PNAS
www.pnas.org

Supplementary Information for
From Driverless Dilemmas to More Practical Common-Sense Tests for Automated Vehicles

Julian De Freitas, Andrea Censi, Bryant Walker Smith, Luigi Di Lillo, Sam E. Anthony, Emilio Frazzoli

Julian De Freitas
Email:defreitas@g.harvard.edu

This PDF file includes:

Supplementary text
Tables S1
SI References

Supplementary Information Text

How Ethics Is Relevant to Stakeholders in the Technical, Legal, and Social Spheres

Technical sphere (development and deployment)

From a technical standpoint, the road to creating reliable AVs is longer than often expected, as can be evidenced by the missed deadlines in developers' public roadmaps (1). The development of autonomous systems presents new engineering challenges, especially in achieving and demonstrating safety. In fact, organizations such as the International Organization for Standardization (ISO) and the World Forum for Harmonization of Vehicle Regulations, among many others, are developing new standards and regulations for this purpose (2). Technical stakeholders and their ongoing ethical challenges related to driving behavior include:

- **Manufacturers** try to build technologies that satisfy currently—and perhaps necessarily—vague and uncertain expectations of safety from regulators and the public. They have little concrete guidance on how to solve ethically relevant technical issues (3), and solutions are developed in isolation with little sharing of insight with competitors, with a few exceptions (2, 4). Manufacturers hesitate to discuss ethical concerns with the public, because in the current atmosphere of uncertainty, no position is defensible (5). There have been numerous documented ethical failings of manufacturers, e.g., the Volkswagen Emissions scandal (6), and Ford Pinto Fuel Tank fiasco (7).
- **Engineers** are individual moral entities whose moral values and behaviors may not necessarily align with those of their employers. Societies like the Institute of Electrical and Electronic Engineers clearly state that the engineer's act of design should be ethical and that ethical requirements should be part of the design (8).
- **Academic researchers** in AI generally and robotics specifically are interested in "computable" ethics (9-21), but there is no consensus on what approach to use. Moreover, the existing approaches do not scale to the complexity of AV scenarios, which can involve hundreds of possibly conflicting constraints (3).

Legal sphere (governmental requirements, civil and criminal liability, and insurance)

AV-specific regulations are sparse and incomplete. Many countries, states, and municipalities have expressly authorized or otherwise facilitated AV testing on public roads (22). Fewer have expressly addressed full-scale deployment. Scholars, governments, and international organizations such as the E.U. and the U.N. are still exploring both the process and the content of AV-specific regulation: how existing rules apply to AVs, how new rules should be shaped, how some of these rules might be encoded in a computable way, and how notions of legal responsibility should apply in the context of autonomous physical agents (23-31). Actors and their challenges include:

- **Policymakers** address the relationship between new technologies and the broader conditions and goals of a society (28). They vary in their relevant expertise, resources, and relationships with AV companies. Policymakers in Singapore, for example, have begun working with AV companies on concrete ethical requirements (32). Policymakers may also take different philosophical views of regulation—including on the propriety and the application of the 'precautionary principle' (33).
- **Technical regulators** translate between the technical and legal domains by conducting research, developing standards, supervising type approval (in the E.U.) or

self-certification (in the U.S.) of vehicles and their parts, granting exemptions, undertaking investigations, and overseeing recalls. They may need to justify their decisions to policymakers, courts, and the public at large. The U.S. National Highway Traffic Safety Administration expressly included "ethical considerations" in its original AV policy (34) before removing this from subsequent versions (35).

- **Standards organizations** foster technical consensuses that can implicate issues of regulation and liability. These organizations may be public, private, or quasi-public. Among many others: SAE developed the leading definitions for automated driving (36), ISO has developed influential standards on functional safety and 'safety of the intended function' (2, 4), and IEEE has developed principles for ethically aligned design (37).
- **Testing entities** evaluate the safety of vehicles and their components with reference to standards that may be mandatory or voluntary, and quantitative or qualitative. EuroNCAP now rates certain driver assistance systems, including automated emergency braking (38).
- **Insurance companies** quantify certain economic risks that serve as incomplete and imperfect proxies for societal risks. These determinations, which are partly expressed in premiums, can affect the decisions of regulators, companies, and consumers. Insurers are already evaluating AV risks, and one was reportedly involved in Uber's settlement in Arizona.
- **Courts** evaluate the lawfulness and reasonableness of behavior—by regulators, companies, and even products. In the case of injuries that have already occurred, these evaluations are necessarily retrospective. One court has already seen early litigation over a test AV that allegedly "drove in … a negligent manner" (39).
- **Lawyers**, like engineers, try to accomplish their clients' goals using certain tools (in this case, abstract legal theories and knowledge) subject to certain constraints. Like engineers, they respond to both intrinsic incentives (e.g., commitment to justice and professional ethics) and extrinsic incentives (e.g., licensing and compensation).
- **Legal scholars** assess the relationship between legal frameworks and emerging technologies like automated driving from descriptive, predictive, and prescriptive perspectives in a way that can produce both clarity and confusion in other domains (31). Even a partial list of these lengthy scholarly articles is itself lengthy (40).

Social sphere (perception and acceptance)

The broad public's knowledge of AVs is mainly informed by the public relations materials released by industry and by news reports that sometimes tend toward either an "everything is perfect" angle (41-43) or a "doomsday" angle (44-49). There is a risk that uninformed public opinions combined with fears about a loss of control will lead to delays in the introduction and spread of a potentially useful technology, re-tracing the story of other potentially useful technologies, as in the categorical opposition to genetically modified organisms (50). On the flip side, there is also a risk that the public may fall in love with overhyped technologies that turn out to be dangerous (51, 52).

- **The public (possible AV users and other road users)** is both excited and anxious about the future—and interested in the macabre, as can be seen in reactions to so-called driverless dilemmas (5, 44-49, 53). A widespread worry is that AVs will make incorrect moral decisions in high-stakes scenarios. This is likely due to an understandable, yet unfounded, tendency to moralize new technologies because of their unfamiliarity and threat to a consumer's control (54). Similar responses are found in the broader context of new applications of AI (55). The public's perception of AV riskiness will also be informed by the safety ratings released by consumer organizations.

- **Social scientists** want to know how people perceive the ethical nature of AVs, since these perceptions may answer questions about human nature and inform efforts to increase public acceptance of AVs, as is achieved through usual marketing research efforts. There has recently been a small surge of studies in this area (56-61), with questionable relevance to the field of automated driving (62).
- **Science communicators** have largely focused on communicating potential ethical dilemmas involving AVs, given the public's interest in this topic and the incentive of clickbait headlines. Some alarm-raising articles resonate with public fears (44-49) but rely too heavily on trivialized descriptions of AVs that distract from the more prevalent public health issues.

Table S1. Review of the main affirmative and negative answers that have been offered to the question of whether driverless dilemmas are a good way to think about AV ethics.

No, they are not useful.	**Yes, they are useful.**
Too rare. On real roads, vehicles rarely, if ever, have only two choices, with no way to buy more time or space.	**Not too rare.** Once enough AVs are on the road, the probability of dilemmas will multiply.
Unrealistic. AVs cannot simultaneously have enough control to choose whom to hit but not enough control to avoid the dilemma entirely. Also, AVs cannot gather social characteristics such as whether a pedestrian is a criminal or valuable to society.	**Not unrealistic.** Surely in at least some cases, the AV will only have a limited set of available actions, all with moral consequences. Also, some characteristics, like age, can be reliably inferred from cues like height.
Too simplified. The dilemmas ignore legal obligations and do not grapple with risk and uncertainty.	**Not simplified.** The dilemmas are not meant to capture all aspects of the situation, but only to highlight the interesting ethical ones.
Lack of public support. Less than 20% of people condone the idea of AVs consulting social characteristics when making ethical decisions (62, 63). This suggests that data gathered from moral dilemma studies should not be considered as 'preferences', since they are the result of cornering people into a forced-choice. In short, this "ethical" test is itself unethical.	**Presence of public support.** We should distinguish between what people say and how they behave in practice. For instance, even though people generally think that AVs should sacrifice their passengers for the greater good, they *themselves* prefer to ride in AVs that protect their passengers at all costs (57). So, even if people say they are against AVs having moral preferences, they may still condone these preferences at an emotional level, e.g., saving a child over an adult (although this too is fraught (61)).

Wrong safety goal. Planning for trolley problems requires identifying when you are in one in the first place, yet usually you are not. Telling the difference between a trolley-like dilemma and an ordinary emergency requires significant abilities. Instead of planning for such 'special' situations, just focus on the more tractable goal of avoiding harm.	**Right safety goal.** Expecting AVs to be *perfect* at identifying trolley problems is too high of a bar. As long as AVs make the best moral choice in light of existing evidence, that is sufficient—much as airbags are sufficiently good at deploying only in emergencies.
Disturbing. We already know that majorities are likely to directly or indirectly discriminate against minorities (64, 65). AVs should be designed to overcome rather than to extend these biases.	**Fair.** We can avoid such discrimination by democratically crowdsourcing ethical preferences for AVs.

SI References

1. Walker Smith B (2020) How reporters can evaluate automated driving announcements. *J. Law Mob.* 2020(1):1-16.
2. ISO (2019) PAS 21448-Road vehicles-safety of the intended functionality. International Organization for Standardization.
3. Censi A*, et al.* (2019) Liability, ethics, and culture-aware behavior specification using rulebooks. *2019 International Conference on Robotics and Automation (ICRA)*:8536-8542.
4. ISO I (2011) 26262: Road vehicles-Functional safety. International Standard ISO/FDIS. 26262.
5. Nowak P (2018) The ethical dilemmas of self-driving cars. https://www.theglobeandmail.com/globe-drive/culture/technology/the-ethical-dilemmas-of-self-drivingcars/article37803470/
6. Hotten R (2015) Volkswagen: The scandal explained. BBC News. https://www.bbc.com/news/business-34324772.
7. Dowie M (1977) Pinto madness. Mother Jones https://www.motherjones.com/politics/1977/09/pinto-madness/.
8. How JP (2018) Ethically aligned design [From the Editor]. *IEEE Control Systems Magazine* 38(3):3-4.
9. Gensler HJ (2002) *Formal Ethics* (Routledge).
10. Anderson M & Anderson SL (2011) *Machine ethics* (Cambridge University Press).
11. Lin P, Abney K, & Jenkins R (2017) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (Oxford University Press).
12. Anderson M & Anderson SL (2018) GenEth: A general ethical dilemma analyzer. *Paladyn* 9(1):337-357.
13. Kasenberg D & Scheutz M (2018) Norm conflict resolution in stochastic domains. *Thirty-Second AAAI Conference on Artificial Intelligence.*
14. Wachter S, Mittelstadt B, & Floridi L (2017) Transparent, explainable, and accountable AI for robotics. *Sci. Robot.* 2(6).
15. Charisi V*, et al.* (2017) Towards moral autonomous systems. *arXiv preprint arXiv:1703.04741*.

16. Yu H*, et al.* (2018) Building ethics into artificial intelligence. *arXiv preprint arXiv:1812.02953*.

17. Yilmaz L (2017) Verification and validation of ethical decision-making in autonomous systems. *Proceedings of the Symposium on Modeling and Simulation of Complexity in Intelligent, Adaptive and Autonomous Systems*, pp 1-12.

18. Dennis LA, Fisher M, & Winfield A (2015) Towards verifiably ethical robot behaviour. *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.*

19. Fisher M, List C, Slavkovik M, & Winfield A (2016) Engineering moral machines. *Informatik-Spektrum, Springer.*

20. Dennis L, Fisher M, Slavkovik M, & Webster M (2016) Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77:1-14.

21. Alvarez MC*, et al.* (2017) Implementing Asimov's first law of robotics. *30th Norsk Informatikkonferanse, NIK*:27-29.

22. Hao K (2017) At least 47 cities around the world are piloting self-driving cars. Quartz. https://qz.com/1146038/at-least-47-cities-around-the-world-are-piloting-self-driving-cars/.

23. Abraham KS & Rabin RL (2019) Automated vehicles and manufacturer responsibility for accidents: A new legal regime for a new era. *Va. L. Rev.* 105:127.

24. Law Commission of England and Wales SLC (2018) Automated Vehicles: A joint preliminary consultation paper. Nov. https://www.lawcom.gov.uk/project/automated-vehicles/.

25. Pagallo U (2013) *The laws of robots: crimes, contracts, and torts* (Springer).

26. Hage J (2017) Theoretical foundations for the responsibility of autonomous agents. *Artificial Intelligence and Law* 25(3):255-271.

27. Smith BW (2017) Automated driving and product liability. *Mich. St. L. Rev.*:1.

28. Smith BW (2017) How governments can promote automated driving. *NML Rev.* 47:99-138.

29. Choi BH (2019) Crashworthy code. *Wash. L. Rev.* 94:39.

30. Smith BW (2016) The Trolley and the Pinto: Cost-benefit analysis in automated driving and other cyber-physical systems. *Tex. A&M L. Rev.* 4:197-208.

31. Smith BW (2016) Lawyers and engineers should speak the same robot language. *Robot Law*, eds Calo R, Froomkin AM, & Kerr I (Edward Elgar Publishing), pp 78-101.

32. Council SS (2019) TR 68: Autonomous vehicles.

33. Bourguignon D (2015) The precautionary principle: definitions, applications and governance. *Eur. Parliam. Res. Serv* 573:876.

34. Administration NHTS (2016) Federal automated vehicles policy. U.S. Department of Transportation https://www.transportation.gov/sites/dot.gov/files/docs/AV%20policy%20guidance%20PDF.pdf.

35. Administration NHTS (2017) Automated driving systems. U.S. Department of Transportation. https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf.

36. SAE (2016) Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. SAE International.

37. IEEE (2019) Ethically aligned design. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/dam/ieee-

standards/standards/web/documents/other/ead1e.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined.

38. EuroNCAP (2019) Driver assistance systems. EuroNCAP https://www.euroncap.com/en/ratings-rewards/driver-assistance-systems/.

39. Nilsson v. General Motors Co. 3:18-cv-00471 (U.S. District Court for the Northern District of California 2018).

40. Smith BW & Neznamov A (2019) It's not the robot's fault: Russian and American perspectives on responsibility for robot harms. *Duke J. Comp. & Int'l L.* 30:143-163.

41. Hawkins AJ (2020) Here are Elon Musk's wildest predictions about Tesla's self-driving cars. The Verge. https://www.theverge.com/2019/4/22/18510828/tesla-elon-musk-autonomy-day-investor-comments-self-driving-cars-predictions.

42. Hawkins AJ (2019) Waymo's driverless car: ghost-riding in the back seat of a robot taxi. The Verge. https://www.theverge.com/2019/12/9/21000085/waymo-fully-driverless-car-self-driving-ride-hail-service-phoenix-arizona.

43. Boudette NE (2019) Despite high hopes, self-driving cars are 'way in the future'. New York Times. https://www.nytimes.com/2019/07/17/business/self-driving-autonomous-cars.html.

44. Donde J (2017) Self-driving cars will kill people. Who decides who dies? Wired. https://www.wired.com/story/self-driving-cars-will-kill-people-who-decides-who-dies/

45. Edmonds D (2018) Cars without drivers still need a moral compass. But what kind? The Guardian. https://www.theguardian.com/commentisfree/2018/nov/14/cars-drivers-ethical-dilemmas-machines

46. Johnson CY (2018) Self-driving cars will have to decide who should live and who should die. Here's who humans would kill. The Washington Post. https://www.washingtonpost.com/science/2018/10/24/self-driving-cars-will-have-decide-who-should-live-who-should-die-heres-who-humans-would-kill/?noredirect=on&utm_term=.49cd910f606a

47. Lester CA (2019) A study on driverless-car ethics offers a troubling look into our values. The New Yorker. nyer.cm/jGCFyZA

48. Markoff J (2016) Should your driverless car hit a pedestrian to save your life? New York Times. https://www.nytimes.com/2016/06/24/technology/should-your-driverless-car-hit-a-pedestrian-to-save-your-life.html.

49. Shariff A, Rahwan I, & Bonnefon J-F (2016) Whose life should your car save? New York Times. https://www.nytimes.com/2016/11/06/opinion/sunday/whose-life-should-your-car-save.html.

50. Scott SE, Inbar Y, & Rozin P (2016) Evidence for absolute moral opposition to genetically modified food in the United States. *Perspect. Psychol. Sci.* 11(3):315-324.

51. Freiherr G (2014) The eclectic history of medical imaging. Imaging Technology News. https://www.itnonline.com/article/eclectic-history-medical-imaging.

52. Hunt DE, Kuck S, & Truitt L (2006) *Methamphetamine Use: Lessons Learned* (Abt Associates Cambridge, MA).

53. Lin P (2013) The ethics of autonomous cars. The Atlantic. https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/

54. Assis S (2018) Seven out of 10 U.S. drivers fear self-driving cars, AAA says. Market Watch. https://www.marketwatch.com/story/seven-out-of-10-us-drivers-fear-self-driving-cars-aaa-says-2018-05-22
55. Leslie D (2019) Human compatible: Artificial intelligence and the problem of control. *Nature* 574(7776):32-34.
56. Awad E*, et al.* (2019) Drivers are blamed more than their automated cars when both make mistakes. *Nat. Hum. Behav.*:1-10.
57. Bonnefon J-F, Shariff A, & Rahwan I (2016) The social dilemma of autonomous vehicles. *Science* 352(6293):1573-1576.
58. Shariff A, Bonnefon J-F, & Rahwan I (2017) Psychological roadblocks to the adoption of self-driving vehicles. *Nat. Hum. Behav.* 1(10):694-696.
59. Li J, Zhao X, Cho M-J, Ju W, & Malle BF (2016) From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. SAE Technical Paper.
60. Savulescu J, Kahane G, & Gyngell C (2019) From public preferences to ethical policy. *Nat. Hum. Behav.* 3(12):1241-1243.
61. De Freitas J & Cikara M (2020) Deliberately prejudiced self-driving cars elicit the most outrage. *Cognition* 208:104555.
62. De Freitas J, Anthony SE, Censi A, & Alvarez G (2020) Doubting driverless dilemmas. *Perspect. Psychol. Sci.* 15:1284-1288.
63. Bigman YE & Gray K (2020) Life and death decisions of autonomous vehicles. *Nature* 579(7797):E1-E2.
64. Pratto F, Sidanius J, & Levin S (2006) Social dominance theory and the dynamics of intergroup relations: Taking stock and looking forward. *Eur. Rev. Soc. Psychol.* 17(1):271-320.
65. Schmidt H (2020) The way we ration ventilators is biased. New York Times. https://www.nytimes.com/2020/04/15/opinion/covid-ventilator-rationing-blacks.html.