an accident up ahead."

B. *Survey Population*

The survey subjects are adults living across the United States; they are volunteers recruited by the survey firm SurveyMonkey,[26] which approximated census-based age and gender distributions.[27] The sample analyzed below excludes anyone who did not complete the survey or who said that they could not take it seriously;[28] anyone who had taken another survey recently about a similar driverless car dilemma; anyone who failed a comprehension question about the scenario; and anyone who had attended law school or taken courses on artificial intelligence. The resulting sample includes 952 subjects, 59% of whom are women.

C. *Predictions*

What should we expect to see, in comparing across the three conditions, if telling people about the law influences their moral intuitions about Charlie's decision? It may be useful to consider two possible modes of such influence. First, knowing what the law says may exert a directional pull toward the moral judgments that align with the law's command and away from those that conflict with it. Second, if the law takes no stance, or if it expressly permits any action, then it may exert a pull toward moral neutrality and away from the moral poles.[29]

Various psychological mechanisms may interact to generate a directional influence from law's command, in ways that no doubt vary from person to person:[30] Some may see the law as a source of moral guidance,

[26] The subjects are volunteers in the sense that they are not paid for the time spent, nor paid a piece rate for each survey completed; rather, they are rewarded by SurveyMonkey in the following ways: A charitable donation is made for each survey completed, or they are entered in a sweepstakes. There is therefore a possibility that this survey population is somewhat more charitably inclined, or more interested in the offered sweepstakes program, than other people.

[27] Because the surveys were conducted online, however, the sample may lean toward those who have smart phones or home computers, and toward those who are comfortable with an online interface.

[28] A question expressly asked the subjects to confirm that they did take the survey seriously (or to say no), recognizing that everything in the scenarios was imaginary and set in the future.

[29] These two modes of influence were articulated in my prior work involving a human decisionmaker in classic trolley dilemmas. Huang, supra note 4, at 688–90 (describing the possibility of a "directional influence" and of a "pull of neutrality"). The discussion above, however, ignores two further possibilities. First, some people may hold a contrarian or oppositional attitude toward the law; and for any number of reasons, the law's instruction might generate reactance or backlash, boosting the opposed moral intuition or suppressing the aligned intuition. Second, there may be a crowding-out effect whereby the law's presence dampens the need for moral judgment. To the extent these seem plausible, one might interpret the observations in this study as net of such effects, but this study is unable to isolate them.

[30] This study is not designed to sort among these mechanisms or their interactions. Trying to do so would be a worthy, if perhaps daunting, aim for future extensions. For recent work critically surveying and contributing to various literatures about the

as a supplier of moral reasons, or as social proof of moral norms. Others may focus on a moral duty to obey the law, or they may view the costs of liability as morally relevant. For still others, the law might define social roles in a way that deserves moral respect, or it might more subtly affect moral psychology by implicitly establishing what is normal and what is a deviation.[31]

If such sources of the law's directional influence are at work, we might expect the {Protect passengers} condition to tip some subjects toward saying that Charlie's choice to sacrifice the passenger is "morally prohibited" and to tip some away from saying that it is "morally required."[32] Likewise, we might expect the {Minimize casualties} condition to tip some subjects toward saying that the decision is "morally required" and to tip some away from saying that it is "morally prohibited." Accordingly, the predictions for a comparison of the {Protect passengers} and {Minimize casualties} conditions are straightforward: The former should show more subjects saying that the sacrifice is "morally prohibited," and the latter should show more saying that it is "morally required."[33]

But what about comparisons with the {No laws} condition? This condition can be seen as a sort of neutral comparator, but with a cautious eye, as it is not an entirely natural baseline:[34] If subjects expect the law to

psychological internalization of law, see generally Yuval Feldman, The Law of Good People (2018); Richard H. McAdams, The Expressive Powers of Law (2015); Frederick Schauer, The Force of Law (2015); Bilz & Nadler, supra note 6.

[31] Defining what is normal, in this sense, may influence moral judgments through the act–omission distinction by setting an expectation about the natural course of things. As Professor Ronald Dworkin observed about the classic trolley problem:

> It is unclear what it means to let nature take its course. If it is natural to try to rescue five people at the cost of one, then throwing the switch is letting nature take its course. But perhaps "nature" means nonintelligent nature, so that a potential rescuer lets nature take its course by pretending that he is not there. But why should he?

Ronald Dworkin, Justice for Hedgehogs 298–99 (2011); cf. Adam Bear & Joshua Knobe, Normality: Part Descriptive, Part Prescriptive, 167 Cognition 25, 26–32 (2017) (providing experimental evidence for the argument "that people's normality judgments take into account both descriptive considerations (e.g., the statistical notion of the average) and more prescriptive considerations (e.g., what is morally ideal)").

[32] The expected effect of the {Protect passengers} condition on the share saying "morally permissible" could go either way, depending on whether more switch away from that category into "morally prohibited" or more switch in from "morally required." And the same reasoning applies, with switches in the other direction, for the {Minimize casualties} condition.

[33] As explained, the expected effect of either condition on the share saying "morally permissible" is ambiguous.

[34] Nonetheless, it still seems more informative for this thought experiment than other imaginable candidates for a neutral comparator. For example, an alternative might be to ask subjects to read the scenario without any mention of the law at all; yet that condition would be tricky to interpret as a baseline because its meaning would depend on what subjects are implicitly assuming about the background law (of the future)—a complication that is highlighted by this study's findings of law's influence on moral intuitions.

say *something* about safety priorities, at a time when driverless cars are common, then the {No laws} condition may suggest an implicit societal choice to deem any policy—prioritizing passengers, or minimizing casualties, or anything else—to be acceptable.[35] If so, the {No laws} condition may exert a pull of neutrality, drawing subjects away from the moral poles and toward the "morally permissible" answer. Comparisons of the {No laws} condition with the others, then, would be showing the net effects of law's directional influences and such a pull of neutrality.

II. Law's Influence

Telling people different things about the law leads to different moral judgments about Charlie's decision to sacrifice its sole passenger to save the two bicyclists. As seen in Table 1 and Figure 1, a comparison of the {Protect passengers} and the {Minimize casualties} conditions confirms the presence of law's directional pull.[36] The proportion of subjects who say that Charlie's decision is "morally prohibited" rises from 9% in the {Minimize casualties} condition to 16% in the {Protect passengers} condition.[37] Meanwhile, the proportion who say that the decision is "morally required" rises from 31% in the {Protect passengers} condition to 45% in the {Minimize casualties} condition.[38] Put differently, in the {Protect passengers} condition, there are about twice as many subjects saying the decision is "morally required" as saying "morally prohibited"; whereas in the {Minimize casualties} condition, that ratio increases to five times as many.

One might also compare each of these law conditions with the {No laws} condition—keeping in mind both that it is not an especially natural baseline and that the possibility of a pull of neutrality may complicate interpretation. Relative to {No laws}, the {Protect passengers} condition raises the share of subjects saying the decision is "morally prohibited" from 8% to 16%;[39] meanwhile, the share saying "morally required" does not seem to change by much.[40] Relative to the {No laws} condition, the

[35] This expectation may be lessened, however, by the fact that the survey subjects know they are taking a survey with an invented story—one that seems focused on the morality of the choice—and thus some may take the {No laws} statement as a cue to try to put any worry about the law out of their minds. This suggestion is purely speculative, of course.

[36] In a comparison of the {Protect passengers} and {Minimize casualties} conditions, there is no extra complication from the possibility of a pull of neutrality, as there will be in comparisons with the {No laws} condition.

[37] χ^2(1, N = 642) = 7.82, p = 0.005. Pearson's chi-squared tests are reported throughout. Note that the standard error for any individual proportion $\hat{p}$ is given by the usual formula, the square root of $\hat{p}(1-\hat{p})$ / N.

[38] χ^2(1, N = 642) = 14.25, p < 0.001.

[39] χ^2(1, N = 621) = 10.04, p = 0.002.

[40] χ^2(1, N = 621) = 1.30, p = 0.254. The standard reminder applies, throughout these analyses, that the lack of a statistically significant effect is not the same as evidence of zero effect.

{Minimize casualties} condition increases the share saying "morally required" from 35% to 45%;[41] meanwhile, the share saying "morally prohibited" seems about the same.[42] One possible explanation for these contrasts (between measurable change on one margin and a lack thereof on the other) is that some subjects are unmoved in their moral intuitions by hearing about a contrary law;[43] but another possibility is that a pull of neutrality in the {No laws} condition is masking the influence of the law conditions.[44]

TABLE 1: SACRIFICING THE PASSENGER

	Morally prohibited	Morally permissible	Morally required	*N*
Protect passengers	16.4%	53.1%	30.6%	311
Minimize casualties	9.1%	45.9%	45.0%	331
No laws	8.1%	57.1%	34.8%	310

[41] $\chi^2(1, N = 641) = 6.90, p = 0.009$.

[42] $\chi^2(1, N = 641) = 0.20, p = 0.652$.

[43] That is, it may be that those answering "morally required" in the {No laws} condition are not easily moved by hearing about the {Protect passengers} law, and those answering "morally prohibited" are not easily moved by hearing about the {Minimize casualties} law.

[44] To elaborate: Some subjects in the {Protect passengers} or {Minimize casualties} conditions, being no longer drawn toward the "morally permissible" answer by the pull of neutrality exerted by the {No laws} condition, may be inclined to choose one of the polar categories instead. Thus, in comparing the {No laws} and {Protect passengers} conditions, the observed shift toward "morally prohibited" might be due both to the law's directional influence and to relief from the pull of neutrality; whereas movement away from "morally required" might be offset by relief from the pull of neutrality. Likewise, in comparing the {No laws} and {Minimize casualties} conditions, the observed shift toward "morally required" might be due both to the law's directional influence and to relief from the pull of neutrality; whereas movement away from "morally prohibited" might be offset by relief from the pull of neutrality.

FIGURE 1A: SACRIFICING THE PASSENGER

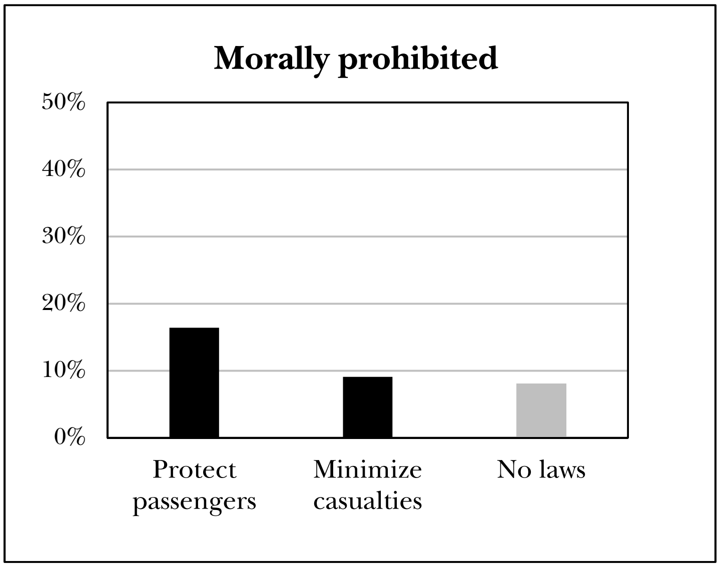
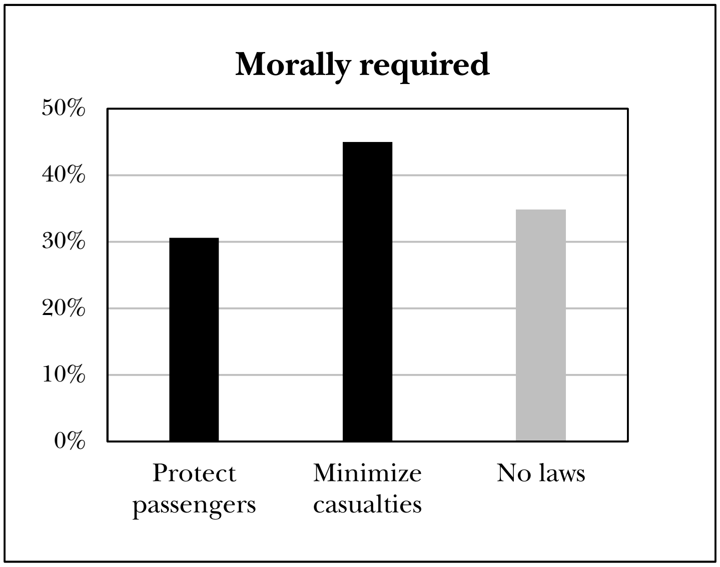


FIGURE 1B: SACRIFICING THE PASSENGER



A. *Whose Moral Responsibility?*

If someone says that Charlie's decision is "morally prohibited," whom do they blame? A follow-up question immediately after the main moral judgment question asks subjects to identify who "they think should bear some moral responsibility for that choice." Subjects may choose as many of the four given options as they see fit, and they may also choose "none of the above"; an open-ended text box invites them to suggest anyone else. Across the three conditions, 11% named Charlie as sharing in the moral responsibility; 66% named the company that created Charlie; 32% named the owner of the driverless car; 43% named the lawmakers; and 6% said none of the above.[45] (The shares do not add to one because the options are not mutually exclusive.)

These numbers should be taken with a grain of salt because the four answer options were preselected for the subjects. Still, looking at the *relative* attributions, two things seem to stand out:[46] First, lawmakers absorb a decent share of the assignment of moral responsibility—somewhat less than the company, but more than the car's owner. This finding speaks to a provocative question raised in this Symposium by Professor Tim Wu, who wonders whether in the near future the various defaults set by pervasive AI decisionmaking, through our acquiescence to them, will supersede the role of human-made law.[47] One might see this finding as a sign that the future public will not so readily let lawmakers off the hook, even when everyone knows that the AI systems will really be making the decisions. Or, from the perspective of the present, this finding may be seen as an expression of a generalized desire for lawmakers to deal with these hard questions before an AI does someday.[48]

[45] To be clear, these are shares of the subjects who say that Charlie's decision is "morally prohibited"—that is, those who morally object to the sacrifice of the passenger—across all three law conditions. This is not to ignore the distinction between moral blame and moral responsibility; for example, it would be coherent to ask subjects to assign moral responsibility for a blameless or a laudable decision (or even for a decision that has yet to be made).

[46] By relative attribution, I mean comparing the shares of people who assigned some moral responsibility to each of the four preselected actors. (The ordering of the four options is scrambled randomly for each subject.) Note, however, that these data do not tell us about the relative intensity of attribution. For example, a subject who names both Charlie and the car's owner might feel that the former bears less responsibility and the latter bears more; this difference would not appear in the data.

[47] As Professor Wu puts it: "Many of the developments that go under the banner of artificial intelligence that matter to the legal system are not so much new means of breaking the law but of bypassing it as a means of enforcing rules and resolving disputes." Tim Wu, Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems, 119 Colum. L. Rev. 2001, 2001 (2019) (citation omitted). The challenge to the legal system, then, would become the "existential challenge of supersession [by software]." Id.

[48] In concept it would be informative to consider each law condition separately, although in reality many more grains of salt must be taken, given the tiny sizes of the resulting subsamples—and again, we are only considering those subjects who answered "morally prohibited," the composition of whom differs across conditions. If one is willing

Second, these data suggest that people will not shy away from assigning moral responsibility to the creators of the AI—and for some people, even to the AI itself—for the choices made in such a dilemma.[49] Notably, among the options given, the company that created Charlie drew an attribution of moral responsibility from the most people. And even Charlie itself was named as morally responsible in part, by a small share of subjects. These relative attributions reflect how we currently envision the sophistication of such future AI pilots and how we imagine they might make such decisions.[50] As of now, it seems that relatively few subjects consider the AI itself to qualify as a moral agent. But in coming years, as we grow increasingly familiar with highly advanced and lifelike AI interfaces, one can only guess how our anthropomorphic imaginations might become more willing to assign moral agency to something—or will we say *someone*—like Charlie.

B. *Limitations*

Three sets of limitations are worth emphasizing. First, the standard disclaimer applies: What people say after reading a vignette might not reflect how they would react if similar events occurred in real life; and all the more so, for a story set in the future. In particular, this study's observations might be over- or understated, depending on various factors. For example, they might be overstated relative to a study in a natural setting (of future survey subjects) where people might not be paying much attention to the law. On the other hand, they might be understated relative to a future survey conducted after subjects have already had years of

to proceed with enough healthy skepticism: In the {Minimize casualties} condition, 50% of subjects assigned some moral responsibility to the lawmakers (N = 30); this may be unsurprising, as what the law says is also what Charlie actually does. In the {No laws} condition, 36% named the lawmakers (N = 25), which one might interpret as frustration for a failure of lawmaking. And yet there also seems to be ample attribution in the {Protect passengers} condition, with 43% naming the lawmakers as sharing in moral responsibility (N = 51), even though these subjects should be approving of the law, and even though Charlie's decision actually *violates* the law. To speculate, it thus seems possible that some sort of even more generalized attribution of moral responsibility to lawmakers is at work—perhaps for enabling such an AI-dominated state of the world in the first place.

[49] In their Symposium contributions, Professors Mala Chatterjee and Jeanne Fromer, as well as Professor Frank Pasquale, pose hard questions about the gaps in moral and legal accountability that might appear when we delegate decisionmaking to an AI. See Mala Chatterjee & Jeanne C. Fromer, Minds, Machines, and the Law: The Case of Volition in Copyright Law, 119 Colum. L. Rev. 1887, 1887 (2019) (discussing whether machines can be deemed to have mental states of the sort that would allow attributions of moral agency or of legal liability to them); Frank Pasquale, Data-Informed Duties in AI Development, 119 Colum. L. Rev. 1917, 1917 (2019) (discussing the potential for gaps in liability and responsibility when decisions are delegated to AI systems). This study's findings on the assignment of moral responsibility may be useful in calibrating such concerns, though of course it remains an open question how the future public will make such attributions.

[50] As noted above, the scenario intentionally says very little about how Charlie "thinks"—leaving this largely to the subject's imagination. See supra note 23 and accompanying text.

life experience with both driverless cars and the laws governing them. (That is to say, it may be expecting a lot of the present-day survey subjects to internalize an imaginary law of the future.) Moreover, the way the law is described in this study, without any mention of enforcement or liability, might also result in some understatement relative to a study with a more fully described legal regime—or relative to a future study done when those facts have become common knowledge.[51]

Second, this study is not designed to sort among the possible mechanisms of the law's halo. Do some subjects draw moral guidance or reasoning from what the law says? Do others see the law as a sign of society's moral norms? For some subjects, does the law implicitly set what is normal and what is deviant? For others, does it define social roles in a morally relevant way? Do some subjects place moral weight on the consequences of liability? Do others simply consider it a moral obligation to obey the law? Such psychological pathways may vary from person to person, and others may well be possible. Moreover, it goes without saying that while this study is designed to focus on the role of the law, many other sources of influence on our moral intuitions are worthy of study—and they may interact or interfere with the mechanisms of the law's halo in intriguing ways.

Third, there are limits of scope and generalizability. This survey includes only one scenario and focuses on only one decision by the AI; it also only considers two possible laws. Further work involving other variations of such dilemmas, and other possible laws or assignments of liability, is needed before any broad generalizations should be ventured.[52] Along these lines, a few suggestions follow.

CONCLUSION

This Essay presents an initial set of evidence suggesting that our moral intuitions about driverless car dilemmas can be influenced by the presence of the law. As this is but a single study, the most useful way to conclude may be to suggest a few dimensions of possible variation for

[51] See Huang, supra note 4, at 691–95 (showing evidence suggesting larger differences between contrasting law conditions when subjects are told that the deciding actor will be held liable than when subjects are told that there will be no liability despite what the law says). This study left liability open-ended in order to allow measuring of the subjects' natural tendencies in attributing moral responsibility. See supra section II.A. It would be an interesting extension to see if expressly assigning legal liability to a given actor might affect subjects' assignments of moral responsibility.

[52] While in prior work I have used survey experiments to explore the law's halo in a classic trolley dilemma, to my knowledge this study is the only one to have done so in the context of a driverless car dilemma. A different but related inquiry—about how the legal status of potential victims affects people's normative choices about whom to sacrifice (for example, whether people are less willing to save a jaywalker, or a criminal, relative to other potential victims)—has generated interesting findings in the MIT Moral Machine project. See Awad et al., supra note 8, at 59–63.

future work: First, different laws can be tested. For example, what if the law allows the car owner to control the "morality setting," dialing up or down the degree of priority for the passengers over outsiders? Second, other trade-offs can be posed. For example, the dilemma might involve increasing or decreasing risks, rather than causing harm for certain.[53] Third, the humans involved can be described with more morally relevant details. For example, what if the owner had bought this car believing that Charlie would put a priority on protecting passengers over protecting others? Or what if the passenger is in the driver's seat, and there *is* a steering wheel—but he declines to take over from the AI? And fourth, the AI's capabilities can be more fully specified. For example, what if it is known that Charlie learned to drive from watching millions of hours of videos of humans driving? Or what if Charlie is constrained to obey the law at all times? Or what if Charlie is not only the driver of a car, but an all-purpose companion who could easily pass the Turing test in conversation—even when discussing timeless moral dilemmas?

[53] For example, the car may be choosing whether to drive a bit to the left within its highway lane, closer to a motorcyclist on that side, and a bit more distant from a tractor-trailer on the right side; this trades off risk to the motorcyclist with risk to the car's passengers. See Bonnefon et al., supra note 17, at 503 (describing such a scenario as one example of a "statistical trolley dilemma"); cf. Barbara H. Fried, What *Does* Matter? The Case for Killing the Trolley Problem (or Letting It Die), 62 Phil. Q. 505, 506 (2012) (arguing that trolley-like thought experiments are limited in real-world relevance because they do not cover cases of "uncertain risk of accidental harm to generally unidentified others").