Ethical Algorithms in Autonomous Vehicles: Reflections on a Workshop

Nicholas G. Evans(✉)

Department of Philosophy, University of Massachusetts Lowell, Lowell, USA
Nicholas_Evans@uml.edu

Abstract. This chapter summarizes and expands on the breakout session at the 2019 Autonomous Vehicles Symposium in Orlando, FL, titled "Ethical Algorithms in Autonomous Vehicles." First, the content of the workshop presentations is summarized, covering technical and nontechnical detail. Second, the discussion during, and on the margins of the workshop is summarized from the perspective of the authors, consistent with the Chatham House Rule. Conclusions are posed for industry, academia, and government readers.

Keywords: Autonomous vehicles · Ethics · Crash algorithms · Safety data · Algorithmic bias · Risk management

1 Introduction

This chapter summarizes and expands on the breakout session at the 2019 Autonomous Vehicles Symposium (AVS) in Orlando, FL, titled "Ethical Algorithms in Autonomous Vehicles." This workshop-styled session was funded through the National Science Foundation grant by the same name, and was centered around ethical issues arising from the development and implementation of autonomous vehicles (AVs). The workshop took place over two days of the AVS, Monday 15th and Tuesday 16th of July.

The primary aim of the workshop was to bring together internationally recognized and emerging scholars in ethics and policy to present new work in machine ethics and the ethics of AVs. The two days of programming featured discussion and conceptual innovation in the ethics of autonomous vehicles, including an open forum to identify emerging issues and develop collaborations for future work. Presenters were drawn from diverse career levels; and from both private and academic enterprise.

The central motivation for such a workshop was two-fold. First, although considerable attention has been paid to the basics of ethics in AVs, very little work has been done to determine a) what specific ethical theories say about machine behaviors; and b) what kinds of behaviors and traits of AVs are of greatest concern. To date, a large proportion of work has focused on surveys of consumer preferences (Bonnefon et al. 2016; Awad et al. 2018), or preliminary work establishing the case *that* there is something ethically important about autonomous vehicles (Goodall 2014a, b, 2016a, b; Lin 2015). The aim of this workshop was to extend this analysis, and that of the

G. Meyer and S. Beiker (Eds.): AVS 2019, LNMOB, pp. 161–169, 2020.
https://doi.org/10.1007/978-3-030-52840-9_15

previous year's breakout session (Goodall et al. 2018), and allow for a frank discussion of ethical challenges facing autonomous vehicles, informed by the expert community AVS attracts.

While individual papers and their content are available on the AVS 2019 website, discussion took place under the Chatham House Rule (CHR). CHR is a common tool in security and diplomatic discussions, in which comments from participants may be quoted, but may not be attributed or identified in any way. This has a marked advantage over closed workshops in which information may be spoken freely, but never disclosed, by providing nonparticipants a record of the conversation that occurred. However, limitations arise because unless authors choose to personally identify themselves and their comments, it is difficult to vet information.

This chapter proceeds as follows, and with the following methodological turn. In the next section, we survey the conference proceedings, covering the information relayed by the invited speakers at the workshop. We then turn to the discussion at, or on, the margins of the workshop, and thoughts by speakers and participants on future work in the ethics of AVs. We conclude with some recommendations for those working in industry, academia, and government.

2 Workshop Content

The workshop was broken into two parts, covering two full breakout sessions of content. Each day began with opening remarks and logistics by the conference organizers, followed by a series of talks, each with time allocated for open discussion.

The first session consisted of a keynote describing preliminary research into modelling ethical decisions for autonomous vehicles using naturalistic crash data. This presentation, by Nicholas G. Evans and Rocco Casagrande, argued that consumer preference models of autonomous vehicles, including the MIT Moral Machine project, fail to give accurate insights into ethical issues on two counts. The first count being that these projects typically stipulate pairwise comparisons between two scenarios, without considering a full range of options, or exploring what options might exist for AVs. Secondly, consumer preferences fail to give ethical guidance because they simply display preferences, rather than considered reasons for what autonomous vehicles should do. The presentation concluded with an examination of an empirical case, showing how parametric models could be used to explore a full range of ethical features for important individual cases.

Katherine Evans then spoke about if, and how, the AVs should be limited in the kind of information they have on hand. With the arrival of 5G technology, as well as the increased efficiency of vehicle-to-vehicle communication, vehicle-to-device communication, and the Internet of Things, looms a second ethical question: what morality requires an AV to know about its decision context, and—perhaps even more importantly—what an AV should not know about the users in its environment. Embracing a liberal view on data protection laws and technical capability, the future of AV decisions could look more like a real-life rendition of the Moral Machine Experiment; a world where AVs may be able to identify the old, the rich, and the criminals in their midst, and incorporate these features into their moral deliberation. Other features may also

seem topically salient: the health status of different pedestrians, the credit score of a driver, or even their tacit and explicit social and political affiliations. The collection and exchange of user data across different devices could afford an AV a quasi-omniscient perspective from which to make moral decisions, but it is a separate question to know which types of information, if any, should make a moral difference.

A panel discussion then ensued examining particular challenges in the ethics and policy of designing navigation algorithms for AVs. First, Duncan Purves explored whether the asymmetry in public opinion between "autonomous weapons" and AVs is coherent. He articulated some obvious and subtle reasons for the asymmetry by way of addressing several recent arguments for banning autonomous weapons systems, advanced by some academics and NGOs, and found that—perhaps surprisingly—some of these objections to autonomous weapons systems also seem to apply to AVs. He then suggested that the difference in public attitudes about AVs and autonomous weapons systems is perhaps best explained by our feelings about the morality of the larger enterprises in which they are deployed: transportation and war. This compels us to justify the larger transportation system, of which AVs are only a recent innovation.

Next, Damien Williams tackling the problem of the dominance of Western ethical perspectives in the development of AVs. In considering the question of what AVs ought to do, Williams argued, designers, coders, and trainers will need to develop new ways of training and categorizing the decision-making processes of the algorithmic systems at work in AVs to account for the cultural and moral concerns of nonwestern societies. By including different nonwestern understandings of nonhuman agency from Asian and West African societies, we can explore new ways of thinking about assemblages of human and machine action, toward the maintenance and enrichment of human and nonhuman life. With this done, we can train AVS to make decisions in a wide variety of global cultural contexts, to address the needs and concerns a wide range of stakeholders.

Finally, Sarah Thornton showcased the Designing for Human Values (DHV) framework including a set of concepts, methodologies and tools for addressing ethical considerations throughout the engineering of a technology. DHV has its roots in Value Sensitive Design (VSD), which is an open-ended design framework that helps to analyze technology in terms of the human values that technology expresses. As a framework, VSD prompts the designer to focus on a broad set of stakeholders impacted by the technology under consideration (e.g. users, policy makers, the environment, the public), the values attached to those stakeholders (e.g. privacy, trust, profit), and the value tensions that can arise between different competing values and associated stakeholders. DHV adds structure to the underlying VSD framework so engineers can incorporate ethical considerations efficiently, reliably and consistently when designing a technology: exposing relevant ethical issues related to a particular technology upstream in the engineering design process; and prompting engineers to identify and reason through design options in more detail, and with a more informed, nuanced and critical eye, than they would have otherwise. She presented results from several 3-h DHV Workshops with diverse teams from industry and academia in order to demonstrate the use of the DHV framework.

Geoff Keeling then presented ongoing work on the kinds of confidence AVs ought to have in certain objects. First, AVs are morally required to exercise *due caution*

around vulnerable road-users such as pedestrians and cyclists. Presumably, this amounts to reducing speed and performing maneuvers to lower the risk of colliding with the road-user within some morally acceptable range. Second, this same degree of caution is not, in general, required towards road-users *inside* other vehicles. The asymmetry in the amount of caution required is explained by the fact that the expected harm to vulnerable road-users is significantly greater than the expected harm to non-vulnerable road-users in collisions with similar impact velocities. Hence the true class for an object, e.g. pedestrian, is morally relevant insofar as the morally right level of caution for the AV to exercise towards an object depends on what kind of thing the object is. Third, in scenarios like the Tempe collision, where the AV is uncertain about the classification for an object, the morally right level of caution is also uncertain. My aim in this paper is to investigate the degree of caution which AVs are morally required to exercise in scenarios like these. The view that I defend is deontological. Roughly, my thesis is that the AV should behave *as if* an object is a vulnerable road-user just in case it is reasonable for an epistemic agent with the same evidence as the AV to believe that the object is a vulnerable road-user. Conversely, the AV is permitted to behave *as if* an object is a non-vulnerable road-user just in case it is reasonable for an epistemic agent with the same evidence as the AV to believe that the object is a non-vulnerable road-user. I spell-out the meaning of these conditions with reference to the AV's probability distribution over the different classes which an object might belong to.

Kendra Chilson presented on the issue of consumer trust, and its importance in the development of autonomous vehicles (A.V.s). Without this trust, A.V.s would not only under-perform, they could be unusable and dangerous. Chilson gave an epistemic account of trustworthiness that beyond the conditions for manufacturers to cultivate consumers' subjective trust. Instead, her account identified a "robust trustworthiness," which fully justifies consumers' trust in A.V.s, based on appropriate indicators of trustworthiness. She developed six desiderata for autonomous systems, based on analogy to automatic technologies:

1) Repeatability—whether the system can be put back into the same state and produce the same outcome
2) Predictability—whether an expert can determine, based on input, what the system will output
3) Reliability—whether anyone can depend on the system's behavior by forming expectations about its behavior through repeated interactions
4) Transparency—whether an expert can understand what is happening within the system in real time
5) Re-constructability—whether the system's trajectory toward producing a certain output can be reconstructed after the fact (even if that trajectory is non-repeatable)
6) Explicability—whether the system can give a high-level explanation of what it is doing and why that is accessible to anyone

She then argued that these concepts can be divided along two orthogonal axes: System Externality vs. Internality, and Expert vs. Non-expert groupings. I will show why, to establish appropriate trust for non-expert consumers, both external and internal non-expert indicators must be present, and internal expert indicators must be accessible by appropriate authorities. Then through conceptual analysis, I will show that several of

these desiderata entail the other indicators needed to establish robust trustworthiness for A.V.s (given certain background assumptions), and give recommendations for incorporating these features into A.V.s.

The final panel concerned the wider effects of AVs on society. William Bauer argued that the widespread introduction of AVs presents macro-level socioeconomic concerns in addition to micro-level 'ethics on the road' dilemmas. His paper examined the impacts of AVs on our basic socioeconomic structure. Given the advent of self-driving freight trucks and taxis, for instance, millions of drivers could become unemployed very quickly. Citizens and leaders should address these kinds of macro-level social justice issues in advance in order to forestall the worst outcomes. To plausibly address these issues, he and his coauthor applied the principles of John Rawls' theory of distributive justice to the question of AVs. Focusing on the cases of truck and taxi drivers, we argue that the principles of Justice as Fairness support several possible policy-guiding norms that can be used to develop specific regulatory policies and ensure smoother economic transition as AVs are implemented on a broad scale.

Next, Johannas Himmelreich examined passenger settings for AVs. He identified conflicts between values in AV programming, and argued that passengers should be allowed to set the parameters to solve such value conflicts. Importantly, however, the parameter setting must be interdependent and passengers should not be allowed to solve conflicts independently of each other. Figuratively speaking, a passenger should have only one control dial to solve the value conflicts instead of multiple dials. The two conflicts Himmelreich identified were between mobility (e.g. time expected to arrive at a destination) and safety (e.g. route planning and speed control in navigating around obstacles); and between passenger-interests and outsider-interests, including collision management and speed control for passenger comfort. He provides an ethical analysis of these four values in conflict and draw on basic microeconomics to formalize the conflict. He defended a dependent passenger parameter setting because this promotes the meta-values of pluralism, human agency, and fairness.

Finally, Carole Turley Voulgaris examined the potential for connected and autonomous vehicles' (CAVs') potential to improve users' quality of life by reducing the frustration and inefficiency associated with traffic congestion. Traffic congestion is a function of the ratio of the number of vehicles using a roadway (volume) and the maximum number of vehicles that the roadway can accommodate (capacity). Vehicle connectivity and autonomy could indeed reduce congestion by enabling fleets of vehicles to coordinate their movements more efficiently, thereby increasing the effective capacity of a roadway. However, since CAV users—freed from the task of vehicle operation— could use their travel time for more pleasant or productive activities, automation would also increase travelers' tolerance for traffic congestion, increasing the demand for motorized travel and likely returning congestion to (and even beyond) levels experienced prior to the introduction of CAVs. The negative effects of vehicular congestion extend beyond vehicle users' lost time to other harms shared with non-users, such as pollution exposure, climate change, and hostile land development patterns. By increasing travelers' tolerance for congestion, CAVs have the potential to shift the burden of congestion-related harms from vehicle users to non-users. Since vehicle ownership is highly correlated with income —and this relationship may be even stronger for CAVs—this would represent a benefit to higher-income households

at the expense of lower-income households. Modifications to vehicle routing algorithms, well-designed roadway user fees, proactive land-use planning, and policies to encourage vehicle sharing could facilitate a more just distribution of the benefits of CAVs.

3 Workshop Discussion

As a preliminary discussion, the appearance of the 2019 novel coronavirus outbreak disrupted the timeline for the drafting of this chapter due to commitments from some of the workshop panelists and conveners. As such, and on a reduced timeframe, this discussion is based primarily on the comprehensive minutes collected by the second author, combined with the reflections of the primary author.

"Before The Crash" Ethics

A primary concern raised by participants at the workshop was the continued focus of ethical analysis on crash algorithms, rather than what one participant described as "the ethics of the crash, before the crash." That is, rather than attempting to resolve tradeoffs that arise in emergent scenarios (including crashes), members of the audience were interested in whether algorithms could be trained to recognize potential crashes and navigate around those situations without having to resolve them at the point of catastrophe. In the case provided by Evans and Casagrande, involving a human tailgater, the ideal would be to resolve the tailgating quickly rather than wait until an emergent case occurred.

This is something that ethicists have begun to consider, as part of a broader attitude towards the ethics of risk in the setting of AVs (Goodall 2016b). Such a strategy presents novel ethical issues, such as the degree to which an AV could hold up traffic in an effort to slow down to minimize the chance of an injurious collision with the tailgater; or speed up, potentially past the speed limit, to get out of their way (e.g. Jiang et al. 2005). We might call this a problem of *vagueness* regarding our emergent cases: attempting to avoid the case leads to a novel set of risk management problems, which carry with them their own ethical issues.

Importantly, however, considering strategies to avoid these emergent cases does not mean we can ignore them. The primary reason for this is that some kind of emergent case will surely exist even in the most optimized system. In the case of the tailgater, we could provide an example of a networked set of cars (including, potentially, human driven cars) that obviates the need for the decisions described by Evans and Casagrande. At the same time, however, such a network may have cascading effects that arise from small misalignments in the decisions of vehicles—in the same way that stock trades between algorithms can cause sudden shocks in stock prices. This is a novel potential emergent even, albeit different from the example of the tailgater. The central point here is that while good engineering might eliminate some ethical dilemmas, they will almost certainly leave some open (and even introduce new ones).

Concrete Analysis of Ethical Problems

A second issue raised during the workshop concerned the lack of specificity about what constitutes and ethical theory about AV behavior. While "applied ethics" is full of research *describing* problems, it is famously unclear about what comprehensive moral views people ought to take about those problems. The rare exception to this is military ethics, in which the combination of rigorous moral analysis and the long history of International Humanitarian Law provide a basis to think about the ethics of risk and come up with robust, detailed conclusions.

Rather, what has arisen in the ethics of AVs are broad assessments of "issues" and a larger vacuum of critical analysis, high quality empirical ethics work, and decision-making tools. Each has their own place, but none of the work in either is responsive to any of the others. This is a significant methodological problem, as neither program of work—conceptual, empirical, or decision-making—has the tools, in isolation, to provide necessary guidance to OEMs and other stakeholders. Deep conceptual work, like the work done in military ethics, is needed to very precisely articulate what our obligations are to road users, pedestrians, and the public at large, when considering how we impose risk through the deployment of AVs. Empirical work is needed to supplement the elements of conceptual work that rely on evidence to motivate one conclusion or another. And decision-making tools are required to interpret both into schemes that engineers and other specialists can apply to their work, without requiring a PhD in some other field (or multiple fields).

Larger Context

A tension proceeded as the workshop went on, between participants who favored discussion of ethics and navigation/risk management algorithms used to pilot AVs, and those who favored discussion of the wider context of AVs in society. This tension is reflected in the current ethics literature, which while overwhelming favoring the former acknowledges the latter as in need of urgent debate. The workshop brought to light some of these emerging issues, including the impact of AVs on congestion, on labor rights, and on cultural sentiments.

The need for larger debate, however, entails a need for a broader set of participants at AVS and elsewhere. In particular, ethical issues pertaining to broad social effects are typically the domain of political philosophers, who consider the basic structure of social institutions as part of their work. The policy conversation around these issues, moreover, requires policy*makers* from agencies such as the Department of Labor, Department of Commerce, and others. Expanding the sphere of concern around AVs requires inclusion of a greater representation of stakeholders. This carries logistical burdens, but would also create new opportunities and benefits to design AVs for society—and better design society for AVs.

Resources and Funding

As a final note, considerable diversity was added to the breakout session over previous years, with the inclusion of emerging scholars in the field of applied ethics, and novel methodologies transplanted from other disciplines into the subject of AV ethics. Continuing this trajectory was judged to be desirable by attendees of the workshop. However, without continued funding, the cost of producing these breakout

sessions—while modest in absolute terms—is very difficult in relative terms. This signals the need for greater funding of AV ethics that seeks to develop concrete solutions for OEMs and policymakers.

4 Conclusions

This chapter outlined a series of works produced as part of the 2019 Autonomous Vehicles Symposium, and reflections on the private discussion within that breakout session. The session included varied presentations on the ethics of AV decision algorithms, their relation to other technologies, and their broader implications. Participants, including presenters, were drawn from private industry, academia, and government; and from early career and established practitioners. Observed points of deliberation concerned: the need for "before the crash" ethical algorithmic decision-making; concrete and robust ethical theories of AV action; a focus on the larger context for AVs; and better resources and financing for the development of ethics and AVs.

This provides a program for further work for a variety of stakeholders. For researchers and practitioners engaged with AV ethics, it provides two central calls to action to better develop a range of ethical decision-making tools, including those that anticipate ethical dilemmas; and motivate a more serious treatment of ethics beyond canvassing issues. For practitioners and policymakers, it provides a guide to expand the sphere of the conversation around the ethics of AVs. And for funders—private or public—it invites the creation of streams of funding for interdisciplinary research into the ethics of AVs that favors collaboration between empirical and conceptual researchers, and decision-makers.

Acknowledgement. Many thanks to Morgan Avera and Pamela Robinson for their work in organizing the workshop and taking minutes during the proceedings, on which this report is based.

References

Awad, E., Dsouza, S., Kim, R., et al.: The moral machine experiment. Nature **563**, 59–64 (2018). https://doi.org/10.1038/s41586-018-0637-6

Bonnefon, J.F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. Science **352**, 1573–1576 (2016). https://doi.org/10.1126/science.aaf2654

Goodall, N.: Ethical decision making during automated vehicle crashes. Transp. Res. Rec. J. Transp. Res. Board **2424**, 58–65 (2014a). https://doi.org/10.3141/2424-07

Goodall, N.J.: Can you program ethics into a self-driving car? IEEE Spect. **53**, 28–58 (2016a). https://doi.org/10.1109/mspec.2016.7473149

Goodall, N.J.: Away from trolley problems and toward risk management. Appl. Artif. Intell. **30**, 810–821 (2016b). https://doi.org/10.1080/08839514.2016.1229922

Goodall, N.J.: Machine ethics and automated vehiclesvehicles. In: Meyer, G., Beiker, S. (eds.) Road Vehicle Automation, pp. 93–102. Springer, Dordecht (2014b)

Goodall, N.J., Santoni Di Sio, F., Mecacci, G., et al.: Ethical and Social Implications of Autonomous Vehicles (2018)

Jiang, L., Xie, Y., Chen, D., Li, T., Evans, N.G.: Dampen the stop-and-go traffic with connected and automated vehicles – a deep reinforcement learning approach. arXiv:2005.08245 [Cs, Eess], 17 May 2020

Lin, P.: Why Ethics Matters for Autonomous Cars. In: Maurer, M., Gerdes, J., Lenz, B., Winner, H. (eds.) Autonomes Fahren, pp. 69–85. Springer, Heidelberg (2015)