BOOK REVIEW

Autonomous Vehicle Ethics: The Trolley Problem and Beyond, edited by Ryan Jenkins, David Černý and Tomáš Hříbek (2022) Oxford University Press, New York, 496 pp., £29.99 (hardback), ISBN 978-0-19-7639191

While the 'trolley problem', of which more later, features in the title of this book, the editors note that this publication is a 'substantial and purposeful effort' (p.xxi) to look beyond the confines of the conundrum that is the trolley problem. What challenges do autonomous vehicles (AVs) and their advocates face as they are deployed in the modern world? They encounter a social space populated by people, vehicles (some autonomous, some not), awash with unpredictable behaviours and complex, demanding and dynamic, physical infrastructures. The contents of this book will resonate with those interested in technologies beyond AVs as many of the issues raised by the contributors are along similar lines to concerns in artificial intelligence (AIs). AVs and AI are not at the opposite ends of a technological spectrum, but AVs represent a very physical manifestation of technologies whereas AI seems obscure and ethereal to many people. Both AVs and AI are conceptually, and sometimes physically, entwined with the Internet of Things (IoT). One might cite certain commonalities to create a typology of technologies, but the dynamic nature of these technologies and their applications mean that attempts to define them in any absolute manner will almost certainly be problematic and have a short shelf life. Such an observation might offend the sensibilities of engineers and scientists, but for perhaps most of the population AVs, AI and other complex technologies might just as well be magical. As Arthur C Clarke noted, 'Any sufficiently advanced technology is indistinguishable from magic' (1973, p.39).

Just because a subject may be challenging and complex does not mean one cannot hold a positive perspective and we are granted insights into the stance of the editors when we learn that Ryan Jenkins wants to 'bend the arc of technological progress to minimise human suffering' (p.v) that results from automobile accident injuries and deaths, and Tomáš Hříbek 'dedicates the book to all those who are tired of being drivers and hope to be liberated by AV technology' (p.v). Minimizing human suffering is an idea that runs throughout the book and one that predates our current technological level. The concept of liberating drivers is worth noting since automobiles were often seen as opportunities for personal freedom and liberation from social constraints in the early days of their adoption. Human-driven vehicles offered private social spaces secure from the surveillance of others. Yet, like many technologies, they are now seen (by some at least) as restricting liberty. Drivers imprisoned in the routine of commuting might agree.

Simply listing all the chapter titles would be of low value, but it is worth noting that the book is split into four main parts: Part 1 Autonomous vehicles and trolley problems, Part 2 Ethical issues beyond the trolley problem, Part 3 Perspectives from political philosophy and Part 4 Autonomous vehicle technology in the city. Each of these sections is introduced by one of the editors and offers a succinct signposting of the main points to be covered in that part of the book. The order of reading of the individual chapters should be guided by the reader's own level of expertise and interest in specific topics.

This review will include several examples of the works included in the volume in order to give a sense of the style and focus of the material. Given there are 40 authors contributing to 26 chapters, these will be indicative of what readers might expect from the book. The editors wanted to produce a multi-disciplinary, globalized, examination of autonomous vehicles; the range of authors and their disciplinary backgrounds reflect this aspiration. Good to see an approach, reflected in comments from the editors, that notes the value of such thought experiments as the trolley

problem, but which recognize that these autonomous entities are already operating in complex socially constructed environments. Some of these places will be societies where equality of human life is not taken for granted. In other words, whatever the merits of arguing that all lives are equal, this is simply not the case in practice. Some individuals and groups are favoured over others in many societies, and this has implications for autonomous vehicles operating in the real world. This book focuses on vehicles, but the core concerns could be expanded to consider how we might integrate drones, robots and other categories of mobile automated systems yet to be developed.

Before we go any further, we should address the trolley-related elephant in the room (the trolleyphant*?*). This deceptively simple scenario has become an enduring example of an ethical dilemma; one of the few to have crossed over into popular discourse, even featuring as a narrative device in mainstream movies. There are variations on the trolley problem, and these are covered in some detail in the book. There is also a substantial body of literature 'trolleyology' (p.xix). The trolley is a runaway tram or trolley car, and the dilemma sees a person having to allow the deaths of five people or to actively kill one. Do nothing and the runaway trolley will hit and kill five people: or intervene by diverting the trolley to another spur of the line and killing one person but saving five (and it is typically five). The trolley problem is a great catalyst for discussions of active (divert and kill one person) or passive harms (do nothing and kill five people). Many variations on the original idea have emerged, but central to each is a simple ethical question. Variations of the original problem include dilemmas potentially faced by doctors, where a patient's life might be sacrificed to save the lives of many others, a useful route into discussions about scarce medical resources.

From which intellectual trolley depot did the trolley problem emerge? It is widely recognized that the credit should go to Philippa Foot, who looked at decision dilemmas, especially ethical aspects of abortion, and what is called the 'doctrine of double effect'. Foot's role is acknowledged in a paper by Judith Jarvis Thompson (1976), worth reading for both its originality and clarity. Interesting to note that the early pioneers of this approach to engaging with ethics were both women. Most of the contributors to this collection are men.

Part 1: Autonomous vehicles and trolley problems

Part 1 is introduced by David Černý, who makes clear that the contributing authors, including Černý himself, will move beyond abstract conceptualizations of automated vehicular ethics to considerations of social factors, including discrimination and the differentiation of sections of the population. The first chapter is titled 'Ethics and risk distribution for autonomous vehicles' and is written by Nichols G. Evans. He begins by asking, 'Autonomous vehicles (AVs) will be on our roads soon. How should they be programmed to behave?' (p.7). AVs are already here. Their exact numbers and nature are open to discussion, partly because there are several industry definitions around autonomy, there are different degrees of autonomy, and there are debates as to which vehicles have achieved them. Considerable uncertainty surrounds the claims of car maker Tesla, but these are in some ways a distraction from the fact that many other manufacturers are achieving high levels of autonomy in large-scale trials on public roads. Volvo is working on autonomous trucks for the long-haul market while Starship Technologies specializes in small delivery robots. Crucially, this chapter notes the importance of discussions on the risks posed by AVs, some of which might be informed by ethical thought experiments.

Why should we want to introduce AVs into our traffic systems? At the beginning of Chapter 2, co-editor David Černý notes that we need to justify their introduction and can do so if we think in terms of positive factors and negative factors. Where positive factors prevail over negative ones, the introduction is justified. This is like the concept of net benefits often espoused by economists. Whether expressed as factors or benefits, the notion of something being beneficial, on balance, is one recognized in several disciplinary areas. Such a concept can act as a useful reminder why we collectively embrace practices and technologies that are problematic – on balance, we are still collectively and individually beneficiaries of technological advancements.

There may be some engineering teams that include ethicists, but some (many?) will not, and these teams will take an approach that is routed in empiricism rather than ethics and they will be analysing large sets of quantitative data derived from pilot projects. The question of whether engineering teams should include ethicists, other philosophers and social scientists is raised in Chapter 4. Some of the discussions in this book, while focused on AVs, apply equally to discussions around AI. It is worth asking what we should compare AVs against in terms of risk. Mainstream media articles about AVs tend to focus on the failures of AVs in real traffic systems. They rarely report the success of AVs completing many miles without incident (it is not good copy). Do AVs need to be as good, or even better than, the average human driver? Might we see benefits even where AVs are not as good as the average driver? I am deliberately repeating the word average as a discussion needs to be had about typical drivers in different parts of the world, at different times of the day etc., rather than about some abstract conceptualization of the average driver. There are incomplete but nonetheless useful datasets available on driver performance and risk level; some of the best will be held by insurance companies with arrays of new datasets emerging from the data being gathered by engineering teams. Maybe we do not need AVs to be as good as the best subsets of drivers before they will offer clear net benefits to society? Once AVs are as good as the poorer subsets of drivers, there will be a strong social and commercial case for adopting AVs at a much greater scale than at present.

How AVs classify objects may not seem thrilling, but it is vitally important, and this is brought home in the opening of Geoff Keeling's chapter where he describes the death of Elaine Herzberg, the first official fatality involving an autonomous vehicle. It is also interesting to note that Uber (whose car it was) was not held liable, but the safety driver (a human back-up to the computerized system) was. By way of context, the official death toll on American roads in 2018 was 36,560 along with recorded injuries to a further 2,710,000 people (National Center for Statistics and Analysis, 2020).

Chapter 4 introduces the fictitious Annabelle as a member of an engineering team who is aware that ethics may have a role to play in designing outcomes for accident scenarios. Accidents will inevitably happen at some point in the pursuance of AVs and possibly in the life cycle of individual AVs. The authors, Jeff Behrends and John Basl, point out that engineers currently have a certain degree of autonomy (how apt), but may have less in the future as policymakers take more of an interest in the subject. That very autonomy brings with it responsibility. The chapter goes on to explore the interplay of algorithms, machine learning and the trolley problem in the context of AVs. The authors offer a useful critique of both technological and philosophical approaches to the topic. Chapter 5 takes a different approach and offers evidence from research conducted by Akira Inoue *et al*. to ascertain whether people adhere to the expected norms of behaviour when faced with the trolley problem. Does their behaviour change if they are being observed?

Chapter 6 brings in a non-Western perspective as Hongladarom and Novotný look at autonomous vehicles and driving schools. Can we bring in guiding principles from for example Buddhism to help AVs lead a good life? Around the world there are different driving styles, and different expectations from other agents such as pedestrians (think about attitudes to jaywalking). What is to happen when AVs become commonplace and the balance of power with other actors changes? Chapter 7 considers autonomous vehicles and normative pluralism. Saul Smilansky asks questions not just of the AVs themselves but of our attitudes to different vehicle types, ownership models and transportation infrastructure.

Derek Leben discusses discrimination in algorithmic trolley problems and how we can discriminate among classes of things; for example, cardboard boxes from citizens. He goes on to consider whether we would be morally right to discriminate between different categories of people; between passengers and pedestrians, for instance. H notes that autonomous vehicle navigation and collision rules of operation (algorithms) will have to come within the purview of the criminal justice system.

Part 2: Ethical issues beyond the trolley problem

Ryan Jenkins reiterates the desire of the editors to move beyond the trolley problem, as valuable as it may be, into the complexities likely to arise with the large-scale integrations of AVs into human society. Technologies are released not into a vacuum but a thick atmospheric soup, where chemical compounds are replaced by social ones – cultural, legal, psychological, socio-economic. Jenkins takes us on a safari of the section offering snapshots of its chapters. I should stress the value of reading the introduction to each of the parts of the book. These are well done and will aid readers in prioritizing chapters.

Part 3: Perspectives from political philosophy

With an introduction by Tomáš Hříbek, this part of the book widens the focus to more broadly defined political perspectives. Authors consider what we might mean by fairness and the distribution of risks and benefits of AVs as they become embedded in the economy. Hříbek deftly details the key points covered. Of particular interest to me are two chapters focused on the long-haul freight industry and the potential social impacts of AVs. This is an area I have researched myself so I was keen to read their take on the topic. Dubljević and Bauer examine ethical considerations of large-scale future adoption of AVs through the frame of Rawls's theory of justice and future instances of economic injustices. What might this mean in practice? You might choose not to make universal use of AVs, reserving them for long haul and keeping drivers for shorter parts of journeys that start or end in urban centres. This would mean the full unfairness of unemployment would not fall upon truckers. This interested me as I had foreseen a similar approach for the mid-term where AVs would complete long, relatively unchallenging, road trips, and human drivers would handle the more complex urban environments, much as highly skilled ship captains make use of pilots for specialist local knowledge when entering and departing ports. Dubljević and Bauer also note the potential impact on the social infrastructure of human truckers – truck stops and service stations. Many people are employed to feed, fuel and refresh drivers or their trucks, but automated haulage trucks require little more than fuel. Yet truck stops can also be focal points for crime, so negative social activities might be disrupted or displaced.

Clearly there are major social challenges that will arise from the adoption of AVs. AVs are a good example of what Joseph Schumpeter (1954) refers to as creative destruction. When writing a previous review of a book about drone technology (Wane, 2022), I cited Schumpeter because there are certain technologies that have the potential to be creatively destructive. The internet, telephone and automobile are good examples, as are personal computers, television and satellites. All these are now familiar and uncritically accepted by most people. Drones, robots, artificial intelligence and the subject of this book, autonomous vehicles, are constituents of a new creatively destructive technological wave. Further, as futurologist Roy Amara noted, 'we tend to overestimate the effect of technology in the short run and underestimate the effect in the long run' (see Lin, 2024). The wave may take some time to arrive, but arrive it will.

Part 4: Autonomous vehicle technology in the city

Each of the individual chapters focuses on a particular issue and the cumulative effect is a section that develops the dialogue around AVs and how they might form part of future solutions to current problems. That is possibly an over-simplification, but the value of this collection within the volume is that it opens up routes for reflections on the nature of transport in general and how transport is integral to the functioning of cities. Given most of the world's population now lives in cities, transport infrastructure and the challenges of optimizing transportation modes between cities is a worthy inclusion. Once again, while the focus is on AVs, the chapters raise wider questions about how we might tackle transport challenges *per se* and which sections of society should pick up the bill. For

example, the challenges of integrating AVs into crowded urban environments are to a large extent the same challenges currently faced by those seeking to integrate drones into crowded airspace. We see struggles to integrate the non-autonomous but innovative; for example, electric and hybrid vehicles into the existing transport environment. In the UK and elsewhere, we see the benefits of reduced pollution along major routes and at key nodes (such as city centres), but local authorities then face lost revenue as vehicles exempt from charges make up an increasing proportion of the traffic. Social engineering prompts may drive the adoption of physical engineering innovation, but there may be both intended and unintended socio-economic outcomes. A good example of this is the Ultra Low Emissions Zone (ULEZ) operating in London, which has proven very successful in reducing vehicle numbers. However, there are opponents to the expansion of the scheme on both social (fairness) and economic grounds. Borenstein, Bucher and Herkert note that vehicle numbers rebounded somewhat as drivers switched to alternatively powered vehicles or made more use of ride-sharing services, such as those offered by Uber and Lyft. Are people gaming the system or doing what was asked of them only to find solutions within the (new) rules that suit them? Critics of ULEZ-style schemes also note the disproportionate burden of fixed charges – now £12.50 per day in London (Transport for London, 2023) – falling on low-income households as they drive legacy vehicles for longer.

I recommend this book to anyone interested in AVs and ethical aspects of technology. The core focus is on vehicles that have autonomy and the resulting societal implications, but there are observations and reflections here for anyone wanting to engage more widely with the ethical aspects of new technologies, whether drones and robots or AI. The editors and many contributors have created an excellent academic vehicle for a journey into the topic of autonomous vehicle ethics. The book promises a lot and, like one of the many autonomous delivery vehicles now deployed in the real world, it quietly and efficiently delivers.

References

Clarke, A. (1962/1973) *Profiles of the Future*, Macmillan, London, available at https://openlibrary.org/books/OL10487554M/Profiles_of_the_Future (accessed March 2025).

Lin, P. (2024) ‘Amara’s Law and its place in the future of tech’, *IEEE Computer Society*, available at https://www.computer.org/publications/tech-news/trends/amaras-law-and-tech-future (accessed March 2025).

National Center for Statistics and Analysis (2020) *Summary of Motor Vehicle Crashes 2018 Data (Traffic Safety Facts Report DOT HS 812 961)*, National Highway Traffic Safety Administration, Washington DC, available at file:///C:/Users/Stuart/Downloads/2018%20SUMMARY%20OF%20MOTOR%20VEHICLE%20TRAFFIC%20CRASHES%20Traffic%20Safety%20Fact%20Sheet.pdf

Schumpeter, J. (1954) *Capitalism, Socialism and Democracy*, Unwin University Books, London.

Thomson, J. J. (1976) ‘Killing, letting die, and the trolley problem’, *The Monist*, 59, 2, pp.204–17, available at http://www.jstor.org/stable/27902416 (accessed March 2025).

Transport for London (2023) *Ultra Low Emission Zone* website at https://tfl.gov.uk/modes/driving/ultra-low-emission-zone/cars?intcmp=52215 (accessed March 2025).

Wane, P. (2022) Review of *The Drone Age: How Drone Technology Will Change War and Peace*, *Prometheus*, 38, 3, pp.365–70.

Philip Wane
School of Social Sciences, Nottingham Trent University
philip.wane@ntu.ac.uk