The ethical apparatus: The material-discursive shaping of ethics, autonomy, and the driverless car.

Faculty of Humanities and Social Sciences
Leuphana University, Lüneburg

Submitted as a requirement for the award of the
title of
Doctor of Philosophy
Dr.phil in Cultural Sciences

Approved thesis by Indira Ganesh
born March 25, 1975 in New Delhi, India

Declaration

I hereby declare that I have neither undertaken nor applied to undertake any other doctoral assessment.

I hereby declare that the thesis entitled *The ethical apparatus: The material-discursive shaping of ethics, autonomy and the driverless car* has not yet been submitted to any other academic, that I have submitted the thesis only as part of this and of no other doctoral assessment, and that I have not previously failed any other doctoral assessments.

I assure that I have written the thesis submitted *The ethical apparatus: The material-discursive shaping of ethics, autonomy and the driverless car* is all my own work and has been produced without any unauthorised assistance. I have not used any aids or material other than that specified. I have referenced all sources used.

Berlin, June 5, 2021.

Funding disclosures

Between 2016-2020, I received funds from the following institutions that directly and indirectly supported this research. This support ranged from short-term consultancy contracts, travel and administrative support to attend conferences, to research fellowship positions. They have all shaped opportunities to pursue different parts of this work. These include: Association for Computing Machinery (ACM); Berggruen Institute, Los Angeles; Cornell University; Daimler AG; Data Justice Lab at the School of Journalism, Media and Culture at Cardiff University, Wales; MECS Institute for Advanced Study, Leuphana University, Lüneburg; Milieux Centre for Art, Technology and Media at Concordia University; Nachwuchsförderfonds, Graduate School, Leuphana University, Lüneburg; Rockefeller Foundation; Transmediale Festival for Digital Arts and Culture, Berlin; University of California, Irvine; Volkswagen Stiftung's 'AI and the Society of the Future' grant to the Media Theory department at the Hochschule für Gestaltung (HfG) Karlsruhe.

Parts of this work published elsewhere

1. Ganesh, M.I. (2020) The ironies of autonomy. *Nature Humanit Soc Sci Commun* **7,** 157 (2020). https://doi.org/10.1057/s41599-020-00646-0
2. Ganesh M.I., 2019. Tipping the scale: Notes on the topologies of big data platforms. In *Platform Politick* for the Policy Frameworks for Digital Platforms: From Openness to Inclusion project. Anita Gurumurthy, Deepti Bharthur and Nandini Chami (Eds). IT for Change. Bangalore.
3. Ganesh M.I. (2018) A-Words: Accountability, Automation, Agency, AI in *The State of Responsible IoT Report 2018* P Bihr, S Höher and M Krüger (Eds).
4. Ganesh M.I. (2017) Entanglement:Machine Learning and Human Ethics in Driverless Car Crashes In *APRJA:Machine Research,* Vol 6, Issue 1, 2017. CU Andersen & G Cox (Eds). ISSN 2245-7755
5. Ganesh, M.I. (2016) 'Entering the factory', *The Society Pages: Cyborgology*. https://thesocietypages.org/cyborgology/2016/10/28/entering-the-factory/ (October 28, 2016)

Acknowledgments

I read an interview with the French philosopher, Catherine Malabou, in which she narrates an amusing anecdote involving Michel Foucault. Apparently, people would say, "Oh, Foucault, one day you're working on that, the following day you're working on something else, so what is the unity of your work?" And apparently, Foucault responded, "the object of my work is my own transformation." I started this project at a time of transition and intellectual restlessness; I was unsure of how to change the direction of the path I was on. Things were not adding up. But, when I stopped adding and subtracting, I realised that my zigzag and multiform journey, often overflowing, is actually one of life's great gifts and challenges: an abundance of ideas, and opportunities for learning and personal transformation. The following people have made this transformation possible and I am grateful to each of them.

All this started over breakfast with Dr. Nishant Shah on January 1, 2015, at a Portuguese cafe in Hamburg. I had been in Germany for just over six months, having moved from India where Nishant knew my work as a researcher, writer, and activist in the feminist, and digital rights movements. I am very grateful that he suggested I pursue a PhD at Leuphana. This has been both a challenging and rewarding life experience. It has given me what I believe I was seeking at the time: an opportunity to learn and deepen my skills as a researcher and theorist, and to start making a shift towards a different kind of career. Nishant has been a friend, advisor, and collaborator over the years and I am honoured and delighted at our past, present, and future collaborations. Nishant introduced me to Professor Dr. Martin Warnke who has been a patient and thoughtful supervisor right through this. I like to tell people that Martin always has a way of seeing right through to the heart of the matter, and giving the most incisive and necessary advice. His insights put me on the right track with this work, and I am very thankful to him for this. Further, I am grateful that I received a MECS Fellowship to pursue this work. It was a much-needed space to think and learn alongside others.

The Centre for Digital Cultures at Leuphana University continues to be a fantastic space for Media Studies scholarship. I am so happy that I could work with Clemens Apprich and Inga Luchs to guest edit the Spheres special issue on 'Spectres of AI'. I believe we pushed some boundaries on the borders between different fields. Thank you to Lisa Conrad, Nelly Pinkrah, and Paula Bialski, for feedback, encouragement, and support over the years.

Professor Dr. Matteo Pasquinelli at the Hochschule für Gestaltung, Karlsruhe, offered me a research associate position at the launch of the Kunstliche Intelligenz und Medientheorie (KIM) research group, funded by the Volkswagen Stiftung. It was a valuable experience that has shaped many of the ideas here; Working with Matteo taught me a lot about looking at knowledge-making in and by AI's technologies. Big thank yous to both Matteo and to Ariana Dongus for the time we spent working and learning together.

Professor Dr. Orit Halpern at Concordia University, Montreal, has been a supportive mentor, giving me opportunities to develop my research through inclusion in her own work and pedagogical practice. Orit's keen historical research is a touchstone for scholars of computing today; she shows us how to reconcile with the legacies of the past. As a teacher, she puts in so much extra labour for her students. The Planetary Futures Summer School (PFSS) 2017 was an introduction to a community of scholars that I have been lucky to learn from and befriend: Blair Bainbridge, Daniel Rourke, Garrett Lockhart, Harshvardhan Bhat, Hisham Awad, Katherine Buse, Max Symuleski, Nadia Christidi,.Thank you especially to Nadia for hosting me and taking care of me in Cambridge, MA on a research trip. PFSS provoked me to engage with literature and concerns about the speculative futures of AI and autonomous technologies in terms of the histories and futures of the planet, and the humans and nonhumans on it.

Professor Dr. Jack Stilgoe at University College London runs the ‘Driverless Futures?’ research project, which was a really nice group of autonomous vehicles researchers that I was happy to engage with and learn from. I attended one of their workshops in 2019 that pushed me to develop an entire chapter here, (‘The Ironies of Autonomy’) and prepare it for Jack’s forthcoming edited special issue for a prestigious journal. Jack’s support, editing, and insights have helped shape some of this work; his own STS scholarship continues to inspire and impress.

Senior researchers and experts in the US, Canada, and Germany, have been so generous with their time and guidance: Brandon Schoettle, Cansu Canca, Jason Millar, Judith Simon, Patrick Lin, Pak Hang Wong, Sven Beiker, and others who wish not to be named. Many of them have informed this work and are cited here. Thank you to everyone who took time to talk to me about my research. I am especially grateful to Alexander Mankowsky and Sabine Engelhardt from the Daimler Futures Research group for including me in their research on the future of autonomous driving. Their blend of cultural, technical, and design research practice in shaping the future of autonomous driving has been a remarkable process to have some insight into.

A group of academic friends, who are also inspiring scholars, have generously offered guidance, ideas, opportunities to present and workshop this thesis, and edits. I have learned so much from them in developing this work; and about being an academic. I am thinking of: Ahmed Ansari, Alex Hanna; Becky Faith, Daphne Dragona, Georgina Voss, Jeremy Packer, Johannes Bruder, Laura Forlano, Linnet Taylor, Nishant Shah, Noopur Raval, Pedro Oliveira, Sarah Sharma, Seda Gürses, Tanya Notley, and Wesley Goatley. Special thanks to Tanya and Becky for solidarity from the Before Times.

The Rockefeller Foundation's Bellagio Center Thematic Residency on AI was a key space for me to get some of this work done. It was also where I met inspiring thinkers, artists, and scholars working on the cultural and social . I'm thankful to Rumman Chowdhury and Sherry Wong for opening this door for me. And once inside, I met people there who became friends, colleagues, and peers; thank you Joy Buolamwini, Noah Levenson, Roya Pakzad, Rumman Chowdhury, and Sherry Wong.

Seda Gürses, Zeerak Waseem, and Francien Dechesne have been influential collaborators. They are thoughtful scientists who I have learned so much from about the "ethics of algorithms", and even more about what it means to be critical, hardworking, and politicallyminded scientist. So much of my thinking and learning has been inspired by the questions they have been asking and continue to ask of their own discipline. Special thanks to Seda for her extremely thoughtful and supportive welcome of me and so many voices rarely heard in the shaping of tech.

Two good friends have been my wing-men through this process; Johannes Bruder and Wesley Goatley have had my back in the past few years. They have patiently responded to my small and big questions about the nuts and bolts of a dissertation, and gently pulled me away from my anxieties and self-doubt. And this is aside from the intellectual work of editing texts, talking about ideas, doing research together, and teaching me about how to read and how to learn. It is no surprise that they are both very popular teachers, and I am lucky to have them in my life. A very special thank you to my weekly Zoom writing club, Maria Yablonina and George Voss, for showing up every week and helping me get through this past year under lockdown.

Thank you Stephen Tattum for the editorial support.

Then there are my friends who have kept me grounded, happy, and sane over the years, and especially during the Corona lockdown when much of this was written. Mercifully, many of them are not academics. Thank you: Ally, Dave, Heba, Junko, Lucy, Olu, Rebecca, Justin, Teena, Tuhina, and Varoon. My WCC crew distributed around the world, Kavi, Sarah, Tina, and Rubina, have known me since I was 18 years old and have always supported me in big and small ways. Paulina Bozek, Sarah Malieckal, Srimoyi Mitra, and Tina Jose were especially generous in offering me space and care when I visited the US on research trips. Special thanks to my childhood best friend, Dr. Tuhina Raman, who ended up buying a Tesla even though I did not think it was the best idea. Thanks for the test ride, it was an important research experience that I got to write about

Some collaborations have substantially contributed to this research and my own clarity: Thanks to Jenny Bourne, Nils Gilman, and Tobias Rees, for the Berggruen Institute fellowship; Katrin Klingan, Nick Houde, Johanna Schindler at the Haus der Kulturen der Welt in Berlin; and Tin Geber and the Digital Earth community at Hivos.

It does not matter if I mention my family first or last, because they are always everywhere. My parents Alka and Ganesh, and my sister Gayatri, have made everything possible; all of this. They are my rock, a constant source of strength, love, encouragement, and support in material and immaterial ways. I am happy to dedicate this work to them. And, thank you Laird Brown for your unceasing belief in me especially in those times when I could not believe in myself. We have both come a long way together, and not just from Bangalore to Berlin. Your patience, support, love, and wit have seen us both through this. I reckon no one else will be happier and more proud than these four people that I have actually finished this work, and transformed myself into the one they know I am.

Abstract

This research argues that the emergent driverless car, as a kind of autonomous vehicle, is a Foucault-ian 'ethical apparatus', working as an epistemic device to materially embody and enable discursive power by generating notions of 'autonomy' and 'ethical decision-making'. The ethical implications of AI, algorithmic, and autonomous technologies are topics of current regulatory and academic concern. This concern relates to the lack of meaningful oversight of black boxes inside AI systems, liabilities for manufacturers, and inadequate frameworks to hold AI-based socio-technical systems to account. One recent artefact, the driverless car, has taken on these concerns quite literally in the shaping of a niche discourse of the 'ethics of autonomous driving'. Ambitions to produce a fully autonomous vehicle based on AI technologies are constrained by speculative concerns that its decision-making in unexpected accident situations cannot be assumed to protect humans. 'The ethics of autonomous driving' evaluates proposals to build 'ethical machines' by examining the relationship between structures of human values and moral decision-making, and how they comport to computational architectures for decision-making. This is the first case this work takes up, chiefly organised around an analysis of a thought experiment, the Trolley Problem, and the online game, Moral Machine, that crowdsourced values to suggest approaches to an 'ethics of autonomous driving'. Rather than evaluate the feasibility or appropriateness of these two approaches, this work attends to the more critical issue that ethics is being proposed in terms of technologies turning on the logics of risk, speculation, and probabilistic correlations that are fundamental to how machine learning makes decisions. The concern in this work is less a normative framework or approach for a *better* or more appropriate ethics of autonomous driving. Rather,

this work argues that what we understand as ‘the ethical’ is being transforming when architected by, through, and for AI/autonomous technologies to become their own regulators. Hence the production of autonomous driving necessitates computational infrastructures that are creating a world legible to and for the navigation of a driverless car. I argue that this is fostering computational governance that has implications for human bodies and social relations, chiefly that conventional approaches to regulation and accountability attend to human values and decision-making rather than computational ones. A second case that this research examines is that of driverless car crashes, to examine how ‘autonomous’ driving requires substantial embodied human knowledge and micro-work. Taken together, these two cases - the ethics of autonomous driving, and crashes - make an argument for how myriad practices of knowledge-production are translating the human world into something legible to the navigational needs of the *car;* producing changes in the human world through the actions of the car on that basis, and advancing notions of ‘autonomy’. This work concludes with arguments for a critical reconceptualisation of ethics and ethical decision-making in AI/autonomous systems.

Abbreviations and terms used

AI	Artificial intelligence
AV	Autonomous vehicles
AS	Autonomous systems
DC	Driverless car
FATML	Fairness Accountability and Transparency in Machine Learning
FRT	Facial recognition technologies
IEEE	Institute for Electronics and Electrical Engineers
LAWS	Lethal autonomous weapons systems
ML	Machine Learning
NTSB	National Transport Safety Board
NHTSA	National Highway Transport Safety Authority
SAE	Society of Automotive Engineers

List of images

CHAPTER 1. AUTO-CORRECT

Introduction

> Decision-making by automated systems will produce new relations of power for which we have as yet inadequate legal frameworks or modes of political resistance—and, perhaps even more importantly, insufficient collective understanding as to how such decisions will actually be made and upon what grounds. Further, demands for public accountability will require much greater participation in the epistemological frameworks that organize and manage these new techno-social systems, and that may be a formidable challenge for all of us (Schuppli, 2014, p. 7).

This work asks, How is the material and discursive shaping of ethical decision-making and autonomous driving generating computational knowledge about the world? What are the implications of this knowledge-making for human societies?

This work is a critical cultural analysis of the making of meanings and knowledge about a future technology, fully autonomous driving, that does not yet exist yet. I argue that this "socio-technical imaginary" (Jasanoff and Kim, 2015) is a techno-solutionist response to the problem of a lack of road safety owing to poor human driving. I argue that in the emergence of this speculative new technology, a demonstration of concern for safety is generating computational infrastructures of accountability and decision-making in modes and on terms that are legible, eventually, only to these computational systems. This is taking place through a discursive construction called the 'ethics of autonomous driving'. The concern for ethical decision-making, which we might think of as a personal or collective human practice of reasoning, is being rendered as computation; but, it is transforming what we understand as the 'ethical' into forms of computational, algorithmic, and statistical knowledge-making, and governance, of human social relations, spaces, and bodies. This is concerning because policy and regulatory narratives about the regulation of technologies that do not exist (yet) will follow from the language, narratives, and imaginaries being laid down now, and may not, as I show here, fully confront how algorithmic and AI technologies interact in and shape the world. The ethics of autonomous driving is concerned with 'alignment' between the architectures and processes of computational decision-making, and human values It is organised around crashes imagined to be legible to computer models; but, when crashes have occurred, 'ethics' defer to conventional accident accountability approaches in which the human operator is liable and responsible. Contemporaneously, the discourse of the ethics of autonomous driving emerges in relation to another discursive construction: 'AI Ethics'. AI Ethics is, a tech industry-led constellation of, variously, performative, academic, and industrial-cultural correctives to minimising harms associated with present and future AI technologies. AI Ethics signals concerns for safety, while enabling self-regulation with minimal external scrutiny through non-stop industrial R&D. The ethics of autonomous driving has become a capsule case study in the popular and expert discussions of 'AI Ethics'.

Autonomous driving as a techno-solutionist response to poor human driving is not new; it has been imagined by engineers since the 1930s; the designer Norman Bel Geddes[1], describing how he thought driving would change said, “Everything will be designed by engineering, not by legislation, not in piecemeal fashion, but as a complete job. The two, the car and the road, are both essential to the realisation of automatic safety” (quoted in Seilor, 2008, Loc 1709). Similarly, in a 1975 oral history interview, the engineer Vladimir Zworykin explained his motivation for building the ‘autonomous highway’ in the late 1950s: “This growing number of automobiles and people killed in accidents meant something should be done. My idea was that control of automobiles should be done by the road.” (in Ackerman 2016, np) Autonomous driving has been in technical development by DARPA and US university researchers since 2004 (DARPA, 2007); and by Daimler in Germany since the 1980s.[2] However, ‘autonomy’ in the sense of ‘self-driving’ does not exist, it is a popular imaginary, closely connected to AI, automatons, and robots.

While we have autonomous trucks, autonomous deep sea exploration robots, and unmanned aerial vehicles (UAVs), the driverless car (DC) or ‘self-driving’ car is still speculative. Cars and the open road are symbolic embodiments of independence and autonomy, such as in the shaping of national identity in the United States during the Cold War (Seiler 2008); and in

[1] Norman Bel Geddes was a US designer who is significant in this context because one of his best-remembered designs was General Motors-sponsored Futurama exhibit at the New York World's Fair (1939–40). ‘Futurama’ was important because it was a master plan for the United States twenty years into the future (1959-1960) imagined in terms of automobility and suburban living. Bel Geddes was an important discursive influence on the automobile and cities. Bel Geddes’ developed the idea of “magic motorways” as central to American flourishing and modern social life.

[2] Daimler’s ‘networked mobility’ project, Prometheus, was piloted in 1986. https://media.daimler.com/marsMediaSite/en/instance/ko/The-PROMETHEUS-project-launched-in-1986-Pioneering-autonomous-driving.xhtml?oid=13744534

present-day Saudi Arabia where women struggle for the right to drive. In the Hollywood film *Thelma and Louise* (Ridley Scott, 1991) the female protagonists find that freedom exists only by driving off a cliff; the law, and the disappointment of relationships with men, can otherwise never really be escaped. Sometimes the merging can be macabre, literal. J.G Ballard's novel, *Crash* (1973), drifts from one erotically charged description to another of mangled human bodies and machines fused together in the moment of a car crash. The antagonist, the "TV scientist", Vaughan, and his motley crew of car-crash fetishists seek out crashes in-the-making, even causing them, just for the thrill of it. Vaughan's ultimate fantasy is to die in a head-on collision with the actress Elizabeth Taylor. Media theorist, Marshall McLuhan, refers to the car as an item of clothing (1964/1994, p. 217): "the car has become an article of dress without which we feel uncertain, unclad, and incomplete in the urban compound." Thus the automobile is a significant media artefact of the twentieth century.

The dramatic spectacle of a 'robot' car that is making decisions on its own plays on multiple anxieties, chiefly that of a long-held concern about AI/robot technologies exceeding human control, a popular theme in Science Fiction. There are rich narratives about deviant machines, that might be delightful at first but that then eventually break down, become corrupt, or go rogue, bringing harm to humans. It is no surprise then that the emergence of driverless cars, or any new technology innovation, has to foreground its safety to various publics and stakeholders, from potential future users, to regulators, and auto manufacturers, among others, if they want it to catch on. This has taken shape in terms of a provocative thought experiment called the Trolley Problem that positions a concern for safety and accountability in terms of *ethics:* that to design a fully autonomous vehicle (AV) would have to take into consideration that it might be faced with traffic situations that put multiple human lives at risk. If autonomous driving meant the replacement of the human driver, then the car would have to be imagined as making split second decisions that jeopardise either the people in the car, or the people outside it, and either saving the life of one person, or many people. How is a car to (be designed to) choose?

In this work I find that the Trolley Problem was only ever intended as a provocation for engineers and manufacturers in the design of AVs; but this technology and its attendant discourses have emerged through the significant online, popular, media, policy, and academic presence of the Trolley Problem in a discourse of 'the ethics of autonomous driving'. As such, it has enjoyed some media attention and policy responses[3] as a framework for advancing what AV safety would look like. As I will show, this is 'discourse' because it is "both an expression and a constitutional prerequisite of the social; become [ing] real through the actions of social actors, supply[ing] specific knowledge claims…[they] crystallize and constitute themes in a particular form as social interpretation and action issues" (Keller, 2011, p. 52). And while discourse has a sociological dimension as Keller lays out, it is also linguistic and representational, setting down the terms through which we refer to something, and often deploying epistemology. This work investigates how the ethics of autonomous driving is an intentionally, materially shaped discourse necessitating technologies for ethical decision-

[3] In 2017, the German Federal Ministry for Transport and Digital Infrastructure (BMVI) was one of the first to release a document about the ethics of autonomous driving, which reiterates a commitment to valuing human lives. https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html Many other countries have their own ethics of autonomous driving policies and reports

making by future autonomous vehicles. As such, this is already re-shaping the interaction of AI/autonomous technologies in relation to the human, social world, but in ways that we struggle to control or understand. The ethics of autonomous driving takes a machine ethics approach to architecting rules sets that will enable automated ethical decision-making. Also known as ‘value alignment’, machine ethics intends to create ‘ethical machines’ that will adhere to existing human value systems. This raises a number of concerns about how exactly these value systems will be identified to be widely applicable to the largest number of people, and how they will inform computational development. In this work I will take a detailed look at machine ethics approaches, and the Trolley Problem in the context of this.

I explicitly connect the concerns associated with the Trolley Problem and machine ethics with how popular works of fiction portrayed robots and AI making decisions that bring harm to humans. For example, Isaac Asimov’s 1942 short story collection, *Runaround*[4] features the Three Laws of Robotics that are popular in the discussion of autonomous driving ethics. These laws are a set of rules governing robot-human relations. First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm. Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. There is also a 'zeroth' law preceding the first but was added later; this relates to 'humanity' rather than the singular human in the first law. In the early days of this research, it was difficult to come across a mainstream media or academic article that did *not* mention the Laws of Robotics. And while this is fic-

[4] Source: https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

tion, much of how we understand AI is in fact through metaphor, fiction, speculative figurations, and imaginaries because we use metaphor to constitute the uncertain and the unfamiliar. (Gilman and Ganesh, 2020)[5]

While the Trolley Problem and its proponents might have achieved mainstream media and TED-talk visibility, academics and researchers are coming up with serious proposals for machine ethics. The most well-known of these is the Moral Machine[6], a project from MIT Media Lab. This is an online 'serious game' that invited the general public from around the world to respond to 13 scenarios loosely modelled on the original provocation of the Trolley Problem. The Moral Machine has assembled a large global dataset of moral values about how a driverless car should prioritise different human and nonhuman lives in the case of an unexpected accident. What is crucial about the Moral Machine is that it shifts the framing of ethical decision-making from one of value alignment rules sets, to statistically modelled risk assessments. I argue that the Moral Machine project presages a form of computational governance turning on the logics of risk, speculation, and probabilistic correlations that we equate with machine learning-based technologies, many of which already constitute autonomous driving-in-development. Central to this is a concern with mapping how our understanding of the ethical itself is shifting in this process. Hence, this work does *not* follow the normative formulation: 'What is the ethics we need to regulate the development of this technology?' Nor do I evaluate individual ethical decision-making proposals from the machine ethics tradition. In-

[5] There is an extensive Wikipedia page dedicated to autonomous vehicles in science fiction cinema, television and literature: https://en.wikipedia.org/wiki/Self-driving_car#In_fiction

[6] https://www.moralmachine.net/

stead, I am arguing that in bringing autonomous driving into the world, proposals for computational ethics as governance of the car and by the car are becoming computational governance of the world around and outside the car. This is because ethical decision-making is being wrought in computational terms only accessible to the computational technologies that enable autonomous driving. As I will show, this has implications for human social relations, bodies, and knowledge.

While the Moral Machine has been published in *Nature* and *Science*, and featured in leading newspapers around the world, the adaptation of the Trolley Problem in the autonomous driving context has been widely challenged, and then largely ignored across various stakeholder groups and industrial communities.[7] And, driverless car crashes have occurred but not along the lines imagined by the ethics of autonomous driving discourse. Crashes are critical moments from which to unpack the social and discursive constructions of AV 'autonomy' shaping human subjectivity in the big data infrastructures[8] that sustain driving. While autonomous driving promotes the erasure or replacement of the error-prone human driver, I find that autonomous driving is in fact deeply socio-technical, and constituted by human and machines working together, as in the case of auto-pilot. The name of the setting, 'auto-pilot', evokes airplanes and flight; the legacy of aviation safety design and engineering has a significant influence on our understanding of the place and role of the human in high-end engineering.

[7] In 2018, the World Economic Forum came out publicly saying that this was not the right approach to regulating self-driving cars and should not slow down progress: https://www.weforum.org/agenda/2018/01/why-we-have-the-ethics-of-self-driving-cars-all-wrong/

[8] When I refer to 'big data infrastructure' in this text I am referring to a large, large-scale distributed computational infrastructure of infrastructures and not just to 'Big Data', which is a distinct nomenclature and domain of study.

Auto-pilot is the setting that makes it appear that the car is driving itself; It frees up drivers to not drive and attend to something else. Crash forensics reveal the cars were in auto-pilot mode, but when computer vision systems failed to correctly recognise objects in the environment around the car, control was handed back from auto-pilot to human control. But the inattentive human driver took too long in taking control back. There is what I refer to as an irony of autonomy emerging here: That even though humans are supposed to be poor drivers, and 'freed up' to not drive, humans are in fact still in the pilot/operator's seat, and must be ready to take over when the system fails; and if they do not, face either liability—or death. Further, this irony is compounded by its obfuscation despite being of great commercial value; it is also under-valued.

The failed computer vision in these systems are at base statistical models that rely on a complex global and distributed configuration of computational infrastructures. Within these are "heteromated" (Ekbia and Nardi, 2014) humans, people who perform micro-work within these infrastructures. The crash becomes a moment to identify the constitution of the infrastructures of autonomous driving that rely on embodied, situated, human cognition, and knowledge. It reveals an emergent artefact, the driverless car, that is entirely native to the planetary-scale computational governance that Benjamin Bratton refers to as the 'stack'; the stack is characterised by "oscillation" between configurations of human and machine in spaces that are "real-but-as-yet-unnamed" and "imagined-but-as-yet-not-real" (Bratton, 2015, p. 13). In other words, we are ill equipped to frame these emergent spaces and relationships through conventional approaches to law, policy, governance and 'ethics'. And yet, driverless cars are framed in terms of the language of human-computer interaction and automation from aviation. I argue that the uncertainty and in-computability of the highly automated, distributed, big data-based computational infrastructures of autonomous driving will require, as Schuppli predicts, new approaches to public accountability. I argue that they do not currently exist and machine ethics is not likely to give us the answers. Yet, conventional approaches to regulation continue: Humans are held accountable for crashes, and must, in deeply embodied ways, fall in line with the demands of autonomous driving. For example, surveillance technologies now literally measure and monitor the human driver's attention to prevent inattention that could result in not being able to take over from a confused computer vision system. Thus, I argue that the constitution of autonomous driving is significantly human.

Emergent social-technological systems like AI and autonomous driving technologies are imbued with faith, fetishization, ideology, and hence eventually power, which we must critically evaluate (Chun, 2011, pp 17-18) given the transformations that I have argued are underway. This research is organised around two inter-related cases: the ethics of autonomous driving discourse, and the constitution and shaping of autonomous driving through crashes. I develop an analytical frame of a Foucauldian-inspired *ethical apparatus* to map the social, cultural, and computational emergence of autonomous driving. The ethical apparatus is valuable and urgent for it allows me to look beyond the artefact of the autonomous vehicle, which, in discourses of ethical decision-making tend to be the focus of autonomous driving. In expanding my perspective to the world around the car, I am challenging ethics and autonomy as housed in the computational systems of AI technologies, and as being shaped through and in society and culture as well. Autonomous driving is not a given; In drawing attention to its social, material, and discursive shaping, I want to emphasise vectors like imaginaries, symbols, institutions, cultural narratives, epistemological practices, and computation that are making and stabilising the meaning of 'autonomous' driving. 'Apparatus' can refer to a *device*, as well as to the Foucauldian *dispositif*, and I engage both these mode to analyse the work of bringing a technology into the world. Devices are epistemological instruments, they make knowledge through practices of measurement, which create worlds and universes. But devices as measures are also situated, and subtly establish the world, and the terms on which we know that world; in other words, discursivity. I argue that autonomous driving and ethical decision-making can be studied as devices *and* as the discourses of 'autonomy' and the ethical thereof. Thus this research takes a 'material-discursive' approach, seeking out empirical and material sites of how these ideas are being shaped. To engage these different sites, I frame the emerging AV in terms of its *multivalent ontologies*: As AI/robot, as a big data infrastructure, and as a conventional vehicle. These multivalent ontologies enable a critical inquiry into how the ethical and the autonomous are constituted by many dense layers of history, culture, and society.

On terminology and language

So far, I have used different terminologies and language in laying out the key arguments and sites of this research. I clarify them here because as a text about the shaping of representations and discourse, there are uses of language that I am trying to critically interrogate, but also just need to use them to refer to the matter at hand. Hence, there is a frequent use of single quotation marks, '..', which indicate discursive shaping and handling of a term. Additionally, the use of these quotation marks refers to the inherently representational nature of this entire domain of AI technologies. A common metaphor about 'seeing' in critical studies of AI is that of looking inside the black box, of mapping the 'anatomy of AI' (Crawford and Joler, 2018), or visualising the layers of decision-making in black-boxed machine learning (Pasquinelli and Joler, 2020). The black box metaphor is often presented in terms of the flat notation: input —> black box —> output. But, we also talk about computer models and machine learning in terms of *depth*, and *layers* of neurons; We might as well put air quotes around all these words. Words like depth, layers, and neuron, or even 'network' are only linguistic and visual *representations* of the spatialisation of mathematical relationship; This is not how things *actually are* (Offert and Bell, 2020; Zer-Aviv, 2016).

'Autonomous vehicle' can refer to driverless cars, buses, unmanned aerial vehicles like drones (civilian and military-use drones). In this work I refer to autonomous vehicles (AV) as a general class of technology, and driverless cars (DC) as the specific instance of this class. However, concerns around auto-pilot and automated decision systems cut across different kinds of 'autonomous' systems that have human operated elements, and there is relevant literature from studies of these other applications that I refer to. The acronym A/IS refers to Autonomous and Intelligent Systems. When represented in this way, it refers to the overlapping concerns about the "design, development, deployment, decommissioning, and adoption of autonomous or intelligent software when installed into other software and/or hardware systems that are able to exercise independent reasoning, decision-making, intention forming, and motivating skills according to self-defined principles." (IEEE 2018) In this work I build on this definition using the terms AI/Autonomous Vehicles (or AI/AV)to refer to such overlapping concerns.

There are a variety of words currently in use in popular and tech media, academic and grey literature, policy documents, and everyday speech to refer to the driverless car: Robot car, self-driving car, semi-and fully-autonomous vehicle, driverless car, and the latest of these is, 'connected cars'. None of them are perfect, and this is part of the problem, as I indicate here. Naming new and emergent technologies can be notoriously difficult in any age; the emergence of the word 'automobile' in the late 1800s was the subject of scornful discussion in the popular press at the time (La France, 2016). Using the words 'autonomous' and 'driverless' allow us to imagine that the car is moving 'on its own'; the purpose of this dissertation is to show how these ideas are being enabled, and what the implications of this are.

Lethal autonomous weapon systems (LAWS) refer to weapons systems that include munitions devices, drones, tanks, platforms (eg. a ship or plane capable of selecting targets and firing munitions at those targets on its own), or, broadly, an operational system rather than a discrete vehicle. Literature on LAWS is informative, but I do not analyse any specific LAWS.

AI refers to a constellation of computational technologies such as machine learning, computer vision, natural language processing, machine perception, automated and algorithmic decision-making, and the distributed infrastructures of planetary scale computation such as cloud architectures and distributed human-enabled annotation or moderation. Wherever possible I refer to the technology (eg. machine learning or computer vision) rather than words like intelligence, which are representations, and loaded, historically situated words. AI is entangled with 'agency', 'autonomy', 'accountability' and 'ethics' in current research and popular media. AI is also what powers 'autonomous' driving. All of these words make an appearance in this work. Part of my intention here is to show how some of them come to be entangled and co-constitutive, like autonomy and ethics. All of the others are as well but I do not attend to all of these inter-connections.

In this research I use the words 'machine', 'computer', and 'system' fairly interchangeably to refer to computers as social and cultural artefacts, and that perform computation. I often refer to computational processes as 'machinic'. Further, I refer to machine learning and modelling as algorithmic, or as statistical, as it is also these things.

Ethics are perhaps the hardest to define because it is precisely what is under question and being discursively shaped. However, 'ethics' have many other dimensions and meanings. Ethics can refer to codes or rules for living, as in the case of machine ethics that I discuss at length here. It can also refer to Kantian rules that enjoin the self to universal subjectivity (cited in Foucault, 1994, p. 279-280). Outside of the realm of Philosophy and Ethics, ethics in discussions of machine ethics and computational ethics are loosely understood as 'rules for correct action', broadly understood in a deontological, Kantian tradition. Ethics can also be a relation of and with the self, as in a Foucauldian technology of the self (Op Cit, p. 284). Foucault also distinguishes ethics from morality, which he understands as codes or rules for living that are socially imposed and regulated. On the other hand, Judith Butler prefers to think of ethics in relational terms that begin with a self that is always socially situated and relating to others (Butler, 2005). There are also many other traditions that do not make much of an appearance here: Aristotelian ethics, virtue ethics, professional ethics, applied ethics, feminist ethics, and ethics of care. What is perhaps most important to emphasise is how two particular registers of ethics, that of self-hood, and of rules or regulation, are most dominant in the shaping of AI and autonomous driving discourses. There is a convenient collapse that happens here in shaping autonomy: a *self* that *knows* how to act according to the *rules. '*Autonomy' is also understood differently in everyday speech, governance, policy, law, ethics, and philosophy; And is also the subject of this inquiry. In computer science it usually refers to *"*the ability of a computer to follow a complex algorithm in response to environmental inputs, independently of real-time human input." (Etzioni and Etzioni 2016 p. 149)

In the next sections, I elaborate further on the two cases this research addresses: the ethics of autonomous driving discourse, and the (de)constructions of autonomy through car crashes.

A problem with trolleys

Ethical Things (Simone Rebaudengo and Matthieu Cherubini, 2015) is about the increasing 'smart-ification' of objects[9]; a speculative object in this collection is a portable, electric, swivel-mounted fan with a dashboard with dials, connected to the internet. It sits on a table

[9] http://www.simonerebaudengo.com/project/ethicalthings

between two people. The fan records information entered about the two people to determine which one of them it should turn towards and direct its breeze. Data about the people can be entered in arbitrary terms as far as fans and breeze go: Educational levels, weight, religion, and gender. If it cannot make an assessment based on the information it has, then it sends the question to an Amazon Mechanical Turk worker (a 'Turker'). Turkers decide who the fan should turn towards, and they are expected to offer a short justification for their choice. The results are hilarious and bizarre. For example, in the video that accompanies the project, we see a response from a Turker that the heavier of the two people should be fanned because fat people sweat more. In a personal conversation, one of the designers tells me this was meant as a provocation in response to the growing 'ethics of autonomous driving' discourse: They explicitly *want* the results to appear ridiculous.[10] They write:

> How can such systems be designed to accommodate the complicatedness of moral and ethical thought processes, especially when human lives are involved? Just like choosing the color of a car, ethics can become a commodified feature in autonomous vehicles (AVs) that one can buy, change, and repurchase, depending on personal taste.

In *Ethical Things*, the designers are referring to the thought experiment, the Trolley Problem. The Trolley Problem was intended as a provocation for engineers and manufacturers to consider a future with autonomous vehicles (AV) that will make decisions independently, ie.

[10] Personal communication with Matthieu Cherubini at Dutch Design Week, Eindhoven, October 2016. Cherubini and I were both invited speakers and connected because Cherubini had also started, and then discontinued, a PhD about autonomous driving http://automato.farm/portfolio/ethical_autonomous_vehicles/

'autonomously', and without human intervention. It prompts thinking about the following: the decisions to be made in scenarios that a fully autonomous vehicle's AI software will have to compute in the event of a crash that potentially involves the loss of human life or harm to humans, damage to property, or both (Lin, 2013). The Trolley Problem is a thought experiment from Philosophy that positions consequentialist, or Utilitarian approaches to ethics (in which the outcomes matter more) against Kantian, or deontological ethics (in which the rationale for actions proceeds according to contextual rules and duties). Originally developed by the philosopher Philippa Foot (1967) to provoke discussion about women's rights to abortion, it goes like this:

> Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, Mrs. Foot has arranged that there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. Is it morally permissible for you to turn the trolley? (Jarvis Thompson, 1985, p. 1395)

The Trolley Problem has been adapted to the autonomous driving context and has become shorthand for a public and academic discussion about how to "assess the need for a moral component to automated vehicle decision making during unavoidable crashes, and to identify the most promising strategies from the field of machine ethics for application in road vehicle automation." (Goodall, 2014, p. 3) It was predicted to be one of the first frames for 'artificial morality' because driverless trains and cars would be the first "robot technologies" we would encounter in everyday life (Wallach and Allen, 2009, p. 13-14). The noted AI scientist Gary Marcus writes for the *New Yorker* magazine about the development of driverless cars:

> That moment will be significant not just because it will signal the end of one more human niche, but because it will signal the beginning of another: the era in which it will no longer be optional for machines to have ethical systems. (2012, n.p)

Marcus does not tell us how this will happen, or what kinds of ethics and ethical systems we will have, but in the years since he wrote this, many actors are staking a claim to exactly this. The Trolley Problem has generated a great deal of media attention, aided by expert scholars who have had space to foreground it in media and public imagination, and in academic scholarship.

Machine Ethics

The ethics of autonomous driving is associated with a machine ethics approach which "attempt[s] to duplicate or mimic what in people are classified as ethical decisions," or "the modelling of the reasoning processes people use (or idealized people might use) in reaching ethical conclusions…" (McDermott 2008, p. 2). It relates to the endeavour to make an 'ethical machine' or for machines to 'behave ethically', that is, whose behaviour is aligned with human ethical value systems (Cave et al., 2018). There are two parts to this: *how* to encode values or principles in artificial agents so that they function as desired; and the other is normative, and relates to identifying *which* sets of values or principles are appropriate for this encoding in the first place. Identifying sets of rules to direct correct action is the primary ambition of machine ethics, and in that sense, is a perfect match for driving itself, which is a unique set of formal rules, informal, explicit, and tacit knowledge. But arriving at these rules is difficult because of the sheer enormity of the task, and because so much of driving is tacit, or learned adaptively. Machine ethics approaches are attuned to accountability, to prevent and mitigate errors, by identifying rule sets that must recognise potential unforeseen outcomes of building the technology; that identify particular narrow use cases for the technology, and what situations might be appropriate for humans to resolve but not for machines.[11]

Some recent proposals for identifying such rules for ethics include:

-an ethics checklist for programmers and developers called Deon[12];

- an 'ethics knob' for sliding between different ethics 'settings' (Contissa et al, 2017);
- ethics bots that record personal preferences over time and automatically act on that basis (Etzioni and Etzioni, 2016);

[11] However, there is a need to clarify that not all machines are the same; and a conversational chat bot and an ATM machine, and an automatic coffee machine are all different, and have different kinds of decisions they make that could be construed as being ethical.

[12] https://deon.drivendata.org/

- a system for ranking outcomes of accidents, which could become the basis for programming a driverless car (Bhargava and Kim, 2018);
- a voting system for ethics (Noothigattu et al, 2017; Awad et al, 2018, Kim et al, 2018);
- ’Responsibility Sensitive Safety’ that arrives at its own ground truth for safe autonomous driving by proposing rule-sets on the basis of reverse-engineering how the 37 most common kinds of road accidents occur (MobilEye, 2019).

These are similar to the approach of the “ethical governor” in the context of lethal autonomous weapons systems (LAWS) that works on, roughly 150 “rules of warfare [will be] translated into ethical logic” (Arkin, In Chamayou, 2013/2015 p. 208). Hence, a machine ethics approach is literally the automation of laws of combat as sets of rules. This highly specific approach obviously understands warfare as a controlled space where everyone plays by the rules. Automatic driving adopts a similar approach in the perspective created by machine ethics.

In discussing machine ethics in ethical decision-making I argue that this approach works as a discursive instrument establishing ethics in terms that are legible to the (data) infrastructures of autonomous driving, and hence can be thought of regulatory only within such infrastructures. More critically, in encoding ethics in computational terms as rules sets, what we understand as ‘ethical decision-making’ is subject to the computational operations of these systems. Hence, my concern is less that ethics is reduced to an output or a number; all digital decisions eventually are specific outputs. Rather, *what kinds* of thinking and decision-making are these technologies establishing in proposing that machine-made decisions might be accurate or appropriate?How must conventional regulatory systems, and human moral and ethical decision-making practices, evolve to meet such conditions?

We can sense a glimmer of this in the Moral Machine project created by a (now-defunct) research group at MIT. The Moral Machine is an online game featuring 13 elaborate scenarios that a fully driverless car with failed brakes has to confront. This online game has already been played by millions of visitors from around the world, generating a dataset of 39.6 million responses. These millions of responses are 'ground truth' to build a database of driverless car crash responses to create a statistical model of human values that could be used to direct the action of algorithmic and computational agents, to develop moral decision-making algorithms (Awad et al., Op Cit; (Noothigattu et al, 2017; Bonnefon, Shariff, and Rahwan, 2016, 2019). I believe that the intention signalled by the Moral Machine to shift to statistical approaches to decision-making suggests an epistemic challenge that I take up in this work. What are the implications of a project such as the Moral Machine for how we understand regulation by 'moral decision-making algorithms', or for that matter, any other kind of algorithmic decision-making?

Researchers have argued convincingly why the Trolley Problem is an inappropriate frame for autonomous vehicle ethics (Roff, 2018); and that it is not even an appropriate frame for driverless cars, that the moral issue lies not in the moment of the decision but long before the trolley even reaches the track (Nyholm and Smids, 2016). Media attention to the Trolley

Problem has been largely uncritical and reactive, like associating ethical decision-making with programming an on/off switch, as if on a table-top fan, to make a decision about who the driverless car should kill. I refer to some of these examples later in this work. Eventually, its media popularity has declined since the start of this research work (Cassani Davis, 2016; Bogost, 2018). Yet, the ethics of autonomous driving has sparked research across the fields of computer science, moral and social psychology, behavioural economics, mathematics, and philosophy to identify how to make the kinds of choices it provokes; or at least to argue for a different approach.

I also want to bring attention to how the Trolley Problem advances a purported *absence* of ethical decision-making capability, which becomes the *rationale* for ethics; thus there is a construction of 'the ethics of autonomous driving'. This is an instance of 'speculative ethics', because it "adopts without question particular imagined technological futures and extrapolates their ethical implications" (Nordmann 2007, p. 32). In this process, the actual science might evade scrutiny, and the "if and then" positioning becomes the framework for thinking about the future in urgent and immediate terms (ibid). Meaning: *If* we were able to make cars that could literally drive themselves, *then* we would have to develop a way to regulate them. Thus there is a co-constitutive relationship Nordmann identifies emerging through speculative ethics; a particular future becomes a source of ethical concern, which then, in turn, begins to shape the technology along those lines. Understanding the shaping of meanings of ethics and ethical decision-making in terms of the political-economic and commercial interests in developing this technology cannot be ruled out.

Eventually, these criticisms are generally silent on how unreasonable and irrational the Trolley Problem is; or as to why the starting point for ethics in the development of this technology is about finding a way to legitimise 'casualties', or 'unintended consequences'. The critical geographer, Louise Amoore, maps the social and cultural space of algorithms. She refers to the Trolley Problem as a kind of "madness" that operates in the "darkness of non knowledge." (2020, p. 119-122) What she means is that these formulations bring an irresponsible level of attention to the narrow moment a speculative crash that ignores the multiplicities of the world in which that crash might occur; and that expects computational decision-making to produce reasonable results when it is hard to justify impossible choices. No one knows what one might do in a case like that proposed in a thought experiment, nor how to program a driverless car to reach in the face of a such a choice. So, why are we framing ethical decision-making in terms of the unreasonable and unthinkable in the first place? What are the implications of this sort of framing? I argue that the selection of the Trolley Problem is in fact a rather well-thought through suggestion to advance an outsourcing of morality to computation *because of* the complexity of the problem. Next, I turn to a second case that this research addresses, leading with the question: 'what is autonomy?'

What is 'autonomy'?

> "We develop and deploy autonomy at scale. We believe that an approach based on advanced AI for vision and planning, supported by efficient use of inference hardware is the only way to achieve a general solution to full self-driving. Build silicon chips that power our full self-driving software from the ground up, taking every small architectural and micro-architectural improvement into account…

> Perform floor-planning, timing and power analyses on the design. Write robust, randomized tests and scoreboards to verify functionality and performance. Implement compilers and drivers to program and communicate with the chip, with a strong focus on performance optimization and power savings. Finally, validate the silicon chip and bring it to mass production. Apply cutting-edge research to train deep neural networks on problems ranging from perception to control. Our per-camera networks analyze raw images to perform semantic segmentation, object detection and monocular depth estimation. Our birds-eye-view networks take video from all cameras to output the road layout, static infrastructure and 3D objects directly in the top-down view. Our networks learn from the most complicated and diverse scenarios in the world, iteratively sourced from our fleet of nearly 1M vehicles in real time. A full build of Autopilot neural networks involves 48 networks that take 70,000 GPU hours to train. Together, they output 1,000 distinct tensors (predictions) at each timestep. Develop the core algorithms that drive the car by creating a high-fidelity representation of the world and planning trajectories in that space. In order to train the neural networks to predict such representations, algorithmically create accurate and large-scale ground truth data by combining information from the car's sensors across space and time…" (From the Tesla website about their Auto-pilot technology https://www.tesla.com/autopilotAI in June 2021)

Conventional understandings of 'autonomy' refer to 'self-reliance', 'self-determination', or individuality. If there was any doubt about what autonomy means in the case of driving, Tesla dispels it. 'Autonomy' and 'full self-driving' are entirely computational conditions reliant on a mapping of the world and then its replication inside a neural network-based model that only the driverless car can process and use to navigate the world. This high-tech language is dazzling and technical, and entirely inaccessible to someone outside this field. Tesla is the autonomous vehicle manufacturer founded and run by the entrepreneur Elon Musk. It is also a company that already has semi-autonomous vehicle models on the road.[13] Tesla's cars have

[13] At the time of completing this dissertation, Elon Musk and Tesla were in the news for attempting to promote their cars as "fully" self-driving, a claim that is under review by the California department of motor vehicles.

also been involved in fatal crashes. Auto-pilot refers to the technology that generates the sense of the car driving itself, much like in air flight, and relies on the vehicle being able to visually perceive its environment and communicate that data to its relevant processing systems. A popular definition conflates 'autonomous' driving with human-level decision-making: That if a system can achieve human-level decision-making to respond to the complexities of the world without requiring human intervention, then it is intelligent and autonomous. 'Fully autonomous' is taken to mean that there is no human driver; this is even beyond the case of airplanes and rockets where there are always humans at the wheel, so to speak.[14] A standard issued by the Society for Automotive Engineering shows Level 5 as "Full automation: the full-time performance by an automated driving system of all aspects of the dynamic driving task under all roadway and environmental conditions that can be managed by a human driver." (SAE, 2014) This measure is visualised in terms of a linear scale progressing from 0 to 5 depicting a human with their hands on the steering wheel, and with every stage slowly taking their hands off it. There is another a measure of the AV's autonomy called 'disengagements', which are applied in California; in this, AVs-in-testing must log how many miles they have traveled before requiring the human to step in to negotiate a traffic situation that the software cannot. Autonomous driving is proposed as a technical solution to the problem of human driving, which is found to be error-prone and the cause for most accidents. Yet, even though machine-driving is supposed to be more accurate and reliable than human driving, there is a catch: Autonomous driving technology, Tesla's hype notwithstanding, still relies on human effort in more ways than one. Crashes have occurred because computer vision

[14] Yet, there are even projects by companies such as Boeing to develop autonomous flight For example, the Leverhulme Centre for the Future of Intelligence at the University of Cambridge has been assessing public perceptions if autonomous flight. See Belton and Dillon (2020) .

systems have failed, and the human, was who not paying attention to the road, was too slow to take over from the auto-pilot that was in charge. But freedom from attention is exactly what autonomous driving is supposed to offer. This is particularly worrying since, a ‘machine ethics of autonomous driving’ relies on computer vision and auto-pilot; the AV must be able to make sense of its environment and know how to respond if the human does not have their hands on the wheel. Contrary to the popular narrative of autonomous driving, human drivers are not being *replaced* but are being *displaced* into responding to the needs of this emerging technology, to coax it into being as it were. This might be just par for the course, historically speaking; talking about what would make future AVs safe, Andrew Ng, Chief Scientist at Baidu, said that pedestrians need to follow the rules and be “lawful and considerate” (Kahn, 2018). He goes on to say that humans have always re-shaped their behaviour in relation to technology, but the truth is that this has also been pushed by the auto industry. For instance, in the early days of driving and before traffic lights were created, humans, who were only used to horse-drawn carriages, could not gauge the speed of cars; they would run out into the road in front of them. So jaywalking was constructed as a specific category of criminal offence in the United States to discourage this behaviour (Norton 2011). This work investigates similar kinds of transformations in human embodied and social relations in this latest evolution of automobility towards autonomy.

Ethical apparatus

In this work, I examine the relations emerging how the autonomous vehicle, which does not exist yet, and may not, is in fact being brought into existence through practices of “epistemologization" (Foucault, 2002, p. 205) enacted at sites such as the SAE levels of autonomy, disengagements, auto-pilot, computer vision, statistical models, thought experiments, machine ethics for autonomous driving. I refer to the driverless car as, and through, the frame of the *ethical apparatus.* Just as sight, vision, and perception were transformed by instrumentation such as telescopes, microscopes, cameras, and cinema (Halpern, 2014; Ihde, 1995), the era of big data, automation, and AI generate their own devices, "stak[ing] out new terrains of objects, methods of knowing, and definitions of social life" (Boyd and Crawford, 2012, p. 3). I argue that proposals to automate ethical decision-making are a kind of safety and accountability mechanism that are ostensibly enabling safe autonomous driving; and in this process are creating worlds legible and accountable only to this technology and within its own onto-epistemic universe. For instance, how machines process a tree, or a dog are mathematical and computational operations that render that tree or dog into statistical relationships that are legible only to the computational systems inside driverless cars that allow it to negotiate the world of trees and dogs. Additionally, the presumed erasure of the human despite the simultaneous deployment of human bodies and labour in furthering the perception of 'autonomy' requires myriad new infrastructures, measures, heuristics, and cyber-physical systems. Thus, I argue that in setting up the infrastructures for autonomous driving, we are seeing the development of various practices of knowledge-production translating the world into something legible to the navigational needs of the *car,* and thereby producing changes in the world itself through the actions of the car on that basis. So, in this research, I examine what constitutes these social, cultural, human, computational and epistemic practices that are organised to enable the emergence of autonomy, how they function, and how they are transformative of social relations. The *ethical apparatus* is a framework for this, building on Michel Foucault's definition of the apparatus, which is characteristically elaborate, so I reproduce it in full here:

> a thoroughly heterogenous ensemble consisting of discourses, institutions, architectural forms, regulatory decisions, laws, administrative measures, scientific statements, philosophical and moral propositions-in short, the said as much as the unsaid. The apparatus itself is the system of relations that can be established between these elements. Secondly, what I am trying to identify in this apparatus is precisely the nature of the connection that can exist between these heterogeneous elements. Thus, a particular discourse can figure at one time as the program of an institution, and at another it can function as a means of justifying or masking a practice which itself remains silent, or as a secondary re-interpretation of this practice, opening out for it a new field of rationality. In short, between these elements, discursive or non-discursive, there is a sort of interplay of shifts of position and modifications of function, which can also vary very widely. Thirdly, I understand by the term 'apparatus' a sort of-shall we say-formation which has as its major function at a given historical moment that of responding to an *urgent need*. (1980, p 194-195; emphasis in original)

What Foucault is referring to here as the apparatus is in the original French, a *dispositif* that enables and legitimises knowledge.[15] The apparatus is not a map-able, visible, mechanistic assembly of moving parts, but is about the making of practices of language, computation, and representations of the ethical and autonomy through the shaping of the driverless car. What I

[15] By 'apparatus' I do not refer to Marxist, Althusserian, or Gramscian approaches to the State and its institutions as repressive or as ideological apparatuses although these might be valid here too.

want to emphasise are the “system of relations” and “nature of the connection” between these elements of the apparatus; so, less its precise structure and organisation, and more the functions of these relationships and what they achieve in shaping the world of, and around autonomous driving. So, the task at hand is to identify a number of discursive ‘objects’ that would constitute this apparatus, how they surface and emerge, the contexts and moments of their emergence, correlations between them, how they are named and known, and how they reinforce, naturalise, and emphasise ways of thinking and talking about a topic (Foucault, 1972, p. 34-45).

The word *apparatus* contains two meanings within it: The first is apparatus as an ensemble of vectors of *discursive power*; and the other is apparatus as a *device*. I bring together twoinfluences: Karen Barad’s study of apparatuses in terms of instrumentation that create knowledge about the world; and the Foucauldian apparatus as the forces that shape discourse. Scientists have always used apparatuses to measure and know the world, but measurement devices and apparatuses are also always implicated in the creation of a measurable and observable world. For the one doing the measuring and observing is not separated from the world being measured and observed, but is of the world, and directly reveals the world on its terms. The measurement device is thus implicated in the creation of worlds and how we know them. If we unpack them, we might learn about their social and cultural origins. In this work I map various ‘devices’ and measurements, broadly construed, that are establishing what it means for the car to be autonomous or ethical: ’auto-pilot’ , disengagement reports, SAE levels, statistical modelling; computer vision; and the entire machine ethics approach might be considered a device constructing ethics as a measure of autonomy, and developing models of ethical decision-making.

This evokes the other meaning of apparatus as the Foucauldian *dispositif*, or the discursive practices shaping how we come to know and understand these notions, and what brings them power. Discursive power is difficult to pin down and is often hidden, and in that sense is a lot like ideology. Ideologies and politics are legitimised through material practices of language, institutions, bodies and social relations, and create new relations that are derivatives of the past and aspire to particular futures.

Finally, 'ethical' could sound like a personal attribute, or an adjective we might use for an individual or organisation, or practice. But, in machine ethics it refers to something that is programmed to return certain kinds of computational outputs. I believe there is a muddling of these two meanings that are rather convenient in making it seem that objects like table-top fans, homes, cities or cars are thought to *be* autonomous or intelligent. Designers like Rebaudengo and Cherubini are directly provoking the question of what it means to design something as 'ethical' when its responses might be unreasonable, or even unethical. I believe the confusion that this descriptor provokes brings attention to how these differences are constructed, and how we might bring attention to what work these terms are doing as representations.

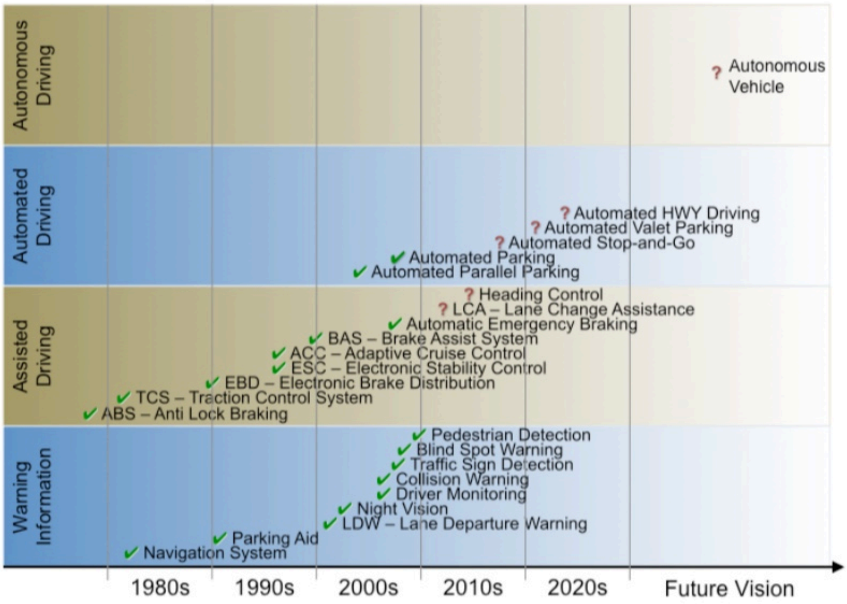


Figure 2: Classification and history of driver assistance systems (author's depiction, source: Wikipedia)

FIGURE 1. TIMELINE FOR THE DEPLOYMENT OF ADVANCED DRIVER ASSISTANCE SYSTEMS WITH THE VISION OF FULLY AUTOMATED DRIVING. COPYRIGHTED IMAGE BY SVEN BEIKER (2016) 'DEPLOYMENT SCENARIOS FOR VEHICLES WITH HIGHER-ORDER AUTOMATION' P 195

Which brings me to the challenge associated with these entanglements; that we are talking about a future technology that emerges from multiple forms and across temporal scales.

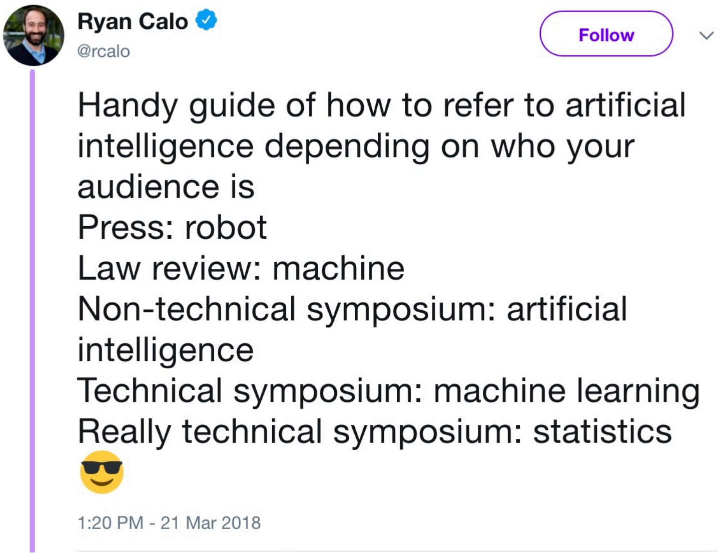
Multivalent cultural ontologies

To ask the seemingly straightforward question, “What is an autonomous vehicle?”, is to undertake a mapping of material practices of knowledge-making, cultural ontologies, metaphors, imaginaries, institutions, and infrastructures that constitute it. I ask this question because when we discuss the autonomous vehicle we are usually referring to a future imaginary; the Trolley Problem and AV ethics imagine a *fully* driverless car. But the future occupies an uncertain time frame; the AV was predicted to be on the roads by 2020 but shows no signs of materialising (Chin, 2020; Hudda, 2013). The question that I have been asked (and that I also routinely asked of autonomous driving experts) is *when* the autonomous vehicle will *be here*. The thing is, the autonomous vehicle is actually already *here* in a myriad of technical innovations and computational processes; but it is also *not here* in a ‘fully autonomous’ form; and even expecting it to take a ‘fully autonomous’ form, as if it were a superhero, is a commitment to a specific ontology. It exists simultaneously across different temporalities; and not just metaphorically, if we consider that infrastructure is being built with the future autonomous vehicle in mind.[16] It also exists in the past, as a car, and within a history and culture of automobility that is still with us today and undergoing transformation. In the study of discursive formations of autonomy, I want to multiply the sites from which to identify technolo-

[16] For example, for the past few years, Germany and its neighbouring countries have been testing and developing highways built to accommodate autonomous driving. These ‘digital test beds’ are being integrated into busy German highways: https://www.bmvi.de/SharedDocs/EN/Articles/DG/digital-motorway-test-bed.html

gies, its politics, and cultural and social relations emerging in the constitution of autonomous driving.

I do not argue that either ethics or ontologies are unstable, only that they are robust and multiple. Also, I do not emphasise *an ethics*, but an apparatus that makes meanings of and about the ethical. Thus my interest is in identifying what kinds of relations emerge in this process. In asking ‘what is an autonomous vehicle’, I take inspiration from a tongue-in-cheek tweet about AI from Dr. Ryan Calo, a Law professor who asks: What is AI? This question is deceptive because it opens up a variety of possible answers indicating its deep technological, social, and situated epistemologies and histories.



Ryan Calo
@rcalo

Handy guide of how to refer to artificial intelligence depending on who your audience is
Press: robot
Law review: machine
Non-technical symposium: artificial intelligence
Technical symposium: machine learning
Really technical symposium: statistics
😎

1:20 PM - 21 Mar 2018

FIGURE 2. SCREENSHOT OF TWEET BY RYAN CALO, WWW.TWITTER.COM/RCALO DATE OF SCREENSHOT: MARCH 21, 2018.

Calo suggests that we are dealing with multiple imaginaries, literacies, discursive anchors, expertise, and epistemic authority figures engaged in the shaping and emergence of AI. And the answer varies depending on whoever is interested in the question, 'what is AI?' He suggests that there is no single object that is 'AI'. To assume that it refers to a synthetic super-intelligence, or a fembot is to concede to a particular and specific imaginary; these are just *one* kind of cultural-ontological manifestation of what we refer to as AI. Just as, perhaps, there is no single entity that we might refer to as *the* driverless car.

In examining the material-discursive practices shaping what is meant by the ethics of autonomous driving and autonomy, I situate the automobile in its cultures and epistemes of automation, imagination, industry, human social relations, automobility, and software. Hence I *expand* the number of nodes from which we might perceive and understand this technology.This complexity is not just about size and scope, but in how they fluidly wrap around and anchor themselves to existing histories, ambitions, infrastructures, and futures. Hence, 'ontological multivalence' is an invitation to refer to dense cultural ontologies, and how different infrastructural formations produce new kinds of material and social realities accruing to discourses and meanings about 'autonomous' driving. In recognising the emergent AV as ontologically multivalent, I believe it might be possible to examine world-making on a larger temporal, and situated scale. The AV as AI/robot is considered a replacement for the human driver. The language of autonomous driving echoes the notion of a computational brain inside a vehicular body, or of automation without humans: driverless car; robot taxi; unmanned vehicular systems (unmanned aerial vehicle refers to drones). 'Self'-driving suggests the vehicle might have a sense of *self*; or that humans see it as having a self because it can navigate itself. Yet, this ‘brain’ is actually a distributed configuration of big data infrastructures of cloud connectivity, 3D maps, LIDAR, cameras, running various AI computational infrastructures. And, this is also an automobile that is a twentieth century media artefact; the driverless car is still a car emerging from a distinct social and cultural history that has significantly shaped twentieth century social, national, and economic life. In that sense, the AV of the future is a steady transformation of the contemporary automobile, which has itself been the site of increasing degrees of computational automation over the past forty years, also known as ‘automated driver assistance systems’ (Leonardi, 2010; Bengler et al, 2014; Beiker, 2016). Thus the AV is already multiple in terms of its constituent cultural ontologies: the future imaginary of the robot car, the connected and networked car-as-big data infrastructure, and the 20th century automobile.. I refer to each one in more detail later.

‘Multivalency’, here, is very loosely inspired by the concept of valency in organic chemistry, which refers to the capacity for atomic combination (specifically the exchange or sharing of electrons between atoms of different elements) resulting in the formation of new compounds. Thus in referring to the AV as *multivalent*, I am referring to the multiple sites of combination of its different technologies, imaginaries, histories, and cultures to dissemble and re-assemble, fragment, and come together in different ways. I am also building on scholars who have already identified the ontological collapses taking place through driving and automobility. The symbolic, affective, cultural, and embodiment of human and automobile fusing have evoked the ‘distinctive ontology of the person-thing’ like a “humanized car… or the automobilized person" (Katz, 2000 cited in Thrift, 2004, p. 47). There is also the collapsed 'map-territory' that refers to how the driverless car's computer vision and computation rely entirely on the map to navigate, thus removing the need for a separate notion of territory. Everything to the driverless car is a map and a territory (Hind and Gekker, 2019, p. 155) Further, there is an argument to be made for a "a new ontological category for robots somewhere between object and agent." (Calo, 2015, p. 119) What I believe this means is to identify how autonomous systems, like embodied robots, embody multiple, conflicting ontologies that must feature in our analysis. This is also evoked in *The Stack: On Software and Sovereignty.* The design theorist Benjamin Bratton elaborates on the condition of 'planetary-scale computation' that necessitates a conceptualisation of a self-perpetuating current and future governance model that simultaneously transcends and is limited by the nation state. What I want to highlight here is his identification of the ontological multivalence of the driverless car in keeping with the contemporary moment of pervasive computation that meets the weight and messiness of the Westphalian State and its traditional governance models. Emphasising the big-data infrastructural aspect of the AV in particular, he expressly identifies:

> the integrated design of driverless cars includes navigation interfaces, computationally intensive and environmentally aware rolling hardware, and street systems that can stage the network effects of hundreds of thousands of speeding robots at once..We see not one totality but the production of multiple and incongruous totalities, some of which are "interfacial regimes" .. may also displace existing geographies. (2015, p.12,)

Hence, critical scholars are motivated in arguing for an analytical lens that acknowledges the increasing complexity of these artefacts, their "incongruous totalities", and that singular narratives are likely to be limited in examining discursive and epistemological work that these artefacts do.

It might seem rather obvious that a single thing is not just one thing but is made up of multiple other things. However, it bears repeating and emphasis when AI is presented as having a single historical lineage or future manifestation, when in fact all computation is a dense, complex assembly of historical systems of systems. In examining the material-discursive shaping of ethics and autonomy, this research is particularly interested in these various dense layers This is the work of discourse, to supply the frameworks and representations that gather and hold these different densities together, and tie up them in seamless representations: Examining how these different ontologies come together, identifying the "gnashing and grating juxtapositions" between them, and the "peculiar new spaces, normal enclaves" that are generated, which "deliberately reorganise[ing] the world." (Bratton, 2015, p. 12) To conclude this introduction, I synthesise these arguments and background in terms of my research questions and sites, followed by an overview of the structure of this thesis.

Research questions and sites

Precisely because AI is a totalising and universalising domain that wants to apply itself to everything (Ernst, Schröter, and Sudmann, 2019), a study of material-discursivity is valuable in understanding how they become so powerful, and what constitutes this power. Research about driverless cars and AVs tends towards, policy and transport research focuses on concerns of regulation, safety, and efficiency, and deterministic popular discourses (Bissell et al, 2020). Social research about autonomous driving focuses on, 'human factors' in the interaction of humans as drivers of conventional vehicles, pedestrians, cyclists, or distracted operators (Schoettle and Sivak, 2015); challenges for policy and innovation (Stilgoe, 2019); and the socio-technical imaginaries influential in the process of technological change (Mladenović et al, 2020; Tennant and Stares, 2020). These worksidentify the AV as innovation on the *car* and as the future of ride-sharing and in mobility, or smart cities. By contrast, the present work engages with the emergent AV's cultures of computation, and the shaping of narratives, language, and knowledge, in the emergence of techno social futures. There are many missing spaces between AI, and society; this research is an effort to bridge that by showing how AI and autonomous technologies are social, cultural, historic, and situated in how they generate knowledge about the world.

The emergence of autonomous driving necessitates computational infrastructures that bring concerns of safety and accountability. According to the Trolley Problem, narrow though it is, this relates to unexpected accident scenarios that the AV will not be able to compute; Or, worse, will compute in such a way that (potentially) brings harm to humans. Even if this thought experiment was only intended as a provocation about what would be required of autonomous vehicles, it has captured media attention, and stimulated scholarship about the 'ethics of autonomous driving'. Machine ethics proposes to architect an ethical machine, chiefly the driverless car, that will act in accordance to human values systems. The Moral Machine is one proposal that assembles a large global dataset of human moral values that its developers believe might architect a 'universal machine ethics', or moral decision-making by algorithmic systems. This project is met with enthusiasm, but I believe the shift it proposes towards statical operations based on such large datasets to architect ethical decision-making has implications for what the ethical is; I argue that we need to investigate this more carefully. Further, I argue that the entire production of autonomous driving necessitates computational infrastructures that are creating a world legible to and for the navigation of a driverless car. These are technologies of knowledge-making like statistical modelling, mapping, computer vision, and automated and algorithmic decision-making. In enabling autonomous driving, these technologies are also shaping the world around the autonomous vehicle. For example, AV crashes have already occurred because of faulty computer vision that does not recognise objects around the AV correctly; and then require human drivers to take control back. These intimate interactions between human and machine in enabling the mirage of autonomy tend to be obscured. Thus, I argue that it is critical to understand how these technologies are operating in the world through the entangled shaping of ethical decision-making and autonomous driving. Thus I bring these concerns together in terms of the question

> **How is the material and discursive shaping of ethical decision-making and autonomous driving generating social and cultural transformations?? What are the implications of this for human bodies, social relations, and societies?**

It is constituted by the following sub-questions:

- What are the social, cultural, and material practices that constitute autonomous driving?
- How is the discourse of the ethics of autonomous driving generating new modes of the ethical?
- What kinds of transformations are taking place in human social relations, spaces, and bodies through the emergence of autonomous driving?

In this work I emphasise materiality, discursivity, and *material-discursivity* to emphasise the situated practices that are shaping autonomous driving as 'autonomous', or ethical decision-making as 'ethical. This can tell us about the representational terms under which we speak about autonomy or ethics in the driving context. Representations are intentionally created, situated, ways of naming, discussing, and knowing about things in the world. There is nothing random about representations, and particularly in relation to how scientific and technical knowledge are situated and shaped (Knorr Cetina, 1983, 1994). I believe this is critically important in the emergence of speculative and new technologies; what autonomous driving means is neither well-known or established; so, meanings must be intentionally made. Many kinds of communities and institutions, histories and cultures, are engaged in this process, and this is what the ethical apparatus allows us to identify. Material-discursivity is about the material, cultural, and social practices enacted in a variety of contexts, in which knowledge and representations are made, creating legitimacy, validity, and the possibility to 'intervene' in the world (Barad, 2007). In identifying material-discursivity, I am interested in the conditions under which terms like autonomy and ethics are created and how they come to intervene and have power in the world. I investigate how these terms are created through material sites, locations and practices: Thought experiments, accident scenarios, scales, heuristics, maps, statistical models, computer vision, among others. I engage these practices through the following sites and locations:

- Machine ethics, the Trolley Problem and its iterations, and the Moral Machine project[17] from MIT Media Lab.
- Cases of recent fatal autonomous vehicle crashes in the United States.
- Technologies of autonomous driving, chiefly computer vision and auto-pilot settings
- 20 in-depth, unstructured interviews with academic, policy, and industry experts from Germany, India, and the United States working in the Law, Computer Science, Design, Mapping, Robotics, and Automotive Engineering.
- Two interviews with Tesla owners in North America, and a test drive with one of them.

Structure of this text

Chapter 2 Introduces the theoretical influences that shape this research, starting with ‘how knowledge is made’ through the lens of the ethical apparatus, wedding the Foucauldian *dispositif*, or ‘heterogenous ensemble’ with measurement and the making of representations as elaborated on by theorist, Karen Barad. I also introduce parallel influences such as the sociology of knowledge approach by German theorists, Berger and Luckmann, that directly

[17] https://www.moralmachine.net/

builds on Foucauldian discursive shaping; and Jasanoff and Kim's socio-technical imaginaries. Then, I discuss 'how AI makes knowledge', introducing its relevant intellectual histories, embodiment, robotics, statistical modelling, the role of game-playing, and error. 'Agency' is a question often asked of AI and autonomous technologies, but I address 'agency' in terms of social shaping of technologies approaches, primarily that of the socio-technical and the technosocial. And use this to reflect more on the debates of how technologies are, or become, ethical. Finally, I conclude with a discussion of the multivalent ontologies of the autonomous vehicle that describe the various cultures of computation and automobility that situate this artefact.

Chapter 3 Details the practical contexts and situations in which this research was undertaken.

Chapter 4 This chapter introduces computer vision and the auto-pilot setting, two key sociotechnical systems constituting autonomous driving . The 'irony of automation' is an observation of the tensions emerging when a computationally superior and efficient machine actually needs human operators to ensure that it is working effectively, and that the human is inevitably always held accountable for errors, even if the machine is more efficient or accurate. This is a legacy of safety engineering and design from aviation, in which the human and machine are understood as separate and having to 'fit' or 'flow' into one another. With the emergence of the autonomous vehicle as simultaneously AI/ robot , automobile, and distributed, big data infrastructural platform, these beliefs are dissolving into the ironies of autonomy:

Humans are invisible, valuable, and minimally rewarded micro-workers, surveilled by data infrastructures, and re-shaped by its information flows. Hence, autonomous driving is about multiple, dense layers of the environmental and the infrastructural, as well as human work and embodiment. This prompts a reconsideration of what we mean by ‘autonomy’ in the context of the complicated, intersecting geographies of the contemporary data landscape.

Chapters 5 This chapter presents the empirical sites that constitute the social, cultural, and political-economic entanglements of the shaping of the ethics of autonomous driving discourse. It begins with an overview of the machine ethics approach, and then introduces research about the Trolley Problem’s reception and impact, including the role of a small cohort of engineers and ethicists at the forefront of shaping this discourse in the media. Then, I turn to an analysis of the Moral Machine from MIT Media Lab that frames ethical decision-making in terms of a statistical answer to an impossibly difficult question.

Chapter 6 The final, concluding chapter in this work reflects on the implications of the discursive shaping of ethics and autonomy, identifies themes for further research, and closes with suggestions for a reconceptualisation of ethics in the context of AI systems and algorithmically mediated life.

CHAPTER 2.

THEORETICAL INFLUENCES

Making discourses and knowledge

This research started with a mixture of disbelief and curiosity about how the humble and familiar car was going to become a *driverless* car; how the engineering was going to work; what made it 'autonomous'; and why the *ethical* was being evoked across the emergent landscape of algorithms and AI that constituted this future autonomous car. The making of knowledge is a key dynamic that creates the possibilities of worlds emerging. Powerful groups of experts, like engineers, scientists, and industrialists, come together to develop or change things with their own highly situated understanding of the world, and largely by wielding knowledge and knowledge-making in its various forms. So, asking how discourse and knowledge are made is necessary in examining the influence on culture and society by this emerging, speculative technology.

I emulate Michel Foucault's approach to 'problematisation', a descriptive and analytical mode that does not only reveal, but rather "rearrang[es] what is already known, of seeking to “make visible what is visible”” (Barnett, 2015, np). Despite the associations of seeing with knowing, I present this work as an unsettling, rather than an exposing. In fact much of this work has relied on the public discursive shaping of a new technology by academic and industry experts who have been given substantial attention in the press and academic venues; hence, there is a lot that is already visible in this domain. So in a Foucauldian tradition, it is not *what* was said, but “how is it that one particular statement appeared rather than another?” (Foucault, 1972/2002, p. 27) which I understand to mean the relationships between these different things like cars, autonomy engineering, ethics, drivers, and AI, among others, that hold them together and make them all connected in how we discuss the future and emergence of autonomous driving.

I have been influenced by two theorists who operate in very different dimensions and scales but are nevertheless connected because of their different renderings of their development of the *apparatus,* in terms of its two entwined meanings, as a measuring device or assembly of devices; and as discursive power. Michel Foucault is a meticulous scholar of over three hundred years of European society in almost all its institutional, philological, meta-epistemological, and interpersonal forms. On the other hand, the feminist technoscientist, Karen Barad,[18] organises her techno scientific explorations of the state of reality and existence through a critical re-reading of one of the most fundamental experiments in Western science of the past century: Niels Bohr’s two-slit *gedanken* experiment that took on the ontology of light itself. I will start by explicating on Barad’s specific influence on my thinking about apparatuses, then move to the Foucauldian approach. Then, I conclude this section by introducing two other

[18] Karen Barad is Professor of Feminist Studies, Philosophy, and History of Consciousness at the University of California at Santa Cruz.

approaches to the shaping of knowledge, both of which are more sociological: the SKAD approach (Sociology of Knowledge Approach to Discourse) that builds on Foucault's work on discourse; and Jasanoff and Kim's 'socio-technical imaginaries'. I introduce relevant literature germane to the topic of the shaping of new technologies, and then move to a more extensive discussion of the AV in terms of being a configuration of AI. Further, I lay out other theoretical influences related to how the social and technological are co-constitutive and entwined; a question that arises frequently in discussions of autonomous vehicles is to what extent they might be considered 'agents'. This is a conflicted and divisive term that I believe needs some clarification because it is claimed by different domains simultaneously. Finally, continuing with the social-technical I introduce a discussion of the relationships between values and technology relevant to this work, in particular a debate that has emerged in the mitigation of algorithmic bias that relates closely to the question of how and if a car can be programmed to be ethical. Through this, I weave in aspects of the emerging driverless car's different cultural ontologies as AI/robot, big data infrastructure, and conventional automobile.

The Baradian apparatus

A full overview of Barad's theory of *Agential Realism* is not possible here nor does it serve this work; however, she addresses the origins of some of the really elemental ideas in Western thinking and pulls them apart, as Niels Bohr did, to reveal that nothing is really apart and only appear that way because of language. In fact, she argues, bordering on the mystical, nothing is fixed, everything is always connected by the potential of moving towards becoming something, because, she argues, matter is not a thing, but a dynamic 'intra-activity'. There is a key reflection in Barad's text *Meeting the Universe Halfway,* that is germane here. Early on in the book there is discussion of the mysterious interactions between the two World War 2 quantum physicists, Werner Heisenberg and Niels Bohr; mysterious, because it was unknown if it was about science, or politics, or a combination of both.[19] But where Heisenberg is known for his famous 'uncertainty principle', Bohr was known for *complementarity*. Heisenberg concludes that there are tradeoffs in how we know something; disturbing one set of values or parameters of our inquiry sets limits to knowing something else in the universe; following from a classical Newtonian tradition in physics, this assumes that everything in the world is sitting around with a set of fixed attributes that can be known by measurement. Extrapolating outwards, this might mean that it is near impossible to know anything conclusively about what we intend or think. Niels Bohr developed a *gedanken* experiment, a thought experiment, which was a fairly standard practice for high-end quantum physicists in the first part of the twentieth century to ponder the nature of existence. Bohr's experiment related to resolving the wave-particle duality of electrons, an experiment that was actually only technically feasible in the mid 1990s, decades after his passing (Barad, 2007, p. 100-102). It is remarkable that he arrived at these conclusions based solely on the rendition of a hypothetical apparatus. What Bohr predicted, correctly, was that the measuring apparatus changed the outcomes of the experiment; electrons exhibit as both wave and particle depending on how they are measured. This vertigo-inducing realisation means that even the most basic elemental stuff of life can change its form based on how it is measured. Hence, the determinate material conditions of the experiment, the apparatus and its setup, influenced the outcome. In other

[19] Barad opens her work with a critical reading of a stage play called *Copenhagen*, a fictionalised account of the events around the meeting of two legendary physicists, Werner Heisenberg and Niels Bohr in 1941. Heisenberg visits Bohr, potentially relate to Heisenberg's potential role in Germany's effort to build an atomic bomb. It is not clear what transpired but that meeting might have been pivotal to Germany not eventually securing that technology. The play is about personal reckoning and morality against this tense political moment; as well as the interior reflections of scientists who had to reconcile the fundamentals of 'science', assumed to be neutral and objective, with the social that was in disarray and entirely corrupted.

words, phenomena are emergent from the conditions of the apparatus that measures and knows them (Op Cit., p. 103-106). And then, extrapolating similarly, Barad concludes, through Bohr, that what accounts for intentionality is that it cannot be understood as a preexisting determinate state with fixed values, and that it "needs to be understood as attributable to a complex network of human and nonhuman agents including historically specific sets of material conditions that exceed the traditional notion of the individual." (Barad, 2007, p. 19) Whereas Heisenberg challenges the nature of *how we know* reality, Bohr challenges the nature of reality itself.

In this research I address the *representationalism* that I believe autonomous driving technologies are shaped by; representationalism is the not-making-clear, and not-acknowledging of the socio-material practices that create representations of things, how we know them and how knowledge about things take form and have power in the world (Barad, 2007, p. 46). My methodological approach emerges from Lucy Suchman and Jutta Weber's critical discussion of how and why we come to think of lethal autonomous weapons systems (LAWS) as autonomous. They argue 'autonomy' is an intentionally wrought representation that obscures its origins, and that thus urge us to "reconceptualiz[e] … autonomy and ethics as always *enacted within, rather than separable from,* particular human-machine configurations." (Suchman and Weber 2016, p. 76, emphasis mine) And "to identify the materialization of subjects, objects, and the relations between them as an effect, more and less durable and contestable, of ongoing socio-material practices." (Suchman, 2009, p. 285) This is how they address representationalism, which in the case of LAWS creates a rationale for a technology that is so obscured in jargon and securitisation that it precludes opportunities for resistance through scrutiny. (I return to the socio-material further ahead in a discussion of the closely related concept of the socio-technical.) Thus in seeking out how representations are being made, I examine practices of measuring and quantification, broadly construed, that are important in making meanings about what it means for a car to be autonomous, or for it to make ethical decisions. The instability of these notions and their reliance on the intentional construction of apparatuses that would identify and measure them, is what Barad's work offers. Apparatuses as measuring devices are neither inert, objective, nor universal, they are productive of phenomena they purportedly measure, and betray their origins if we study them (Barad, 2007. p.146). In other words, apparatuses do not sit apart from the world to passively observe and record it but are as large, small, or expansive as the determinate local conditions of their assembly. Barad writes,

> The apparatus is in and of the world, shaped by language, and this is central in the apparatus as a measurement device. This is how material practices accrue to particular realities and relations: ontologies. This is how the autonomous machine has the power to re-shape our knowledge of the world… to enable dynamic (re)configurings of the world" (Barad, 2007, p169-170).

Hence, a device that measures up to 100 on a scale does not allow for a value of 101 to exist in the universe circumscribed by that device; this does not mean, however, that the value of 101 does not exist. Moreover, where that device was made and how the scale was architected both reveal how and why the number 100 is important to that scale. And if we wanted to identify how the scale was made and why 100 is important, then we would have to move outwards from the moment of measurement into the measurement device and the world around it. An apparatus, Barad argues, is more than just the sum of its parts, but it keeps extending materiality as a counting exercise very far back. She describes a lab experiment and asks how each part of the experimental set-up is connected to something else, from the computer that collects data, to the printer that prints out the results, to the person who loaded paper in the printer, and scientists who read the marks on the paper to make sense of them: "What precisely constitutes the limits of the apparatus that gives meaning to certain concepts at the exclusion of others?" (2007, p. 142-43). So we might examine that the 'apparatus' also suggests an *assembly* of variegated, dispersed, non-linear, relational, fragmented, multi-scalar processes. And what she means here is that these assemblies are entirely social, and have a significant impact on social relations, but not always in ways that can be identified and mapped at the time. Barad goes on to say that these measurements give us language and representations of the world around the device, they make meanings because they limit our knowledge of the world. It is in this historical, situated, materiality of devices of knowing that there is convergence with the work of Michel Foucault and *his* concept of the apparatus.

The Foucauldian apparatus

There could be a variety of ways to map the implications of a Foucauldian apparatus for shaping the emergence of discourse. I limit myself to leveraging the Foucauldian apparatus as the perspective that, in the making of a discourse of ethics or autonomy there are disparate elements that are not just related to cars, but to multiple other objects and ideas that emerge simultaneously at different points in time, coming together in formations that can be studied and mapped. The 'apparatus' brings together the multivalent ontologies of the driverless car, to resist a normative approach to ethics, and to show how it is being framed, measured, and represented by social, historic, cultural, and material actors both human and nonhuman. These are practices of 'epistemologization', the making of epistemologies through situated practices,

There is a close alignment to be found in Barad's and Foucault's apparatuses concerning the production of knowledge. Barad's work actively addresses nonhuman technologies like the ultrasound scanner takes taking a socio-technical approach, examining how technologies are architected like devices, as epistemological instruments that make knowledge through situated practices of measurement and thus subtly establish the world, and the terms on which we know that world. Foucault's work is at a much more expanded scale, examining the histories of the formation of entire epistemologies across centuries in the establishment of social and cultural ways of being and collectivity. I engage both the Baradian device together with the Foucauldian *dispositif* to analyse how this technology comes to be in the world.So I argue that autonomous driving and ethical decision-making can be studied as devices that are establishing in   'autonomy' and the ethical discursive entities. The ethical apparatus shows how different material and relational practices are establishing the ethical and autonomous as inter-related and as discursive; 'autonomy' as the result of AI and automotive engineering; and 'the ethical' imagined as a decisional output; and the implications of these for the shaping of the autonomous vehicle.

A related approach in the German-speaking context refers to the emergence of discourse in sociological and social terms, and that I turn to next, the Sociology of Knowledge Approach to Discourse (SKAD) advanced by Berger and Luckmann. Rainer Keller's discussion of SKAD discusses how this field develops an approach to discourse rooted in the Focauldian

(2011) In this, there is an emphasis on ‘concrete’ forms of the social that become discourse, such as cultural material like books, texts, and media. SKAD also includes actual humans centrally in ‘discourse’, something that Barad and Foucault tend not to. Just as the SKAD approach emphasises people as speakers, and as actors negotiating power through knowledge, in this work I find that autonomous driving is being shaped through experts who are making the language and representational forms of discourses of ethics and autonomy in driving.[20] The newly emergent discursive sub-domain, 'ethics in AI and Big Tech’ suggests a similar dynamic. The expertise of engineers is part of the perceived problem of lack of ethics, and hence the site of its solution as well. A rationale goes that if we fix how computer scientists and engineers architect and develop AI and autonomous technologies broadly, then we are likely to see a shift in how AI is in the worldThis assumes that computer science ‘lacks’ ethics, that teaching ethics to individuals will correct this socio-technical, social shaped technology; this risks ignoring the apparatus that technology emerges from and is shaped by. Engineers are but one part of this system, and hence attempting to influence them has a largely unknown effect. It is also incorrect to assume that entire disciplines or individuals lack morality or ethics. A social shaping approach would locate individuals as organisational and cultural communities embedded in various social-relational-epistemic infrastructures.

Karin Knorr Cetina refers to ‘epistemic cultures’ as the “sets of practices, arrangements and mechanisms bound together by necessity, affinity and historical coincidence which, in a given area of professional expertise, make up how we know what we know” (Knorr Cetina, 2007, p. 363). Here ‘epistemic’ refers to the constitutive cultures and cultural practices of know-

[20] I profile the work of three experts who have been extremely influential in shaping the ethics of autonomous driving discourse: Jason Millar, Patrick Lin, and Iyad Rahwan Interviews with Lin and Millar have been informative in understanding the emergence of ethics and autonomy.

ledge-making. Knorr Cetina's work comes from studies of how physicists and biologists in laboratory settings create knowledge, and is rooted in constructivist approaches to the scientific shaping of knowledge. Her work maps the "machineries of knowing" that create, warrant and legitimise knowledge: a "nexus of lifeworlds" that acknowledges the spiritual, non-material, as well as embodied uses of symbols and meanings in material contexts (p. 364). In other words, the socio-material practices and situations in which discursive ideas get made, but also conveyed. epistemologies take the 'shape' of the socio-material 'containers' they get made in, and discerning what those containers are reveal how diverse social actors-human, rhetorical, affective, political, relational-constitute and influence how we know things. Epistemic cultures alert us to how epistemic objects, like a thought experiment or a big data infrastructure, naturalise certain orders of thinking, and the lifeworlds they bring into being. Why this matters is because it recognises that putative makers of knowledge - scientists, engineers, CEOs, laboratories, inventors - are influenced and constituted by powerful social and political forces. The values and desires of these entanglements pervade technologies.

Socio-technical imaginaries and constructions of technology

The relationship between the making of scientific knowledge and place and situation is further advanced by work in Science and Technology Studies (STS), that emerges in the interstices of History, Sociology, Anthropology, and related disciplines and fields in the Humanities and Social Sciences. STS brings with it a resolute suspicion of hype, surfaces, and affect and situates epistemologies and knowledge as actants, and how they have political and material influence beyond the making of representations. This study refers extensively to STS theorists in the study of models (Galison, Leonardi), the histories of intelligence and the evolution of calculating machines (Daston,), crashes and accidents(Brown, Downer, Galison). A key concept from STS isSheila Jasanoff and Sang- Hyun Kim's 'socio-technical imaginaries', a definition that captures the variety of influences on how technology is 'made' as a: "collectively held, institutionally stabilized, and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology." (2015, p. 3-4)

A number of key ideas stand out in this definition relevant to this work.
The word 'desirable' is of relevance here, suggesting that intentions and visions are also often hidden from us and must be identified through a triangulation of social and political actors, including more discursive and rhetorical ones. Such as, for instance, the question of what constitutes 'the imagination'? Here, we can point to the powerful place of media and cultural artefacts, narratives, and tropes in shaping visions of technology and the future, particularly when it comes to artificial intelligence technologies (Cave et al., 2018) Identifying the autonomous vehicle in terms of the robot imaginary, suggests that how the 'self-driving' car exists in literature and cinema has a place in how visions of the future AV will inform the creation of conditions influencing its actual design and emergence. This is not suggesting a causal or circular link between imagination, discourse, and design, but recognises that future visions also act as self-fulling prophecies. This research analyses the close association in engineers' and marketing rhetoric to the AV as a replacement driver for the human as evidence of how sociotechnical imaginaries work . The 'socio-technical imaginaries' definition identifies institutional actors in 'stabilizing' visions of the future, echoing the notion of stabilisation that

is key to Social Construction of Technologies (SCOT) theories about how technologies emerge. Put forward by Wiebe Bijker and Trevor Pinch, SCOT proposes that powerful social actors, that they refer to as *relevant social groups* (RSGs), have a role in narrowing down the multiple possible perceptions and ideas of what a technology is. Owing to their social power or, epistemic credibility, or both, the *interpretive flexibility* of a technology, its potential to evolve in more fluid and multiple ways through the openness of what it means, is foreclosed. Jasanoff and Kim discuss how multiple deployments of ethics as a regulatory and epistemic actor has shaped *other* new and speculative technologies.

In *Dreamscapes of Modernity*, Benjamin Hurlbut recounts a curious pre-history of a set of AI Ethics principles called 'Asilomar for AI'. In 1975, roughly 140 geneticists, self-appointed leaders in the field, gathered at a conference centre in Asilomar, California to discuss something radical taking place in their field. Gene splicing technologies had just been discovered and this promised what had only existed in the realm of science fiction: The ability to combine and clone genetic materials. These scientists drew up ethical guidelines for genetic engineering research and explicitly named the lack of expertise in government and the lay public as the reason for needing to self-regulate. They wanted to enable future research and development away from government 'interference', thus they drew up a set of guidelines called the Asilomar Principles to regulate research and development in genetic engineering. The result was the boom in the biotech industry that we continue to witness today more than 45 years later, with the exception of brief intervention by the US government when Dolly, a sheep, was cloned in an Edinburgh laboratory (Hurlbut, 2015).

While Asilomar was beneficial for genetic engineering, AI Ethics has been exposed as a status quo-ist industry response to secure its own domain of research and development. ’AI ethics’ is a new phenomenon about governance and regulation of the AI research, development, and production industry. The ‘wrongs’ thought to be associated with AI are a broad amalgam of many things that are problematic with applications of machine learning systems in black boxes, some of which I will discuss ahead: bias and discrimination; denial of individual autonomy, recourse, and rights; non-transparent unexplainable or unjustifiable outcomes; invasions of privacy; isolation and disintegration of social connection; unreliable, unsafe or poor-quality outcomes. (Leslie, 2019, pp. 3-5) This takes shape around AI Ethics Principles that guide and orient research and development of AI technologies. By one count, there are at least 84 (Jobin et al, 2019) and by another 160 (Algorithm Watch 2020) sets of ethical principles and guidelines for AI ethics put out by states, technology corporations, think tanks, civil society groups, scientific societies, professional associations, and private sector alliances. There is also a top-down dimension to industrial AI ethics, which locates the problem, and the potential for change, in individuals with immense social power, wealth, and capital. In the past few years, particularly since the Trump administration, there has been an exodus of “reformed techies”[21] from Silicon Valley, chiefly Silicon Valley entrepreneurs who acknowledge that they have helped to architect products that have now enabled everything from the hollowing out of privacy, to the mass manipulation of democratic processes like elections (Ganesh, 2018).

[21] A phrase coined by the journalist Adrian Chen: https://twitter.com/AdrianChen/status/960945570050228225

AI ethics is eventually shallow, and about industry accruing power to regulate itself. The 'AI ethics' discourse is shaped by Big Tech' possibly as a tactic to evade regulation and continue with business as usual. These principles offer ethics focused on technical design *without* an "explicit focus on normative ends devoted to social justice or equitable human flourishing" and particularly given what is known about bias and discrimination (Greene et al, 2019; Ochigame, 2019).Thus, 'AI ethics' might be a temporary but important discursive moment of collectively being alert to the shaping of the ethical and 'ethics'. Scholars refer to the corporate/ Big Tech interest in ethics as corporate 'ethics-washing' (Wagner, 2018) and as 'toothless' because it approaches ethics as a regulatory mechanism, as if it were law. (Resseguir and Rodgrigues 2020).

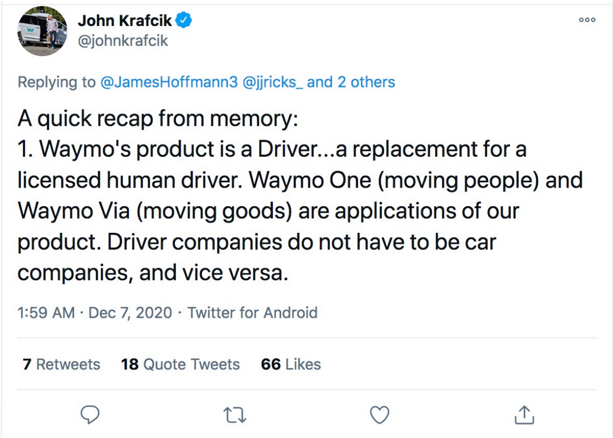
Similarly,, Alfred Nordmann argues that 'speculative ethics' enabled the development of nano ethics as a distinct field when it was difficult to say that there were indeed any potential risks from the development of nanotechnologies. The technology was not under any serious scrutiny but the ethical concerns had raced ahead, generating an entire discursive field. So, nanoethics is a "merely possible" speculative future that turns into something inevitable, displacing the actual, and overwhelming the present (Nordmann and Rip, 2009, p. 273) Strangely, AV ethics and AI Ethics rarely meet over any shared concerns. Why is this so? This dissertation traces the genealogies of both strands to understand why.

How AI makes knowledge

The driverless car as the AI/robot imaginary

The shaping of the driverless car as a 'socio-technical imaginary' (Jasanoff and Kim 2015) and the 'imaginaries' frame reveals the connective tissue between the interior associations we have with technology and its broader social and political-economic dimensions. Autonomous vehicles, broadly understood, already exist in restricted contexts, such as in drones, deep sea exploration, in war and conflict situations, and mobile delivery in controlled environments. The fully autonomous vehicle, with an emphasis on *fully*, does not exist anywhere but in the future. What is the relationship between these past fictions and our present? Is the AV materialising now because its mention in Science Fiction (SF) served as inspirations for scientists and engineers? "Science fiction does not merely anticipate but actively shapes technological futures through its effect on the collective imagination"; and "Science fiction visions appear as prototypes for future technological environments " (Dourish and Bell, 2014, p. 769) However, it would appear that the emergent AV is not a prototype anymore but is an actual artefact being publicly tested. What are some of the narratives and material conditions being generated at the intersection of SF imaginaries and engineering?

In a recent, public discussion of their work, the CEO of Waymo, the Alphabet-owned self-driving car project, John Krafcik[22] says that their product is a replacement for a licensed driver. Interestingly, Waymo never talks about the ethics of autonomous driving explicitly, however, the "replacement of the human driver" that Krafcik refers to, becomes the popular perception of what autonomous driving is. In the tweets included here, Krafcik un-ironically frames 'the human driver' in mechanical and computational terms: a mix of hardware and software. Their driverless car, the Waymo Driver, is similar; however "the AI part is the hardest", he acknowledges. There is an utterly serious way that AI is referred to as the human driver-like element that will animate autonomous driving; on the other hand, this intelligence is in fact many technologies, infrastructures, and human labour.



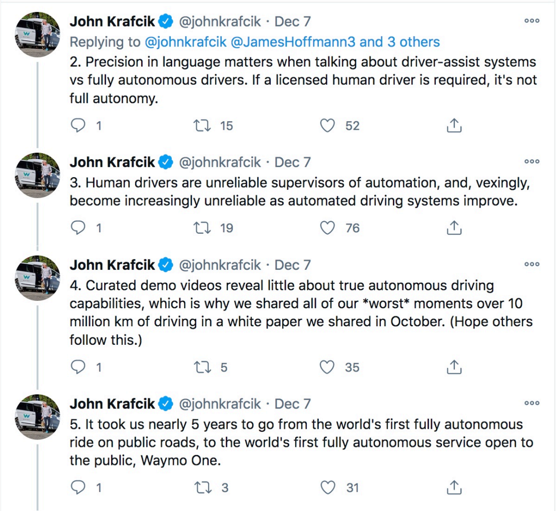
John Krafcik
@johnkrafcik

Replying to @JamesHoffmann3 @jjricks_ and 2 others

A quick recap from memory:
1. Waymo's product is a Driver...a replacement for a licensed human driver. Waymo One (moving people) and Waymo Via (moving goods) are applications of our product. Driver companies do not have to be car companies, and vice versa.

1:59 AM · Dec 7, 2020 · Twitter for Android

7 Retweets 18 Quote Tweets 66 Likes



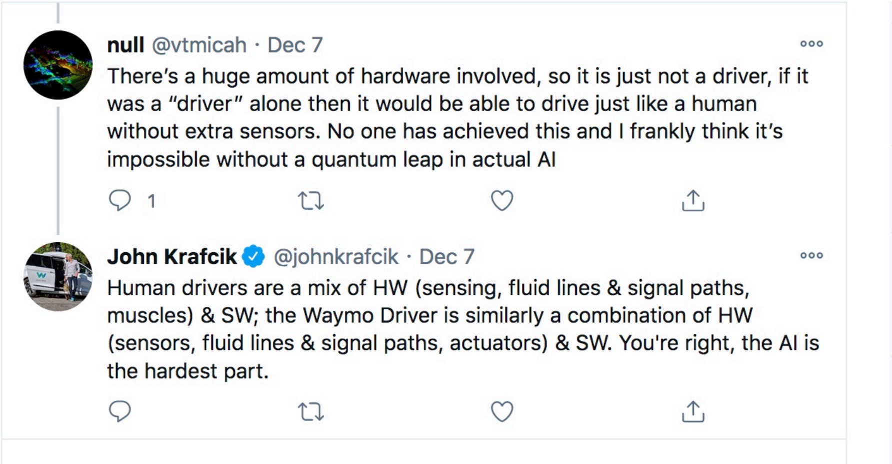
John Krafcik @johnkrafcik · Dec 7
Replying to @johnkrafcik @JamesHoffmann3 and 3 others
2. Precision in language matters when talking about driver-assist systems vs fully autonomous drivers. If a licensed human driver is required, it's not full autonomy.
1 15 52

John Krafcik @johnkrafcik · Dec 7
3. Human drivers are unreliable supervisors of automation, and, vexingly, become increasingly unreliable as automated driving systems improve.
1 19 76

John Krafcik @johnkrafcik · Dec 7
4. Curated demo videos reveal little about true autonomous driving capabilities, which is why we shared all of our *worst* moments over 10 million km of driving in a white paper we shared in October. (Hope others follow this.)
1 5 35

John Krafcik @johnkrafcik · Dec 7
5. It took us nearly 5 years to go from the world's first fully autonomous ride on public roads, to the world's first fully autonomous service open to the public, Waymo One.
1 3 31

[22] Waymo tends to be very private about their self-driving car project, compared to, for example, Elon Musk's very public displays of developments with his Tesla project.



null @vtmicah · Dec 7

There's a huge amount of hardware involved, so it is just not a driver, if it was a "driver" alone then it would be able to drive just like a human without extra sensors. No one has achieved this and I frankly think it's impossible without a quantum leap in actual AI

John Krafcik @johnkrafcik · Dec 7

Human drivers are a mix of HW (sensing, fluid lines & signal paths, muscles) & SW; the Waymo Driver is similarly a combination of HW (sensors, fluid lines & signal paths, actuators) & SW. You're right, the AI is the hardest part.

FIGURE 3. SCREENSHOTS OF TWEETS BY JOHN KRAFCIK, CEO OF WAYMO, ON DECEMBER 7, 2020. DATE OF SCREENSHOT: DECEMBER 12, 2020

Artificial intelligence is constructed through a fertile and messy exchange of metaphors about human and machine. In fact, AI itself has been developed as a metaphor for thinking and intelligence. Metaphors are powerfully entangled with epistemology even when they are not accurate, and are constitutive of theory particularly in young fields of research. Theoretical psychology, for example, is rife with analogies of humans as computers and vice versa, that, "computer metaphors have an indispensable role in the formulation and articulation of theoretical positions" about how the human mind works (Boyd, 1993, p. 487). Notably, the roboticist Rodney Brooks has been deeply critical of this metaphor referring to it as an "intellectual cul de sac" that does not advance AI (Brooks, 2012, np). The AV is imagined as an artefact that is humanoid in processing capabilities in the same way that AI/ artificial intelligence is, an 'awesome thinking machine' that will make decisions for itself, automatically or, 'autonomously' (Natale and Ballatore, 2017). AVs in advertising, cinema, literature, TV programs and industry literature are resolutely anthropomorphic (Kröger, 2015). Drive.Ai, a Silicon Valley software company says that they are “building the brain of driverless cars”; and the BMW Sales and Marketing Lead echoes this: “now we’re in the ‘hands off’ and ‘eyes off’ phase, but only for brief periods. The next phase will be ‘brain off’” (2015). Both are directly modelling autonomy on a separation thought to exist between human body and cognition; a fetishised individuality, an atomised independence, and separation. The driverless car as AI, a machine that exists independently embodies a fetish of individuality; this fetishised individuality is assumed to be human. Is the autonomous vehicle a ‘connected and networked car’[23] that is part of a larger data assembly, or the replacement of the human driver, or a robot that is independent? All of these imaginaries and fantasies co-exist.

The media often rely on specific tropes of the future from genre Science Fiction in particular in the popular discursive shaping of the AV. The place of fiction, imaginaries, and metaphors cannot be understated in their discursive power. In the movie *Minority Report* (Steven Spielberg, 2002) a ‘fully’ autonomous vehicle plays a key role at one of many decisive moments. The protagonist, a detective in the ‘Pre-Crime’ division, John Anderton, is on the run and in a driverless car; except that it identifies him and locates him on the network. When he realises he has to run, a chase scene ensues and Spielberg dazzles us with shots of AVs moving at high speed in dedicated, elevated channels. Yet, they still maintain that satisfying vroom

[23] This is relatively recent terminology; the most up to date is the “circular car” that the World Economic Forum 2021 has coined, to integrate the AV into sustainability mandates.

vroom sound, suggesting that even in the future, the car will not loose its ‘feel’.[24][25] Perhaps what is distinct about *Minority Report* is the world of 2054 that exists in the smallest details, like the cartoons on the side of the cereal box, the newspaper that updates itself, and personalised advertising that jumps out at the protagonist. There is a powerful role that cinema has in generating visions of the future, and even more so when a powerful cultural influencer like Spielberg assembles scientists and engineers to help him create a future where we blend into the medium. David Kirby writes that “diegetic prototypes” are technologies that exist only in fictional world “ — what film scholars call the diegesis — but they exist as fully functioning objects in that world” (2010, p. 41), and thus demonstrate to us, the audience, the “need, viability and benevolence” of those technologies (Ibid). But media and influential technologists work together to amplify certain messages. Such ‘fictions’ can be a valuable jumping-off point to create legitimacy to institutional knowledge and naturalise particular kinds of ontologies and social orders.

From AI making knowledge through imaginaries and fictions that materialise in different ways, I turn to discussing how AI quite literally makes knowledge as computation.

[24] I did not find this to be the case on my test drive in Philadelphia, however, and an interviewee in California tells me that future AVs will not be able to capture the embodied, sensory, and affective satisfaction that comes from driving a car.

[25] Spielberg took great pains to make the year 2054 seem realistic and true to life; before the release of the film he spoke with the film critic Roger Ebert about his process: https://en.wikipedia.org/wiki/Minority_Report_(film)

Symbolic v Connectionist

The current wave of AI applications, and why we even have semi-autonomous vehicles, is thanks in part to the work of Geoffrey Hinton and his team.[26] In 2012, Hinton's team won the Imagenet challenge that had been running since 2007, thus effectively introducing deep learning techniques to the world, and that are now implemented on a large scale. Imagenet is a database of millions of images that are legible to humans, but not to machines; the challenge was to develop machine learning that could annotate, or 'recognise' those images. These object and facial recognition technologies enable the driverless car to 'see'. Since then, the field of machine learning has received substantial attention and testing, being applied in a variety of contexts. What enables this is the access to large amounts of training data. But, it was not always like this. AI suffered its infamous 'winter' because there as not enough data to advance either of approaches that were predominant in the late 1970s/1980s when a wave of AI research came to a standstill.

The most dominant early strands of AI that was developed were about modelling and replicating human intelligence into computers by various approaches to cognitive problem-solving of different kinds. The symbolic approach to AI takes a heuristic approach and is distinguished by two features: first, that language, its structure, semantics, and syntax, has largely shaped the understanding of cognition, and by extension AI computation (Smolensky, 2012 pp 308).

[26] Geoffrey Hinton is an influential scholar who works at Google Research, and at the University of Toronto. His research group at the University of Toronto has revolutionised speech recognition and object classification technologies. He is also the great-great-grandson of the logician, George Boole. https://research.google/people/GeoffreyHinton/

The brain is taken as a machine for formal symbol manipulation, and these symbols are semantically words of English. The second distinctive feature of the symbolic approach is that it relies on the formulation of systems of *rules* that connect things in the world with their representations in semantics and language. In relying on language and rules, the symbolic approach suffers from the fact that meaning in language is not complete, is partial to metaphor and complexity, and that a finite set of rules, ultimately, do not scale. It is not possible to arrive at rules that apply to large and complex situations; it would be akin to mapping literally everything. Herbert Simon and Alan Newell's General Problem Solver, and Logic Theorist machines, received attention as instances of how successful Symbolic AI was thought to be. However, the reliance on a language-based representational system conveyed through elaborate rules resulted in the eventual demise of this approach because it was unwieldy to map the complexity of thought and language without the computational power we take for granted today. For this reason it is known as GOFAI, or 'Good Old Fashioned AI' (Haugland, 1985, p. 112).

A second approach to artificial intelligence, referred to as connectionism, originates in the work of Walter Pitts and Warren McCulloch's 'neural nets'; and Frank Rosenblatt's perceptron. The connectionist approach assumes that a vast array of small, neuron-like computational units linked together can be encoded with information not unlike in the brain, and without rules needing to be specified. Connectionism emphasises learning from sample data from which a model of the world can be inferred statistically; in other words, from the bottom up. Early work on connectionist approaches flourished, but it was the popularity of Rosenblatt's perceptron that sparked a controversy. MIT scientists, Marvin Minsky and Seymour Papert, were instrumental in publishing a book called *Perceptron* that outlined why a connectionist approach would fail, even though they knew it would not . They effectively shut down this line of inquiry and diverted funding to Symbolic AI. At the time, AI was a high-stakes academic and funding game (Katz, 2020, p. 32-33). Eventually, the tables were turned when Symbolic AI failed because it was too cumbersome because it relied on elaborate sets of rules. Connectionist AI has now re-emerged as the dominant form of AI as we understand it now: machine learning, deep learning, neural networks, and all thanks to the explosion of data and computational power. As I will discuss ahead, these two approaches roughly map onto the machine ethics approaches: Symbolic AI is a lot like the top-down value alignment approach, and connectionism is like a bottom-up or hybrid approach. Both approaches are limited in and of themselves because they assume that all knowledge is data that can be accessed through by sets of rules. n .

Phenomenology, embodiment, robotics

Hubert Dreyfus was a key figure in the early history of AI who argued against the dominant symbolic paradigm: that humans do not think according to elaborate rules; we do not have elaborate libraries of associations that we call on when we need to work something out; we do not sit apart from the world and observe and reason. He argued that humans have a way of knowing how to deploy the relevant information at the right time. When we start off with a task that we are unfamiliar with, such as driving, we follow the rules in a fairly precise way;

but, as we become more competent and expert, we become more intuitive and reflective, and cannot actually describe how exactly we go about solving problems; we appear to just *know* (Dreyfus, 1987) Dreyfus' theoretical work adopts Heideggerian and Merleau-Ponty-ian phenomenological approaches to suggest that AI would have to be situated in a context of the world, and that it was shaping, and being shaped by the world. There was no sense in an AI that could identify objects through a set of narrow rules;rather its knowledge would be more robust if it engaged with these objects contextually. Dreyfus' work and phenomenology did not catch on for a number of reasons (see Katz, 2020 pp. 189-192) but influences from cybernetics, biology, and self-organising systems became important critiques of the extremely analytical and rule-obsessed AI of the day. The work of Terry Winograd, Fernando Flores, Lucy Suchman, Humberto Maturana and Francesco Varela, and Philip Agre, are important scholars here. I do not address their work in great detail here for it does not relate directly to the questions that motivate this research. Suffice to say, their works attempt to build on and go beyond phenomenology, to understand how embodied, social, knowing subjects can be in, and build, worlds.

Particularly relevant to this research in this brief tour of early Western AI technologists is the work of roboticist Rodney Brooks from MIT. For Brooks, the traditional AI approach to robotics had great difficulty in producing operational systems because instead of considering the problem of embodied interaction, they diverted their attention to formalising abstract thought through a complex system of rules to identify. For Brooks, the idea of the human brain was not the most apt metaphor for AI and computation; he emphasised interaction, situation, and embodiment instead (Bruder, 2020, pp. 4-5) Brooks decided that a robot would not have a centralised representation of the world, but would slowly build up its own internal cognitive map of the world it was encountering. So he broke from programming an AI system with a top-down representation of the world, and gave it the tools to form its own representations to learn about and navigate the world. He convinced much of the robotics community that the human mind should no longer be the blueprint for robotics, but that the capacity to cope with the real world should be instead. Robotics should follow the line of evolutionary complexity he argued, and only pursue the modeling of human intelligence once animal intelligence is achieved (Brooks, 1991). This work resulted in the invention of the Roomba, the 'autonomous' vacuum cleaner; and its parent company,iRobot, sold its technology to the US military to develop autonomous weapons systems. Lucy Suchman criticises this outsourcing of human "dirty work" to machines, first the work of cleaning, and then killing at a distance (Suchman, 2015, p. 15-16). I return to this point later, when I argue that now we see difficult moral questions being handed off to machines as in the imagination of the driverless car.

Machine learning, algorithms, and data

Machine learning is the computational form we are most often referring to when we think of AI and algorithmic technologies. Models, data, algorithms, and machine learning are all interconnected and co-constitutive software operations;I discus models at some length ahead. Machine learning originated in connectionist schools of AI, and relies on software that identifies patterns in data-sets as the basis for learning how to navigate aspects of the human social and physical geographies, as well as non-human ones.. All machine learning relies on data to be trained on. Training data however is a reflection of the world it emerges from; data about crime are not just data about which crime occurred, where, and who committed it; but are actually data about how crimes were recorded, and how crimes were classified. Something as simple as walking down a street drinking from a bottle of beer is a crime in some places, and is also just a way of life in another. It is not just that context matters, but that entire social and political worlds can be architected around such simple contextual facts.Thus, in a highly global context of technology production, the importance of the origins of data and how they are pushed into the architectures of machine learning cannot be understated. Because, once data is collected, then data are organised into coherent data sets, and must be labelled.e This is a thankless, tedious task that humans must do, or at least oversee, because computers can sense but cannot make sense of the world of data. Hence Amazon has become extremely wealthy and powerful through its Mechanical Turk platform that does such digital 'piecework' for scores of technology production centres around the world. It is on these cleaned and labelled data sets that machine learning algorithms work. Machine learningrefers to how different kinds of neural networks, essentially mathematical functions contained within algorithmic ensembles, identify the relationships between elements of a data set, and architect a model of these inter-relationships. This is where a lot of what is referred to as 'data science' or 'AI as statistics' comes into play.

Machine learning algorithms extract features of a data set by identifying the patterns of relations in datasets, and then translate those patterns into mathematical functions that are then used to parse *new* datasets. Hence, machine learning interacts in new worlds on the basis of what it was trained on. Pasquinelli and Joler (2020) visualise these practices in visual detail, and identify machine learning as a kind of cultural techniquein two senses of German media-theoretical *Kulturtechniken. Cultural techniques are* practices that make distinctions between inside and outside, pure/impure, male/female, human/animal and so on (Siegert, 2015, p. 11); and that generate concepts and ideas in the world (Op Cit, p. 14) Thus, machine learning and algorithmic decision-making, work as discursive actors, creating knowledge and reality (Pasquinelli and Joler, 2020). I discuss this last point in more detail ahead, in the context of building models, which are central to how machine learning moves from being a statistical technique to being a cultural technique that 'knows' the world.

Models and representations

We encounter a great deal of how we know the world through modelling, statistics and projections of these, from the spread of pathogens and the evolution of pandemics, to the rising temperature of the planet, to predicting which kind of person will repay their loans on time. The power to model reality through computerised models and simulations is a defining moment in computation and epistemology; the introduction of computer simulations as a 'natural' way to know things that were too big and too complex resulted in a new way of doing science: the management of the various cognitive activities associated with experimenting became supplanted by a simulation that came to take on an epistemic property, thus eliding the distinction between theory and experiment (Galison, 1998; Pias 2011) As a result, "analytic-ally intractable phenomena" situated computer science in a unique epistemological dimension of being a kind of vehicle for the production of truth and knowledge. So mathematical principles tend to be invisible but become the prime movers of value: "The belief that makes it possible for mathematics to generate value is not simply that numbers are objective but that the market actually obeys mathematical rules." (Poovey, 2003, p. 32) Derivatives and mathematical models 'become', in a sense, what the market is, and even form what it is, giving a quantified shape to something abstract, like risk. Similarly any kind of AI system is a statistical model (to recall the Ryan Calo tweet from the previous chapter) that will enact a particular reality into being.

Modelling risk is familiar to the automotive industry. In the 1970s, with the increasing use of computers, the development of computer based simulation and modelling was applied to crash testing and crash worthiness, thus outsourcing traffic safety to mathematics (Leonardi, 2010). Prior to this it was a practice to build cars and then crash them, and investigate the crash very carefully to understand what happened in the process; the insurance industry contributed to the costs of this crash-testing. Obviously, mathematics-based simulations and modelling were more accurate and cost-effective. A decade-long campaign by automakers resulted in what the Society of Automotive Engineers refers to as "road-to-lab-to-math": an industry-wide belief that mathematics based simulations are more cost-efficient than road and laboratory testing (Op Cit p.247) 'Road to lab to math' to assess 'crashworthiness' is about the co-evolution of industrial, technical, organisational and regulatory change in the automotive and adjacent industries; and was not necessarily something linear or 'natural' in understanding crashes (p. 246). Hence we have seen the establishment of knowledge-practices that are decidedly social and cultural events, particularly in the automotive context. Models and model environments can be valuable spaces to correct and test AVs. Some testing of AVs happens on open roads; but a lot of testing happens within the relative safety of a simulation. For example, within the growing big data industry that supports autonomous driving, Cognata is a company that builds entirely simulated world models for AV makers to test their products. They create “digital twins or entirely synthetic worlds for AV testing and validation” and “Fuzzing to quickly create new variations of scenarios” (Cognata 2020). The purpose of these simulations is for AVs to be tested and validated safely, as if playing a game but without having to do this ‘in the wild’ and with ‘fuzzing’ to create new and challenging scenarios. Thus the simulations built by Cognata will be just one reality inhabited by driverless cars. It is not necessarily *human* reality.

I offered a stripped-down and basic outline of how a model works for the purpose of this discussion. A predictive policing application is thought to be able to identify future crime ‘hot spots’. This system is extrapolating from the patterns it has already identified in the dataset (i.e. a slice of the world) it was trained on: that, for example, in location A, in a time period B, people of the type C, are found to be responsible for crimes of the type D, with weapons of the kind E. The people who make a predictive policing system also must have a clear question in mind: how likely is it that crimes of the type F might be committed by people of type B in location A? However, the collection of data about crime is a highly situated matter. Scholars demonstrate that some communities are overpoliced, and that histories of over-policing enter the taxonomies of crime and criminality, and thus how we collect data about crime. In other words, crime data is as much a snapshot of crime as it is about those who classify and collect data about crime. (Ref) Algorithms (which already constitute the model and are also later produced by the model) run through a cleaned training data set about crime occurrences, identifying the statistical distribution of values in that set, the patterns in the data, as it were; and then develop a mathematical model that captures these patterns. Models allow for the development of new algorithms that identify patterns in new data sets and make predictions on the basis of these patterns. If new data that this algorithm encounters matches the patterns it detects within the normal distribution of the dataset it was originally trained on, then the machine learning is *recognising* the *same* pattern. The algorithm is *generating* a new *pattern* or making a prediction when it uses past data to predict the likelihood of a crime taking place again. The model is thus creating a new set of realities, a "data derivative" (Amoore, 2011) of something that has *not* happened and *may never* happen and only *could* happen. This model is a snapshot of the world as known through data assembled at a particular time and place by certain methods, all of which leave imprints on the data (Gitelman, 2016). Machine learning algorithms are agnostic about the data itself; it does it 'know' that it is reading the details of a crime, or even what a crime is, or what people are.

None of this is linear and everything within this system can be optimised, or mathematically weighted to be more sensitive to something in a dataset, such as location or person type. Hence, models are already biased, not necessarily in the sense of being discriminatory, but in the sense of being *situated*. This does not minimise their impact, but necessitates guardrails and oversight in their use and application. Thus statistical models in AI systems are establishing relationships in the world. Computational decision-making is only eventually representing what computer architectures can capture about the world as understood by the architects of those systems and the affordances of models and software; it is not necessarily the *world as it is.* This mathematical model derived from mapping patterns in the dataset can be used to classify patterns in new datasets, and also make predictions based on a new dataset.

The transformation of data into the actions of a car deciding that an object on the road is a cat, a child, or a banana is what much of its autonomy hinges on, and the socio-technical infrastructures that enable this transformation. As I will discuss ahead, AV crashes have taken place because of failures in the computer vision's statistical model, failures to recognise and classify objects correctly in the environment around the driverless car. We find repeatedly, that the story of an emergent autonomous or AI system is one of apprehending the gaps that open up between models of the world that exist inside AV systems, and the world itself. This is because models are not always up to date, even though there is substantial effort made to build infrastructure that will 'push' real time information about the world to the model accessed by the driverless car, via the cloud. Models may not be architected on data that is representative of diverse contexts. Further, models only mediate between the world and its *representation*, so much so that in performing vast numbers of calculations therein, a "qualitative leap of inexactitude" occurs (Pias 2011, p. 32-33). We see this inexactitude emerging from the socio-material practices underlying the building of AI's models. Statistical models running on big data that have been collected, coded, and labeled according to the subjective understanding and situation of particular people and institutions, make distinctions between those who are more or less likely to return to criminal activity, or who should be hired, or are likely to default on their loans and so on. Thus statistical models establish particular orders through which we come to know the world (Porter 1996). To return to the example of predictive policing, above; the reason why predictive policing applications require oversight (if not being altogether banned) is because the multiple moments of inexactitude could have very serious repercussions for people if wrong.

The algorithmic model of one part of the world ends up re-shaping the world when it encounters another part of the world. Consider what is going on inside the computational system, at the level of computational ontologies. In computer science, 'ontology' refers to the common understanding of the world through organisation of it into categories, and the meanings given to representations of things, so as to be intelligible to machine systems. They are "structural frameworks for knowledge representation about the world or some part of it, which mainly consists of concepts (classes) and the relationships (properties) among them.“ (Zhao et al, 2015) Ontologies are critical to driverless car software (or its 'intelligence') to distinguish between different objects perceived in the environment. This naming and ordering of objects into categories, and valuing them, directs the car's computer vision software to know how to react in various situations. Breakdowns in the tagging of images in computer vision setting in motion the conditions for a crash. One important case occurred in Florida in May 2016 when a test driver was killed when he did not take control from the Tesla that was in Auto-pilot mode and that drove into the light, white side of an 18 wheel truck, mistaking it for the early morning sky. The driver did not respond to alerts to take over, and eventually had three seconds to respond before impact (National Transport Safety Board, 2017; Tesla, 2016). The training data for the car's computer vision system had not been adequately trained to distinguish shades of colour that might exist across the sky, or a painted truck. I return to this point later.

It turns out that it is not just that light coloured trucks are a problem for computer vision systems; research finds that object detection by computer vision systems used in AVs perform poorly "when detecting pedestrians with Fitzpatrick skin types between 4 and 6. This behaviour suggests that future errors made by autonomous vehicles may not be evenly distributed across different demographic groups." (Wilson et al, 2019, p. 1) In other words, people with darker skin tones are less likely to be clearly identified by computer vision used in AVs. The problem of darker, and female phenotypes being misidentified or not seen by computer vision has a precedent, it is not new (Buolamwini and Gebru, 2018). This happens because the training data sets in computer vision systems are not standardised on darker or female phenotypes; the standard tends to be lighter and male phenotypes. This illegibility becomes a slice of the world that the computational system recognises. What fails to be recorded has implications when that model becomes the standard for when and how that computational system negotiates the world. The measurement of human bodies to be legible to machines and for machines to be efficient, only amplifies measurement techniques in knowledge-making.

Error, Games, Learning

AI's intellectual and computational history is entwined with games of Chess (Ensmenger, 2012) and Go (Bruder, 2020); and the Imagenet competition[27] has been significant in bringing machine learning technologies to the level they are at today. While the Turing Test is the most famous 'test' we associate with AI, Alan Turing himself only asked if a machine could pass for human by the achievement of language-based, verbal tasks, and without establishing which tasks those were. Yet, the 'Turing Test' has become widely understood as a simulacrum of human intelligence writ large. In the popular film, *Ex Machina (*Garland, 2014), the test is inverted: the synthetic intelligence, Eva, proves her humanity by displaying the worst kinds of human behaviour in a struggle for her own survival: cunning, deception, malice, and violence. In 1987 a computer was 'intelligent' if it could play Chess well enough to beat a Grandmaster; in 2021, it is expected to drive a car. In fiction and in life, the intelligence of artificial intelligence has been measured by qualitative and quantitative tests attenuated to the scale of human cognitive capacities. The Voigt-Kampff Tests[28] in *Blade Runner* (Ridley Scott, 1982) distinguish 'humans' from 'replicants' on their ability to feel empathy. The 'real world' has always been the target of AI development, and "turning to [it] would challenge the field to encounter and manage that world's full complexity." (Ribes et al, 2019. p. 286) This is what the dream of Artificial General Intelligence is. The autonomous vehicle navigating the 'real' world presents a contemporary real-world challenge; in its earliest iterations, the AV

[27] Instituted by AI scientist, Dr. Fei-fei Li, in 2009 and running till 2017, Imagenet was a computational challenge to develop bring rich, semantic, and accurate human-level annotations to a machine's 'understanding' of an image. It was inspired by a classification system called Wordnet, a lexical database of word meanings (Princeton, 2010). See https://www.excavating.ai/ for a critical overview of Imagenet.

[28] The Ridley Scott 1982 film, Bladerunner, opens with a fictitious empathy test called the Voigt Kampff Test, used to identify 'replicants' who look human but are in fact, synthetic intelligences; they are highly intelligent and strong, but lack empathy.

emerged from DARPA's *on-road* challenges (DARPA, 2007). AI's intelligence has always been a shifting target as the human notion of our own intelligence has evolved.

In all testing and development, error and feedback are essential to improvement. With more and better data, and adjustments based on learning from errors, a machine learning system can only get more refined and accurate. It does not 'know' in the conventional sense what decision it is making or what the outcome of it is, but it is likely to keep becoming better at whatever it does. And, it is the messiness of the world and the impossibility of knowing everything about it and yet reducing what is known to rules, that makes algorithmic techniques and machine learning more robust. AI needs to be spoofed and challenged to get better (Amoore, 2020, p. 121). This is the point of AI that plays games to get better. For an AI to become strong, it must adopt rules from the world and test out these rules in the same world. However, the facts of the world always changing, and *not* being a set of unchanging paths to a valued goal stymies navigating that world, unlike a game. This is why advances in the game of Go are significant in the development of machine learning becoming more 'intuitive' and going past the limits of rigid top-down or bottom-up programming. In the 1980s, Deep Blue was an IBM computer program that was brute-force programmed to play Chess and win, eventually beating the world champion, Gary Kasparov.Nearly every possible permutation and combination of moves that could be made on a 8x8 board with 32 pieces was identified by the minimax algorithm developed for this purpose, but nothing about this development actually moved AI forward; it only resulted in making computers that played Chess against Gary Kasparov very well (Ensmenger, 2012, p. 24) The development of AlphaGo, however, marked a significant shift in the development of AI. A more complex game with millions of possible moves, Go demanded an entirely different approach because it does not rely on methods or rules, but on an intuitive and sensory feel for the board (Bruder, 2002, p. 6). So the AI research lab, Deep Mind, built AlphaGo which 'looked at' hundreds of millions of existing games of Go, and discerned the most successful patterns of combinations of moves that achieved a winning outcome. It was also designed to simultaneously calculate and evaluate the next move while verifying those moves against past experience –and of course, it is a very powerful computational system. What is interesting about the AlphaGo game is that it appears to have a 'situational awareness' for the board; in other words, it actually plays as if intuitive (p. 129-130) There is a famous moment that is repeated in all the news stories about this historic human v machine moment; that the top rated world champions it has beaten, Fan Hui, and Lee Sedol, both marvelled at the program's moves; that they had never encountered moves like that p. 130). That's because it was not really a human playing; nonhuman computation is therefore unique, and how close (if at all) to human thinking and intelligence it really is is an open matter.

These key ideas in the intellectual history of AI, and how machine learning works to model the world e are relevant to the ethical apparatus in terms of their historical influence on decision-making and problem-solving expected of the computational systems imagined to make future automobiles 'autonomous'. Next, I switch gears and turn to theoretical approaches about the social shaping of technologies.

Society-technology

This study of how knowledge, language, and representations about *autonomy* and *driverlessness* emerge through social and material ‘shaping' these. The study of the ‘social shaping’ is not just a matter of “turning the arrow round so that instead of the natural sciences explaining social phenomena, a social explanation of molecules, cells, or bodies is being presented.” (Mol, 2002, p. 157) Instead, the purpose is to show that all of these practices of knowing and knowledge-making about AI and autonomous technologies and their world(s) hang together and must be reckoned with together. These technologies are always already social and cultural. This is a challenging idea because we assume technology artefacts just emerge *naturally*. Or, as discussed in the case of Asilomar, AI ethics, and nanoethics, earlier, the social is also constituted by broader, political-economic, regulatory, and intellectual agendas. Additionally it is hard for many to consider that the social and technical are conjoined and co-constitutive, and not just added on top of the other; they always have been part of each other, and always will be. I reprise what Bruno Latour said decades ago to situate my intention more clearly, that this is not merely an attempt to sketch out “ a hybrid object, a little bit of efficiency and a little bit of sociologizing, but a *sui generis* object: the collective thing…” And he goes on to say that there are “missing masses” in our social theories of technology production of software itself (cf Latour, 1992, p. 254). The car as a sociological innovation is already well-understood, and is in fact what is being leveraged in particular future socio-technical imaginations.

Of all the words about AVs with multiple, conflicted meanings, perhaps ‘agent’ is the most challenging to address because it is so commonly used.

“But does it have ‘agency’?”

In this section I sketch out the different theoretical influences on ‘agents’ and ‘agency’. I identify what kinds of agents autonomous vehicles *are*, and what they are *not*. There are differences in how the word ‘agency’ is used and applied across the different communities and industries that are involved in shaping autonomous driving and ethical decision-making. ‘Agent’ is a word commonly used by AI researchers and roboticists to refer to both embodied *and* entirely digital *and* embodied digital technologies they build. An ‘agent’ can be a software program. A Latourian might conceive of a car door, a screen, or the car itself as an agent. A question that I have been asked many times in the course of this work is if I think driverless cars are moral *agents*, and what this means. This question conveys an anxiety that belongs to both philosophers and lawyers: are driverless cars computational nonhuman agents that could be held responsible but not accountable for their actions? In philosophy, the word ‘agent’ is used in a very precise and technical sense. For example, a moral agent is different from a moral patient in that the former has moral responsibilities and obligations, and the latter refers to those who might have agency but cannot have responsibilities and obligations. So, a baby, an adult in a coma, and even a human-eating leopard might be moral patients but are not moral agents. In a machine ethics approach, this distinction becomes valuable in considering nonhuman and technological systems like autonomous vehicles that might have agency, but not responsibility or obligations.

In the context of autonomous systems, if a complex agent constituted by a complex infrastructural system breaks down, then how do we identify the single point or moment of culpability? This is the question often asked by lawyers and regulators, and particularly in the 'ethics of autonomous driving' discourse related to speculative car crashes (discussed in chapters ahead). Who is responsible for how the system is architected to make decisions about how to react in the case of an accident? If agency is widely distributed into multiple agencies, then how do we think about responsibility and accountability when things go wrong? As, the philosopher and applied ethicist, Deborah Johnson argues, "using the metaphor of autonomous agents may be setting the scene for a collision with moral and legal notions and practices of accountability." (Johnson, 2011, p. 63)

James Moor's categorisation of *ethical agents* brings together questions of design and context. A system with a 'high ethical sensitivity' is a 'full, explicit ethical agent' (Moor, 2006), and Wallach and Allen refer to an 'Autonomous Machine Agent' (AMA) with 'autonomy' (2009, p. 33-34). Moor categorises ethical agents into four types (Moor, 2006; Winfield et al., 2019). In his schema, 'Ethical Impact Agents' are "any machine that can be evaluated for its ethical consequences", and the 'machine' can be anything from a chatbot to a facial recognition system. 'Implicit Ethical Agents' are those machines that are designed to avoid unethical outcomes, such as an automated teller machine (ATM), an auto-pilot setting, and even a pharmacy that has a duty of care to sell only safe and licensed products. Explicit Ethical Agents' are machines that can reason about ethics; and 'Full Ethical Agents' are machines that can make explicit moral judgments and justify them. Moor's categorisation is popular in the machine ethics domain, which I refer to in more detail in further chapters.

The interest in 'agents' is to do with assigning responsibility and accountability in the case of accidents and failures. An industry that strongly influences autonomous driving technologies is aviation, and is one that contains an extensive industry of accident and crash accountability. Accountability for aviation crashes identify a deep entanglement between human actions and the perceived agency of technologies. Finding what triggered or precipitated an accident is much easier than it is to identify where it actually started (Brown, 2006) to the point where we cannot eventually see the entire system of aviation — of humans, nonhumans, technologies, organisation, knowledge, social systems, commercialisation etc. — as separate in any way, but as entirely 'compounded' (pace Bruno Latour). Similarly in her investigation and analysis of the 1986 Challenger space shuttle tragedy, Diane Vaughn names the 'normalisation of deviance' as the culprit rather than blatant corruption, malafide intent, or negligence (Vaughn, 1997). The 'normalisation of deviance' refers to a slow and gradual loosening of standards for the evaluation and acceptance of risk in an engineering context. The O rings on the rocket boosters of Challenger that broke on an unusually cold January morning did so despite considerable knowledge and evidence of their questionable performance in low temperature conditions. The space shuttle's launch date was also repeatedly delayed for this very reason. Every single organisational department and technical or managerial community at NASA made a slightly different assessment of the risks from the weak O rings; this was the ‘normalisation of deviance’. Vaughan’s detailed ethnography of the disaster thus defines agency as distributed, and cultural – as is accountability.

This work argues that driverless cars *are* agents that shape, create, and influence worlds, and in the sense of Latour’s ‘congealed morality’ of speed bumps, seat belts, and heavy hotel room keys that enforce particular kinds of behaviour because of their design (1992); and similarly in Langdon Winner’s politics of the design of technology (1986). In this work, I assume that technologies are already moral in their conception and development. Thus they “help humans be moral agents”, because morality emerges from the “complex and intricate connections between humans and things, which have moral agency as a result rather than as a pregiven ontological characteristic” (Verbeek 2017, p. 84). This necessarily implies accountability in the process of laying bare how these agencies are architected. In their important work on lethal ‘autonomous’ weapon systems, Suchman and Weber (2016) convincingly argue that agency is not a fixed attribute but is constantly being made and re-made through social, cultural, and material conditions, actors, and practices, and their intra-actions. Hence, this study identifies how these social and cultural practices coalesce and bring agency to these artefacts. The discussion of agency in terms of artefacts such as the driverless car is helpful because it resists notions that locate agency inside the artefact itself, and urge us to seek the meaning of autonomy, accountability, and intelligence, through deeper relationships in and with the world. This is exactly what two important theoretical influences - the socio-technical, and the technosocial - that influence this work argue. :

The socio-technical

There is a long-standing discussion about values and technology, that technologies are not neutral, that they carry the imprints of the context of their design, and continue to shape and be shaped by the world. Socio-technical and socio-material approaches introduce necessary discussions about how we understand computation in large-scale, industrial systems like AI and autonomous systems. A socio-technical approach is a body of scholarship against deterministic, top-down, 'effects' of technology, the social shaping of technology ('SST') field, which insists that we must understand and investigate technologies in terms of the patterns and processes of their development and use, so we may understand their role in society. (Mackenzie and Wajcman, 1999; Bijker and Law 1994). An important idea within this identifies technology as political in its design and implementation (Winner 1977, 1980), as "politics pursued by other means" (Latour 1988) and thus either preserving or altering social relations. As renowned media studies and technology studies scholar, Judy Wajcman, puts it in her discussion, feminist approaches understand technology as both a source and consequence of social (gendered) relations, thus making it impossible to separate one from the other (2010).

A socio-technical system is: "a recognition of a recursive (not simultaneous) shaping of abstract social constructs and a technical infrastructure that includes technology's materiality and people's localised responses to it" (Leonardi, 2010, p. 247) The 'socio-technical' has a decades-old lineage in organisational studies to understand the role of technology (i.e mechanisation and automation) in work. It comes from the Tavistock Institute's 1951 study of how new kind of mechanisation in coal mining in Britain influenced the re-shaping of human interaction with technology, and other humans. The lesson from the Tavistock study is that more is disrupted than just the achievement of tasks when technology is introduced; power relations, communication patterns and group dynamics are too. A socio-materialist approach, similarly, discerns the imbrications in a socio-technical system; 'imbrications' refers to how roofs are tiled in layers, inter-locking to create a dense and stable lattice. Technologies may be constituted of materials that stay the same, say an interface or code, but whose functions change through the constraints shaping their use, and the affordances that are stretched to produce new realities of the context of use. In the context of this study, I approach the imbrications of human ideologies, values, culture, and labour with computation and automation within the automated infrastructure of driving. This approach makes it very difficult to see technologies as separate, added-on, or even merely 'hybridised' with human action; and thus requires a different understanding of the new entity that is formed.

Agency from a socio-technical perspective is challenging, and can present as uncertain to the requirements of law and regulation. The example of the auto-pilot setting in a semi autonomous vehicle is instructive. I sketch this out briefly here and will return to it in more detail in future chapters. Auto-pilot is an inherently socio-technical system; future regulation, law, and policy must be aligned to addressing it as such rather than as software or engineering. By its very definition, auto-pilot relieves the human of having to pay attention to the road; However it requires human intervention when it cannot address a particular situation on the road, or lacks the computational inputs required. Crashes have taken place when human drivers, seduced into inattention that an efficient and smoothly-running machine fosters, have to scramble to regain control of a system that falters in making sense of its environment. It is socio-technical in terms of multiple interactional dynamics: how humans become inattentive at an affective, motor, and physiological level; the reaction time to regain control; the alerts in the system that signal the need for human intervention; the communication between human and machine; and how eventually driving itself becomes quite different when in auto-pilot mode. In other words, driving with auto-pilot is not the same as driving without. Ahead, I will narrate the experience of a test-drive in a Tesla Model S where this was palpable. Hence 'agency' cannot be assigned to a thing but "is both enabled and bound by the material and discursive limits of digital architectures while also immanently reconfiguring social borders and boundaries and re-shaping and re-making the bodies and actualities of those orders" (Dixon-Roman, 2016, p. 487).

The driverless car as a big data infrastructure

The ontology of the driverless car as a big data infrastructure identifies the complexity of technical that humans are enmeshed in. What Drive.Ai and BMW refer to as the AV's 'brain' is in fact a vast data-infrastructural network spread over multiple commercial, regulatory, legal, and cloud geographies. The material infrastructures within the emergent AV render it as data platform in itself (Alvarez Leon, 2019) that runs AI technologies like computer vision and automated decision-making based on multiple sources of data processed through machine learning. Metaphors of a computational brain are materialised by a raft of profitable software companies that are building autonomous driving based on AI technologies. The venture capital firm, Comet Labs, detailed the 263 companies working on driverless car technology, many of them small and relatively unknown, like Actility and Braiq but producing important components, sitting next to bigger names like Google and Uber (Stewart 2017; see image below). The AV exists in and as a 'formidable' 'intelligent vehicular grid', a big data-infrastructural platform. It is comprised of sensors capturing

and processing data about the environment, cameras, radar, Lidar[29], myriad data processing functions including machine learning, object recognition, tracking, and coordination, mapping and localisation systems, machine-readable road signs, networking and communication architectures including vehicular cloud computing, computer vision, machine-learning based risk and uncertainty assessments, hard and soft telematics, and driving style analysis among others (Gerla et al., 2014; Yurtsever et al., 2020). It also includes fleet and traffic management, sensor software, 'emotion, fatigue, alcohol detection+distraction avoidance', 'rapid prototyping, 3D printing, modularisation and open source'. At the end of 2017, this was the "hotspot for research and investment in Silicon Valley".

[29] LIDAR, or 'light detection and ranging', is like radar in that it is used to map the environment but it produces more granular and vivid maps. Primarily, driver-less cars have to learn how to identify objects so they know how to respond to them. LIDAR has become fairly well-known now because it creates vivid, colourful three-dimensional depth maps of the environment; these often crop up in image search results for 'driverless car'. Maps for driverless cars require granular detail. Thus the precision and accuracy of driver-less cars comes from computer vision software that 'learns' appropriate driving behaviour — merging, driving around construction zones, waiting for pedestrians — through exposure to large datasets that its algorithms are trained on.

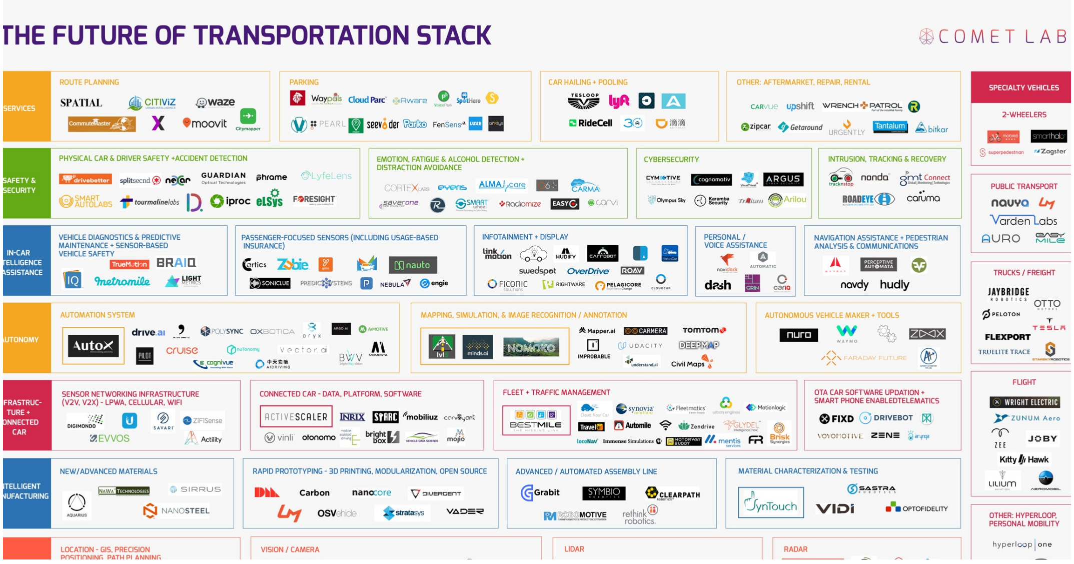

THE FUTURE OF TRANSPORTATION STACK
COMET LAB
SERVICES
ROUTE PLANNING
PARKING
CAR HAILING + POOLING
OTHER: AFTERMARKET, REPAIR, RENTAL
SAFETY & SECURITY
PHYSICAL CAR & DRIVER SAFETY +ACCIDENT DETECTION
EMOTION, FATIGUE & ALCOHOL DETECTION + DISTRACTION AVOIDANCE
CYBERSECURITY
INTRUSION, TRACKING & RECOVERY
IN-CAR INTELLIGENCE ASSISTANCE
VEHICLE DIAGNOSTICS & PREDICTIVE MAINTENANCE + SENSOR-BASED VEHICLE SAFETY
PASSENGER-FOCUSED SENSORS (INCLUDING USAGE-BASED INSURANCE)
INFOTAINMENT + DISPLAY
PERSONAL / VOICE ASSISTANCE
NAVIGATION ASSISTANCE + PEDESTRIAN ANALYSIS & COMMUNICATIONS
AUTONOMY
AUTOMATION SYSTEM
MAPPING, SIMULATION, & IMAGE RECOGNITION / ANNOTATION
AUTONOMOUS VEHICLE MAKER + TOOLS
INFRASTRUCTURE + CONNECTED CAR
SENSOR NETWORKING INFRASTRUCTURE (V2V, V2X) - LPWA, CELLULAR, WIFI
CONNECTED CAR - DATA, PLATFORM, SOFTWARE
FLEET + TRAFFIC MANAGEMENT
OTA CAR SOFTWARE UPDATION + SMART PHONE ENABLEDTELEMATICS
INTELLIGENT MANUFACTURING
NEW/ADVANCED MATERIALS
RAPID PROTOTYPING - 3D PRINTING, MODULARIZATION, OPEN SOURCE
ADVANCED / AUTOMATED ASSEMBLY LINE
MATERIAL CHARACTERIZATION & TESTING
LOCATION - GIS, PRECISION POSITIONING, PATH PLANNING
VISION / CAMERA
LIDAR
RADAR
SPECIALTY VEHICLES
2-WHEELERS
PUBLIC TRANSPORT
TRUCKS / FREIGHT
FLIGHT
OTHER: HYPERLOOP, PERSONAL MOBILITY

FIGURE 4. THE FUTURE OF TRANSPORTATION STACK BY COMET LAB, CITED IN STEWART (2017) MAPPED: THE TOP 263 COMPANIES RACING TOWARD AUTONOMOUS CARS. WIRED 10 MAY 2017. AVAILABLE AT HTTPS://WWW.WIRED.COM/2017/05/MAPPED-TOP-263-COMPANIES-RACING-TOWARD-AUTONOMOUS-CARS/

Thus, the emergent driverless car might be thought of as a computational and data platform, or a ‘data assemblage’ (Kitchin and Lauriault, 2018). Automobility has long been considered hybrid and dynamic, a dense infrastructure of infrastructures, a system of diverse institutional forms from manufacturing and selling automobiles, to highway and “gasoline delivery infrastructures, traffic rules, parking structures, licensing procedures, and sundry regulatory authorities.” (Chella Rajan, 2007, p.79) Similarly Dant’s ‘driver-car assemblage’ summons the social, cultural, historic, and industrial worlds co-existing and co-evolving: “neither a thing nor a person; it is an assembled social being that takes on properties of both and cannot exist without both…The driver-car is socially embedded as that is not going to be foregone or forgotten easily.” (2004, p. 74). Dant’s assemblage is made by “a system of affordances, actor networks and embodiment”, and is thus also of “human design, manufacture and choice” (Op Cit, p. 62) within the automotive-industrial complex. Thus, whether we view the driverless car in terms of its software ‘brain’, or its automotive-and automobility-infrastructures, or both, we are encountering large-scale relational, complex and distributed social-technical infrastructure that is also human. As big data infrastructure, it starts inscribing the world through the logics of its computation; Computer vision, mapping, and human-machine interactions in the auto-pilot setting are distinctive cases in this regard that I take up in detail in this work. Framing the human-machine relationship in the AV in terms of 20th century histories of safety engineering in automated systems, particularly aviation engineering, is problematic, and will continue to stymie AV regulatory policy. Because of what is being regulated, and who is in control, such questions are not as easy to answer. Thus the socio-technical is the re-working and re-shaping of human and technological so that they emerge as *entangled*, but not necessarily as *one*.

The technosocial

If the socio-technical is a perspective on the intimate and fluid interactions at the level of human eye, hand, and attention with the digital to create something entirely unique, then Arturo Escobar's discussion of the technosocial (1994) offers a perspective on this that is just as intimate, but on a different plane. Escobar's development of technosociality and biosociality, taken together to mean 'cyberculture', refers chiefly to "new orders for the production of life, nature, and the body" (Escobar, 1994, p 214). At this exact point in his famous essay, *Welcome to Cyberia*, he is referring chiefly to biotechnologies but also mentions AI as similar and as part of this formulation. If 'Welcome to Cyberia' now feels quaint, it is because it is, and at that time, the internet was still new and uncertain. We might feel a similar way about AI 27 years into the future. The technosociality he sketches out identifies the challenges to the study of how emergent science and technology are re-shaping discourse, knowing, culture, and social formations, marking the limits of existing theoretical frameworks and calling on his discipline to engage with new kinds of cultures and societies of the digital. The technosocial positions the social (rather than singularly, just the human) as constituted by the technological in the excess of information and the anonymisation of individuality into ‘nodes in a network’, and hence the question of what constitutes the real becomes one that is asked in all seriousness. Forit genuinely feels as if everything is virtual, and capable of virtualisation. If this became increasingly palpable with the emergence of social networking sites, then their transformation to platforms through mobile telephony, and the explosion of data that we now refer to as big data, firmly put us in the realm of the technosocial. Writing in 1994 to his home community of anthropology, and with extensive responses from his peers, Escobar lays out how at that time, emergent technologies of the internet and biotechnology were generating new questions about the constitution of society, the body, and reformulating our relationship to nature, to machines, and thus to human subjectivity itself. Questions of disruption and the re-ordering of social life and social formations felt as intense in 2014 as they do now, suggesting, among other things, that perhaps what is of value is to examine the new kinds of technosocial conditions and subjectivities being created and re-shaped (Ito and Okabe, 2005).

Long before a ‘fully' autonomous vehicle comes along, if it ever does, the fact is that our everyday existence is *already* mediated by a number of moral machines and ethical robots that are automating complex decisions at scale in sensitive areas of social and personal life. We don’t always think of them as ‘machines’ however, although they are; we know them by different names: ‘algorithms’, ’facial recognition’, ‘chat bots’, or, ‘recommender systems’. While there are different assemblies of similar technologies[30] underlying each of these systems, this work attends to the discursive positions emerging around what it means to shape and manage computational decision-making that have value-laden effects in the world. The ethics of autonomous driving has found public attention in a moment when there has been an all-round greater awareness of algorithmic harm, lack of transparency, and oversight of algorithmic systems. This is accompanied by the awareness that contemporary legal, social, and political systems do not have mechanisms by which to account for, and reconcile to, the social implications of algorithmic and automated decision-making and knowledge-making. For one, the technologies many of these regulatory systems were architected for are now being transformed, integrated, fused, upgraded, black-boxed, or automated beyond the limits set by the law and society. And with this comes transformation in social spaces, relations, and bodies. Consider: automated weapons systems, smart cities, automated hiring systems, targeted electioneering, machine learning integrated into medical diagnostics, automation and algorithmic management of global logistics and supply chains, biometrics, speculative financial algorithms, to name just a few. Further, trying to effect changes in these systems, even at a personal scale, can be challenging and fraught. For example, in trying to change the kinds of digital news or culture we consume, it is hard to know that we are not already always going to remain in a filter bubble and algorithmically shaped to want and like certain things, or if we can always trust the news and culture we are being directed to; thus we need to develop

[30] Chiefly: environmental sensing and capture of data about the environment; cloud computing and the extensive planetary infrastructure of cloud based services; computer vision systems; natural language processing; varieties of machine learning and the various statistical operations underlying them; the mathematical optimisation functions that derive analytical value from data; systems of big data capture; ontologies of representation and categorisation of data; segmentation of human audiences through algorithmic targeting; and the business, policy, organisational models that sustain and thread these all together.

new literacies and situated digital awareness. The convergence of these technologies are embedded in our everyday lives and interactions, and future lives, in myriad banal and experimental ways. They are also profoundly re-shaping many dominant decades-old, centuries-old institutions, social practices, and practices of the self. As I discuss later in this thesis, the place of humans in the multiple infrastructures of autonomous driving are constituting the 'human factor' of these technologies, but within the paradox of having limited control – and yet putatively having to be entirely in control as the case of auto-pilot reveals. The conditions of remediation of space and time through the profusion of maps and computer vision technologies in the DC are dissolving all territory into a map, one that is constantly being uploaded and updated from the cloud in real time. 'Space' is constantly being made and re-made at the speed of local bandwidth, and in terms of the knowledge shared by a fleet of autonomous Teslas. The socio-technical that lies at our finger tips, telescopes into the technosocial, constituting new subjectivities that we do not yet know how to name, even less to regulate or control. 'Ethics' becomes incredibly challenging to comprehend. Escobar's exhortation is that what we might do as scholars, is closely map out the contours of these new constitutions.

Ethics, values, and the design of technologies.

If the social and technological are so deeply entwined, then this raises questions for how values and ethics become part of the technologies we build. This is where multiple fissures are revealed in the relationships between disciplines in quite fundamental ways, as to what technologies are and how they are constituted. Is it something that gets *designed in* like James

Moor's implicitly ethical agent that has values distinctly baked into design, thus directing behaviour, like a shopping cart whose wheels lock when it goes beyond a certain distance from the supermarket? Jason Millar might refer to this as a paternalistic top-down "moral proxy", and ask how such proxies can be more thoughtfully designed to complement the human to ultimately effect moral behaviour (2015). In digital terms this manipulation of behaviour is baked into a variety of interface design techniques that are collectively known as deceptive patterns: intentional efforts to keep platforms and apps 'sticky' and difficult to leave.[31] The other perspective on 'ethics', is the machine ethics approach that intends to construct computational systems in terms of rules based on existing ethics and value frameworks; this is what Moor refers to as an explicit ethical agent, that makes ethical decisions. But aside from the design of technologies itself, there is a body of work that identifies the role of the social and cultural as part of how technologies are designed and built; so, it is not necessarily about putting all the ethical eggs in the basket of literal computational and technical design, though there can be elements here as well. There is the case, as raised above, and that will come up for discussion later, about how and if giving engineers 'ethics education' might enable more equitable and sensitive technology design. Some might argue that this a fairly top-down approach in itself. The social and cultural context play out as the socio-technical, socio-material, and technosocial, creating the enabling conditions for technology design and development understood as sociological and cultural processes, and not just 'tech'. Systematic and systemic prejudice infuses all of society, and hence it should be of no surprise that social products, including technology come 'pre-loaded' with bias (Benjamin, 2019). Thus, Katie Shilton (2012) identifies "values levers" as being intentional practices in the design process where

[31] A very thorough list of deceptive patterns: https://www.darkpatterns.org/

values come up for discussion. In her work, she reports on how a technology design lab negotiates the privacy and surveillance implications of an app they are building. She finds that what is more effective in integrating ethics into technology is the emphasis on intention that translates into money, and time, to make ethics and values part of the conversations designers have in the process of design. It is the time and money to have dedicated meetings to researching and reflecting on these topics. Batya Friedman and Helen Nissenbaum (1997) have been frontrunners in thinking about the social context of the value and design of technology. Nissenbaum, later, (2005) lays out a thoughtful schema of the technical, philosophical, and empirical modes when addressing the role and place of values in technology design. These three modes bring together the practical realities of design as well as testing if that design worked (the empirical mode) by assessing its place in the world. The philosophical mode is moment of evaluating the push and pull of different approaches to values, the trade-offs in adopting one set of values and not another. Between the extreme of top-down paternalisms in design, and the modes Nissenbaum lays out, there are different socio-technical iterations underway. So, from the social, material, and cultural shaping perspectives we might understand that ethics and values in technology are in the interfaces of use; so, returning to Barad's apparatus that extends further into the world around the device, we might imagine that ethics and values act as prompts and provocations in many different theatres of design, imagination, and implementation. It might include thinking about the very place of the technology itself, how it will impact landscapes and communities, and even bring into focus who those communities are, particularly the most marginal. While these impact assessment happen all the time, the design of algorithmic and AI-based technologies is of particular interest and concern because of the consequences for the technosocial that they engender that are not always within human control, as shopping carts are. In the next section, I introduce the case of algorithmic bias as way to focus on a particular discursive tension in the matter of ethics and values in technology design that reflects on the apparatus as device and as discourse. But before that, I reiterate that the wider social contexts are not just techno social but are also intimately socio-technical, and decisions made somewhere in the distributed networks of the apparatus come to bear on the body and being of the user.

Why was it that for decades African Americans and Black people did not appear as they are in photographs, but were depicted with features exaggerated or washed out? Lorna Roth finds that analog photography's default technical systems social values we subtly coded in: deficiencies that emerged in film emulsion, studio lighting, and TV lighting were based on "a global assumption of "Whiteness" embedded within their architectures and expected ensemble of practices, so the default was White skin and features. Anyone who was not White being photographed on certain kinds of film was considered a deviation in technical terms (Roth, 2009, p. 117). This adherence to an unspoken and unwritten technical norm is how bias has been integrated into technologies, in this case, into the "physics and chemistry" of photography and film emulsion that did not limit a more dynamic range, but merely exhibited a series of cultural choices made. The change that took place in photographic film emulsion and lighting technologies allowed us to think that these biases might be corrected, says Roth, but only if we explicitly acknowledge them. With the emergence of digital technologies and colour balancing options to make different skin tones and colours look good in post-production, the situation has evolved even though the material conditions of the lives of racially marginalised communities have not. Recent films and television shows have to intentionally think about how to make darker skin appear flattering and beautiful on camera (Lewis 2019); but, Roth argues, a subtle and not-so-subtle preference for lightness remains. What the story of this bias indicates is the distinction between values and ethics as explicit outputs, and values as implicit in the design of technologies which would result in more equitable designs and contexts of technology use and application; and to understand that technologies, humans, and the world are in constant interaction and processes of mediation (Verbeek, 2011; Ihde, 1995). Thus, to make ethics a matter of computer architecture rather than society, to decouple ethics from the design of technologies-are all inherently ethical matters in themselves (Parvin and Pollock 2020). Thus we might say there is a distinction between an ethics *of a* technology and an *ethical* technology; so the *ethics of x* and *ethical x.* The 'ethical apparatus' calls out this difference, and underlines the intention to architect an ethical machine by pointing to the myriad social, political, epistemic, historic conditions that emerge around and shape the machine as well.

The FATML case: Can values be designed in to technology to make them 'ethical'?

'Algorithms' have become a subject of focus and study as investigative journalism and academic research show that big data collected on scale and animating algorithmic decision-making are shaping personal, interpersonal, social, and political life. Algorithmic discrimination and manipulation takes place in fields as diverse as the curation of news (Katzenbach and Ulbricht, 2019), gender-biased automated hiring (Dastin, 2018), the victimisation of the poor and economically marginalised (Eubanks, 2019; Karp and Knaus, 2018); and racial bias in the United States (Sweeney, 2013; Angwin, et al 2016; Ingold & Soper, 2016; Noble, 2017) The research that activated the inquiry into racial bias and algorithmic discrimination in the United States is the now-historic study by the Harvard computer scientist, Latanya Sweeney, who, Google-ing herself, found she was being served advertising for bail services related to law enforcement and the prison system. This research is about discrimination in online advertising to 'black-sounding names' like her own (2013) Another important study, Gender Shades, found that Microsoft, IBM, and Face++ had facial recognition software that could not read darker and more female phenotypes, or faces (Buolamwini and Gebru, 2018). Similarly, algorithms in computer vision systems do not 'see' people of colour or a gender, nor acknowledge more than two genders (Costanza- Schock, 2018; Keyes, 2018). This has amounted to active erasure and discrimination, violence, or unwanted hyper-visibility when people are flagged as 'suspicious' at border crossings (Amoore, 2018).

Some early work about the applications of algorithmic technologies in society mapped the normative and epistemic concerns referred to as the (lack of) 'ethics of algorithms' (Mittelstadt et al., 2016; Tsamodos, 2020) A relatively new interdisciplinary academic domain called Fairness, Accountability, and Transparency in Machine Learning (FATML)[32] assembles research about how computational decision-making, the law, and, human and social governance interact in the perpetuation of algorithmic harms such as bias and unfairness, and how these might be mitigated, to create equitable, fair, and accountable algorithmic systems. However, governing algorithms[33] is a fraught exercise because 'algorithm' does not necessarily hold up as a clear analytic category or a singular object of productive inquiry (Barocas et al, 2013).An algorithm is not just the set of rules or a recipe for how a computational process should be executed. Social and cultural scientists tend to make algorithms 'disappear' into 'material history', 'socio-technical systems', or 'culture' (Seaver, 2017) so they become unstable and difficult to study, let alone govern (Ziewitz, 2016, p. 8) On the contrary, Christian Sandvig and

[32] This domain is known in terms of its subjects of study: 'algorithmic bias', 'machine learning bias', 'fair ML' and so on. This field is associated with academic conferences organised under the aegis of the Association of Computing Machinery (ACM), an academic and industry body. Through the course of this research I have been associated with this community by attending conferences and authoring cross-disciplinary tutorials with other scholars.

[33] Attempts to manage algorithmic bias also comes from industry, lawyers, and computer scientists: Accenture's Responsible AI program: https://www.accenture.com/us-en/services/applied-intelligence/ai-ethics-governance ; IEEE's soon-to-be published industrial standard, P7003, on Algorithmic Bias Considerations: https://standards.ieee.org/project/7003.html ; New York City's task force to review automated decision-making adopted by the city: https://www1.nyc.gov/site/adstaskforce/index.page and, a new 'ecosystem' of firms providing AI ethics and bias mitigation consultancies (Hao, 2020). Perhaps nothing else has been as successful as the comprehensive calls for bans on the use of facial recognition technologies in Europe, UK, and the United States by lawyers, scientists, policy experts, and activists (Roussi, 2020)

his colleagues argue for the study of algorithms writ small, finding that ‘the algorithm’ itself is a useful locus of scrutiny for its ethics and legality; because, in doing so, there is in fact an expansion of the methodologies and modes of inquiry that takes place, into social spaces that algorithms inhabit or touch that go beyond the realm of just computer science or the law (2015). Further, that in countries like the United States, the ‘anti-discrimination’ approach is itself problematic because it centres the law, which in itself has been implicated in perpetuating social inequalities (Hoffman, 2019). More critically perhaps, accountability itself is produced, is relational, how its actors are identified, and algorithms become “accountabilia” (Ziewitz, 2011; Woolgar and Neyland cited in Barocas et al 2013).‘Bias’ and ‘unfairness’ are not just matters of machine learning, law, or data-collection, but are also historic and social. This includes, among others, historically biased training data sets, standards governing technology production, the deeply embedded social inequalities that infuse the development of technologies resulting in negative outcomes for already-marginalised communities, and all of these exacerbated by the “technology is neutral” discourse that is popular in Big Tech, the idea that these are ‘unintended consequences’ that cannot be controlled through design (Dobbe and Ames, 2019; Hanna et al, 2020).

The question of *how* we might regulate algorithms to make them less biased and more equitable is central to the FATML field. There are those who argue that algorithms can be less biased and made more fair or non-discriminatory by computational means. They propose methods oriented towards building fairness and non-discrimination into algorithmic systems themselves, broadly speaking. But, a number of social scientists and humanists associated with these fields caution the belief that algorithmic bias can be computed away is a highly seductive one, occluding how algorithmic systems are only exacerbating something biased *already* coded into our technologies and societies (Powles and Nissenbaum, 2018). They argue that algorithms are inherently socio-technical systems anyway, and that fairness and anti-discrimination cannot be understood as computational. (Selbst et al, 2019) In other words, these scholars are arguing that even computationally made value-based, ie ethical or moral, decisions, must evolve within and as part of a wider set of social relations that resting on the historic, contextual, industrial, cultural, socio-economic, and embodied aspects of those relations. Simply put, some emphasise that fairness might be enhanced and bias mitigated through re-architecting algorithmic systems from the top down; others believe that the architecture of algorithmic systems, the data that fed it, the business models that profit from it, the calculation of the optimisation function, among other computational practices, are all inherently social practices; and bias accumulates along the way as the 'fingerprints' of these practices.

Said another way, this is where distinction between *the ethics of algorithms* ('the ethics of x'), and *an ethical algorithm* ('ethical x') becomes blurry and requires clarification. Are we referring to the conditions around the algorithm that keep it ethical or not; or the intention to architect the algorithm to make ethical decisions; or some configuration of the two? The latter, *ethical x*, emphasises computational architectures and systems, and the formalisation of values into rules and representations legible to computation. The former, *the ethics of x,* identifies the myriad social, organisational, cultural, and material contexts and practices of design in which values are negotiated and discussed-either explicitly or not—and that become the very work of the building of technology. Obviously, the *ethical x* emerges from such a process, but the ethics are not located there, or rather, may not achieve as much emphasis when talking about ethics. Notably the *ethics of x* are also distinctly human, social, and organisational processes, slower, and more distributed. I return to a mention of some FATML cases at the end of this work. This case raises questions about how our thinking about values and design play out when it comes to algorithmic and AI systems given the scale and scope of their speed, and the myriad nonhuman ways that they perceive and know the world. Is it conceivable that the shaping of autonomous driving might consider ethics and values outside of the decisional moment of a speculative crash in the *future*? And, perhaps attend to a *current*, *ongoing* harms? Unfortunately, not yet. With this, I bring this chapter to a close and turn to some practical aspects of how this research was conducted.

CHAPTER 3. RESEARCH METHODS AND APPROACHES.

Cartographies and prehistories

In her landmark study of electricity and the telephone in early modern European and North America, Carolyn Marvin found that the study of electronic media as *artefacts* lies like a "great whale" across media history, rather than the *"technical prehistory"* that came before it (1988, p. 4; emphasis mine). This technical prehistory, she says, is critical because it identifies how power to speak authoritatively about technologies is awarded, and how this in turn shapes the technology itself. A study of the "changes in the speed, capacity, and performance" of communication technologies do not really tell us about the "social meanings [that] elaborate themselves undisturbed" (Ibid). Marvin's work shifts the focus from "the instrument to the drama" in which various social groups, from 'housewives' to engineers, negotiate authority, representation, and knowledge (p. 6). I refer to the emerging autonomous vehicle as the ethical apparatus; as such, it is the site of discursive power through knowledge-making practices, the work of expert communities, culture, and data and computational infrastructures, among others; I argue that these constitute both the embodied and discursive shape of autonomous driving. In that sense, this work diverges from Marvin, in that I am interested in both instrument *and* drama. Influenced by Foucauldian and Barbadian approaches to discourse and knowledge, and approaches such as the socio-technical, this work follows practices of knowledge-production and the actors that enable them. Thus I consider this work a "cartography" in the sense of "a theoretically-based and politically-informed account of the present that aims at tracking the production of knowledge and subjectivity… to expose power". (Braidotti, 2019, p. 33) I also position this work as a 'pre-history' of a technology like Marvin does, i.e. one that is expected to evolve into something else. What we refer to as a semi-autonomous, connected, networked, or 'circular' car is expected to eventually *become* a robot, a robo-car, *fully* autonomous, or driverless car. Thus, there are intersecting temporal and social-spatial registers to this work, mapping the spread and organisation of various practices and discourses of knowledge-making emerging from expert communities across histories of automobility, AI, and automation towards a future in which 'full autonomy' is expected to exist. We do not know if this will come to pass, but with an investment of US $80 billion already (Kerry and Karsten, 2017) and nothing on the horizon that lives up to promises made so far, it is to be seen when and how investors give up hope.

How do you study an apparatus?

How do you study an apparatus, practically speaking? What are the tools and capacities at the disposal of the critical cultural scientist to study AI and autonomous systems in driverless cars? I take a cue from Donna Haraway's germinal writings on epistemology and science; as a critical feminist technology researcher I have fashioned this work to be attentive to the power that accumulates to particular kinds of actors, both human and nonhuman[34], to speak about AI/autonomous systems and how they do it. Discerning how the apparatus gives form and legitimacy to knowledge is about identifying places where people speak authoritatively; where knowledge is made and how it is amplified. Identifying these sites of speaking and particular cases were varied, and emerged through the process of research, chiefly through building networks and communities, as I will discuss below.

Research about autonomous driving tends to focus on legal and policy research to support the emergence of this technology in society, studies of responsible technology innovation, communication of emerging science and technology in society, and human factors in safety engineering and design, among others. These I argue, belong to that media history tradition Marvin critiques of emphasising artefacts over their dramas; they assume that driverless cars will become part of the social landscape and want to support this emergence. On the other hand, my research is interested in the claims packaged into the shaping of AI and autonomous technologies that constitute the driverless car as an 'instrument' in society. These claims are made on epistemic and discursive grounds: that ethical decision-making might be verified by particular kinds of decision-making in complex Trolley Problem-type situations, or by scales identifying handovers between human and machine; that autonomous driving is safer because human driving is error-prone; that the human driver can be replaced and have more time to themselves while the car drives itself; that autonomous driving is also intelligent driving emerging from a cognitive-computational process of decision-making. I assess how these

[34] Here, 'nonhuman' refers to computational settings, practices, devices, and knowledge-making, such as optimisation functions, statistical models, algorithms, and computer vision, among others discussed in this work.

claims are given material form and valence through the analytical lens of the ethical apparatus, which is both a *device* and a site of *discursive power.* Thus I am interested in the co-productive and entangled interactions of device and discourse; how a practice of quantification or measurement serves as the basis for language and representations of notions of 'autonomy' or 'intelligence', for example. In other words, how the conditions for the very construction of knowledge about 'autonomy' or 'intelligence' become the terms under which we can, and know how to, talk about it. These representations go on to take on a life of their own, effectively leading that artefact to have power and agency in shaping social, cultural, and human worlds.

Since epistemology and discursivity are the dramas in this work, I shall map the conditions of shaping and emergence of knowledge about what constitutes autonomy and ethics. Hence, I have selected sites that I believe are relevant therein. Given that this work is multi-disciplinary, it draws on influences adjacent to the field of Cultural Studies: Media Studies, Big Data Studies, Critical AI studies including Histories of AI, Science and Technology Studies, Digital and Cultural Anthropology, in particular Infrastructure Studies, Software Studies, Feminist Epistemology.

Empirical methods and contexts

This research is a qualitative, cultural study of the knowledge-making practices that constitute and shape the 'autonomy' of the autonomous vehicle along with notions of ethics and intelligence. Conducted over five years (2016-2020), it is based on a combination of field work and desk research that include the following[35]:

- The construction of ethics through the Trolley Problem and its iterations, primarily the Moral Machine from MIT's Media Lab's (now dissolved) Scalable Co-operation Group.
- Cases of recent fatal autonomous vehicle crashes in the United States.
- 20 in-depth, unstructured interviews with academic, policy, and industry experts from Germany, India, and the United States working in the Law, Computer Science, Design, Mapping, Robotics, and Automotive Engineering. (2016-2020)
- Two interviews with Tesla owners in North America, and a test drive with one of them. (March 2018)

Interviews

I have sought out individuals whom I considered experts, and who I believed were shaping the emergence of the driverless car. Interviews with experts would allow me to make sense of industrial public relations and media material that also constituted 'discourse' and data. Interviews helped me understand how the technology was evolving and the different influences on this evolutionary process. I used a combination of three methods to access interviews: Personal networks, 'cold calls' (or, emails rather) to experts, and by actively participating in this field as a researcher by giving talks at events.[36] This last approach put me in contact with

[35] A comprehensive list of events, sites, and interviews, and their dates and locations, is included in Appendix 2.

[36] Chiefly, 'digital society' cultural events, as well as academic conferences.

communities of people who were working on questions of AI, ethics, and autonomous driving. Many direct requests for interviews with roboticists, scientists, software engineers, and HCI experts went ignored. The people who were most amenable to interviews and discussions were policy makers, innovation policy academics, lawyers, and automotive engineers working in car companies. In other words, people who were trying to make sense of these software and digital technologies too. In the case of my research, I found that, from policy makers to lawyers to engineers, no one is quite sure what to do about the 'problem' of AV 'ethics', or even 'autonomy'. Hence, there is interest, particularly in Germany, in engaging with potential new experts who might have the answers. Personal networks emerged as sources of interviews in two ways: through chance and the generosity of friends; and by reaching out to people who I followed on social media platforms. For example, a vocal and critical 'mobility lawyer' from California who I followed on Twitter was willing to meet me for an interview when I was visiting the Bay Area after attending an academic summer school.

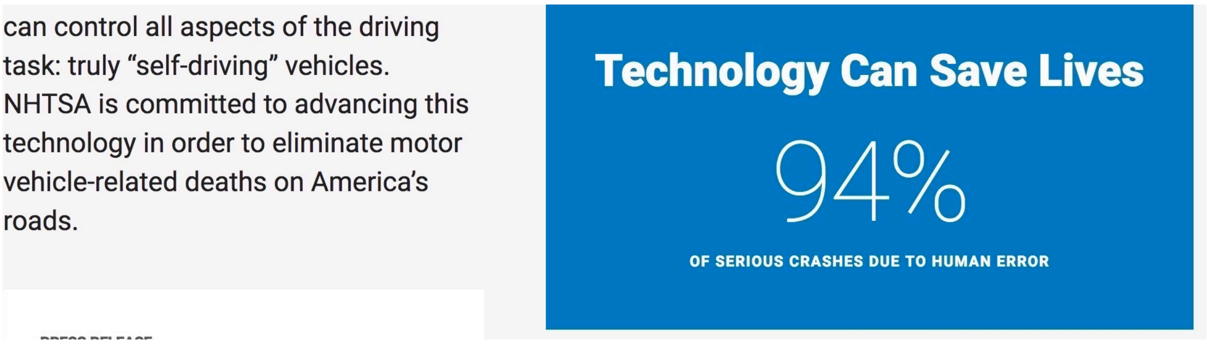
Locations

This research refers primarily to the ecosystem of autonomous automobility development in North America, Australia, the UK and Germany. These regions have been selected either out of proximity, or opportunistically, thanks to professional and academic networks and affiliations in these places as discussed above. However there is a significant emphasis on histories and materialities emerging from North America where driverless car testing is taking place publicly. There is a slightly different trajectory in Germany that is associated with its strong automotive engineering and manufacturing industries. Unfortunately, language barriers made it difficult for me to access written material in German, and do interviews in German. However, many engineers and researchers here are fluent in English as well, which has been extremely helpful.

Processes

I kept notes of my encounters and discussions with experts that I spoke with; they were aware that this was part of a dissertation research project and I confirmed their consent to my writing about our discussion. There was no structured interview guide that I used, therefore it is not possible to consider them all equivalent. Some were unexpected and fleeting conversations, email exchanges; the test drive with a friend was completely unexpected and unplanned. The process I followed in all interviews was similar. Everybody I interviewed was an expert or authority of some kind, either self-appointed or recognised as such. I perceive them all to be in positions of power and unlikely to be disadvantaged by a researcher asking them questions. In some cases, people welcomed talking with me because they believed it was interesting and relevant. Interviewees were asked if they would like to be anonymised. Some did, however most did not. Nonetheless I have partially anonymised some of the people I interviewed by changing their names. The next chapter, Ironies of Autonomy, introduces recent AV crashes to stimulate an empirical and theoretical discussion about the evolving relationship between human and machine in the shaping of autonomous driving. It responds to the claim that autonomous driving is safer and more efficient because the human driver is effectively replaced or erased. I find that the human is not replaced or erased but is in fact taking on new roles, bringing into question a much-needed re-think of ‘autonomy’ itself.

CHAPTER 4. THE IRONIES OF AUTONOMY



can control all aspects of the driving task: truly “self-driving” vehicles. NHTSA is committed to advancing this technology in order to eliminate motor vehicle-related deaths on America’s roads.

Technology Can Save Lives

94%

OF SERIOUS CRASHES DUE TO HUMAN ERROR

FIGURE 5. SCREENSHOT FROM THE NATIONAL HIGHWAY TRANSPORT SAFETY AUTHORITY (NHTSA.ORG) OF THE UNITED STATES. DATE OF SCREENSHOT: 23 OCTOBER 2018.

Blur

The case for autonomous driving is made by showing that humans are error-prone and that software is safer. The US National Highway and Traffic Safety Association proclaims on its website (pictured above) that ‘technology can save lives. 94% of crashes due to human error’. Autonomous vehicles (AVs) are promised to be safer: they follow rules, do not speed nor get drunk and drive, do not check their phones, get sleepy, or distracted. However, four recent crashes involving AVs resulted in human fatalities.

- In Hebei province, China, a driver was killed when his Tesla Model S crashed into a road-sweeping vehicle. The father of the driver claimed that the car was in Auto-pilot[37] mode, however Tesla said that the damage to their vehicle was so severe that it was not possible to retrieve information about how the crash actually occurred (Boudette, 2016)
- In Florida, a test driver, Joshua Brown, was killed when he did not take over control from the Tesla that was in Auto-pilot mode and that drove into the light, white side of an 18 wheel truck, mistaking it for the early morning sky. The driver did not respond to alerts to take over, and eventually had only three seconds to respond before impact (National Transport Safety Board, 2017; Tesla, 2016)
- In Arizona, an Uber test-driver in a Volvo semi-autonomous vehicle did not take control of the car that was in Auto-pilot and hit a pedestrian, Elaine Hertzberg, wheeling her bicycle across the road; the pedestrian was not properly identified by the vehicle's computer vision software; the driver was found to be distracted (National Transport Safety Board 2018);
- In California, a driver was killed when driving a Tesla in Auto-pilot mode that drove into a road works barrier (Shepardson, 2018).

There is a thread in the causes for each of these crashes: faulty handover between the human and the vehicle that was in Auto-pilot mode because of the human driver's distraction, or slow response time in taking over (National Transport Safety Board, 2017, 2018, 2019). With an increase in automation of a task, the human physical and cognitive skills required to com-

[37] 'Auto-pilot' with the A capitalised refers to Tesla's auto-pilot technology; however 'auto-pilot' refers to the technology itself.

plete, monitor, oversee, and eventually step back in to that task become poor, and especially at short notice (Cummings et al., 2013; Cummings and Ryan, 2014). So, it becomes harder for the driver to take control of the system. Ironically, this is the promise of driverless-ness in the first place: handing over to the more efficient and accurate AV, with the freedom to be distracted and inattentive. But, it was not just the handover between human and machine that was a problem, but a failure of the computer vision systems within the Florida and Arizona test AVs[38]. A key insight by John Cheney-Lippold frames these crashes in terms of the AV as a big data infrastructural system, one that we are still struggling to understand. He writes with reference to the Florida crash:

> an ontological gap form[ed] between a white truck crossing into one's lane and an algorithmic interpretation of a white truck crossing into one's lane. One is a collection of elements moving through time and space; the other is a probabilistic evaluation of those elements, represented, as best as the algorithm can, as a deviating new world, intelligible as data, where a white truck ceases to be a white truck and becomes a statistical relationship. It is instead a formal acceptance that the statistics underlying Tesla's Autopilot suite are operational precisely because they are not evaluating some mythical, unmediated "real" but rather are processing the world in line with the necessarily objectifying force of statistics. (Cheney-Lippold 2019, p. 527)

[38] I refer to these two cases in this chapter because there is a substantial amount of publicly available data, such as official forensic reports, published about them.

In each of these accidents, the human who was not paying attention was put in the situation of having to take over because of an error that occurred elsewhere in the big data infrastructure of autonomous driving, that is, in its computer vision system. It is near-impossible to intervene in such a complex industry in such an instance. The recognition of the world through the (not so) “objectifying force of statistics” (Ibid) that mistakes a truck for the light blue, almost-white sky suggests a scale of computational, automated decision-making that diverges from conventional approaches to human-machine interaction. As I will discuss in this chapter, safety engineering and design in AVs can be traced back to the human operator-machine relationships that emerged in aviation safety and accident avoidance. And just as the airplane has become more computational, the AV is already a big data infrastructure. It is not a conventional car, and the socio-technical systems that comprise it produce new human subject positions. Together, these muddy notions of ‘autonomy’ as self-reliance, or independence from humans. For example, the conditions of optimisation and standardisation of the data in the statistical relationships that underlie computer vision have the power to produce multiple, conflicting subjectivities within the AV: that of an accident victim on a dark night or poorly lit street; in the operator’s hot seat and expected to take over at a moment’s notice, but without any control over the contingencies set in motion by the computational infrastructures she is embedded in; and as a ‘heteromated’ worker-cog propping up these material infrastructures, including annotating and labelling visual images for computer vision systems. It has become common for computer vision systems applied in autonomous driving to outsource image annotation as low paid digital piece-work around the world. So even as metrics and computational infrastructures of ‘autonomy’ are created, humans play a significant role in these systems, as I detail here– yet, this significance is associated with accountability and liability, and requires being subject to surveillance and monitoring.

Between the legacy approaches to automation and automobility, the conditions of subjectivity to systems like distributed computer vision, and a future state of 'full' autonomy, lies a hauntology that this chapter addresses. This haunting is a condition of the multivalent ontologies of the autonomous vehicle that I believe is an instance of what Benjamin Bratton refers to as 'blur', within the "accidental megastructure" of computational governance and globalisation that is the 'stack'.[39] (Bratton, 2015, pp. 13-17) The AV's big data infrastructure is an instance of this new planetary-scale computational geopolitical governance, the stack, that introduces bewildering, uncannily inter-twined, fractured, poly-scalar complexities of place, time, and subjectivity that we do not yet know how to name, and that demand their own epistemologies and accountability regimes. Blur, he says, is

> a system in advance of its appearance maps what we can see but cannot articulate, on the one hand, versus what we know to articulate but cannot yet see, on the other. This oscillation between the real-but-as-yet-unnamed and the imagined-but-as-yet-not-real —this blur between them—might sustain the necessary challenges to the imagination

[39] It is difficult to launch a full discussion of Bratton's notion of the stack in a footnote; in the main text, it runs the risk of taking over the narrative. *The Stack: On Software and Sovereignty* is a treatise and speculation on design, political geography, software, theory, architecture, and computation, among others. Bratton describes the stack in terms of computational governance proposing quite simply that it emerges from the notion of the machine as the state. This is where he makes a connection to Foucault: "Just as for Foucault's technologies, its mechanics are not representative of governance; they *are* governance. But unlike for Foucault's archaeology, its primary means and interests are not human discourse and human bodies but, rather, the calculation of all the world's information and of the world itself as information. We, the humans, while included in this mix, are not necessarily its essential agents, and our well-being is not its primary goal." (p. 8; emphasis in original) The stack is comprised of six layers of planetary state computation: earth, cloud, city, address, interface, user.

> and even enforce what it conceives, a giving way to compound images and sectional perspectives (Op Cit, p.13)

The autonomous vehicle project and its ontological multivalence raises the kinds of stack-ed and blur-red conditions that result in a crash: a poorly captured or wrongly annotated image, or a test driver caught up in large-scale, automated decision-making and its cognitive flow, an accountability regime that understands the human as ultimately responsible for anything that goes wrong. But, it is difficult to hold a single individual responsible for a crash (although we do) in terms of conventional accountability mechanisms. How the annotation of an image even makes a difference to all this, or not, is difficult to grasp in familiar terms of geography, industry, software, tech policy, or driving laws. It is also not just that everything is 'just data' now, an entirely cybernetic dream come true where everything is responding to everything else purely in terms of data and statistical operations. These compound and sectional perspectives are what we might perceive and experience as the *ironies of autonomy*, a riff on Lisanne Bainbridge's original *ironies of automation*: "that the automatic control system has been put in because it can do the job better than the operator, but yet the operator is being asked to monitor that it is working effectively." (1983, p. 776). Similarly, in this chapter I discuss how the irony of autonomy emerges when computationally superior and efficient machines actually need human micro-workers, many of them rendered invisible, to ensure that it is working effectively in a display of the mythical status of autonomy. Moreover, the human is inevitably held accountable for errors in this system.

I am not supplanting the ethical apparatus with the stack; the ethical apparatus is a study of discursive formations and the transformation of bodies and spaces. In that sense, the broader Bratton-ian project has resonance for this work. I do not refer to the stack in much more detail, although Bratton's *blur* and the *accident* (that I refer to at the end of this chapter) work as powerful theoretical perspectives on the conditions of materiality and media emerging from AV crashes. The crash is a key moment in understanding of the shaping of autonomy: it is what the Trolley Problem apprehends as central to the construction of autonomy, for one; other definitions of autonomy are discussed in this chapter next. The crash is also a key moment in computation, as media theorist Alexander Galloway observes, when the computer becomes 'non media'; that is, the crash reveals the re-mediation of text and image into computation. (Galloway, 2012, p. 21) Similarly, I believe the study of the AV crash will reveal the constitutive practices of mediation and materiality in how the shaping of 'autonomy' transforms spaces, bodies, and social relations therein. I return to the AV as a media object at the end of this chapter. For now, I turn to remembering that the driverless car is also ontologically, a car.

The driverless car as just a car

The driverless car is also just a car, and as such, a very important media technology of the past century. The automobile as a social and media technology works in two opposing ways; it individualises and fragments human relations and spaces, but it is also strongly associated with cultures of embodiment and human-machine hybridity. As such, the AV as a car also implies that it is a media object and the site of multiple dynamics of mediation. The first dynamic of individualisation and fragmentation was identified as ‘mobile privatisation’, with numerous plays on the meanings of the words, ‘mobile’ and ‘privatisation’. Chiefly this refers to the increase in mobility that enabled the (mid 20th century North Atlantic and European) home to become more private and self-sufficient (Williams, 1974/2003). And of course in tandem with other technologies evolving at the time, from the television set, to packaged margarine[40] that spelled modernity. The notion of mobility was not just literal in terms of car culture and driving, but also includes television and media that offer the outside world as spaces that could be ‘traveled to’ from within the home. Mobile privatisation also made restrictions on mobility stark, such as of older people, or those who had care responsibilities that kept them at home, and thus cut them off; the individual who is private as a result is cocooned (McGuigan, 2013, pp 78-79). Williams was presaging an individualisation and fragmentation that was relatively novel then, and one that we have come to accept as a fact of life now, mediated as our spaces are by mobile phones that create digital bubbles. “Are we there yet?!” is the annoyed refrain of the traveler who must put up with the tedium of travel and close interactions with other humans (family, and other commuters) as the price paid for the value associated with mobility. Media have rushed in to fill this gap and massage the spaces of friction with other humans; we can shut ourselves off in perceptual and spatial bubbles of headphones and personal screens; now we will have masks too. Screens within cars, like video and TV, to our phones, capture our attention — aka data — which translates into algorithmic knowledge about consumer tastes and preferences (Packer and Oswald 2010, p. 314). This is foreseen as one of the profitable projections of autonomous driving. The opportunities for media and entertainment consumption are thought to explode with the emergence

[40] Owing to the post World War 2 butter scarcity, and despite strong resistance from dairy farmers.

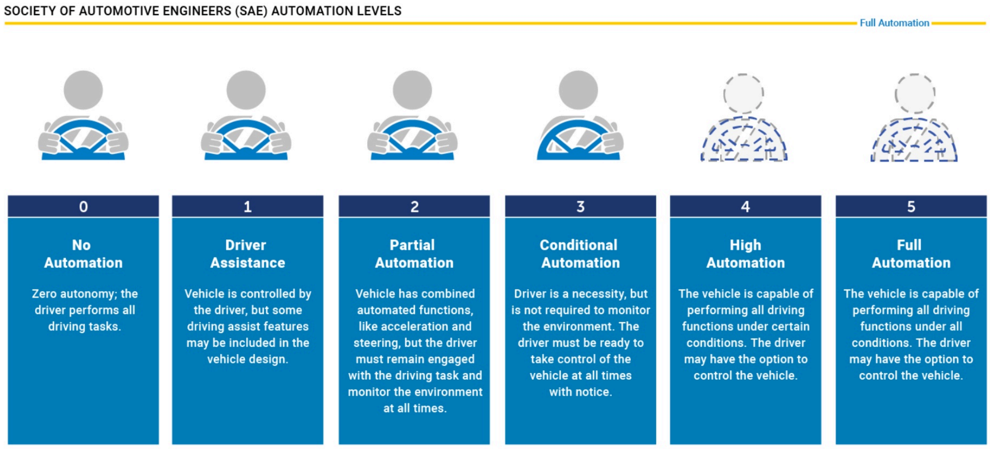
of the fully autonomous vehicle. The consulting firm McKinsey estimates that a “mind blowing one billion hours”, roughly twice the time it took to build the pyramid at Giza, has the potential to generate “global digital-media revenues of €5 billion per year for every additional minute people spend on the mobile Internet while in a car” (Bertoncello and Wee 2015; np) So the car is set for another reinvention as a new social artefact, a more profitable one. In a casual conversation about this work, the social scientist of financialisation, Martha Poon, made a throwaway remark that is uncannily astute in bringing together many elements of this ontologically multivalent technology. She said: “the driverless car is imagined as the perfect little neoliberal subject that willtootle along making decisions for itself.”[41] I believe Poon is referring to the imagination of the AV as *Homo economicus*, the human who makes rational and informed decisions, unfettered from struggle, and disconnected from history, or affect; the human who is always trying to improve themselves and become more smart and successful by making the right decisions. The transformative future of the driverless car captures all these subtle layers of human and machine.

[41] In personal communication, Brussels, Belgium, January 27, 2017.

What makes a car autonomous?

Human-machine handovers

What ‘fully autonomous’ means is a difficult question to answer directly, and this research could be considered an exercise in problematising how and why we ask it in the first place. A vehicle is considered autonomous if it navigates the road like a human driver does, that is, on its own and without a human needing to pay attention to it. There is a standard that captures this relationship in a heuristic that has taken on the status of fact. The initial J3016 standard for automated driving issued by the Society for Automotive Engineering shows Level 5 as “Full automation: the full-time performance by an automated driving system of all aspects of the dynamic driving task under all roadway and environmental conditions that can be managed by a human driver.” (SAE, 2014). This is accompanied by a graphic that has been widely reproduced in popular media, tech writing, legislative and policy documents. In this graphic, autonomous driving is presented as a linear scale from ‘no automation’ to ‘full automation’; at the end of the scale depicting ‘no automation’ (level 0), a humanoid figure, shaded in solid grey colour is at a steering wheel shaded in a solid blue; as the levels of automation increase upwards to 1, 2, 3, the human and the wheel, both lose their solid shading, become clearer, and are bounded by a dotted line, suggesting a change in their status, like instability or disappearance. Recent updates to the standards do not include the image of the humanoid figure (SAE, 2018). Instead, the text describes increasing automation in behavioural and technical terms only. SAE levels can be misleading and dangerous but also ignore the many layers of automation that already exist in driving at present such as parallel parking, rear mirrors, lane-assist and other features (Stayton and Stilgoe, 2020; Roy, 2018). Liza Dixon proposes the term 'autonowashing' "to describe the gap between the way automation capabilities are described to users and the system's actual technical capabilities" (2019). Similarly, the prefixes 'fully' and 'semi' suggest linear stages leading up to 'full' autonomy as a final destination. However, this heuristic is now a widely-cited approach to autonomy.



SOCIETY OF AUTOMOTIVE ENGINEERS (SAE) AUTOMATION LEVELS

Full Automation

0 — **No Automation**: Zero autonomy; the driver performs all driving tasks.

1 — **Driver Assistance**: Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design.

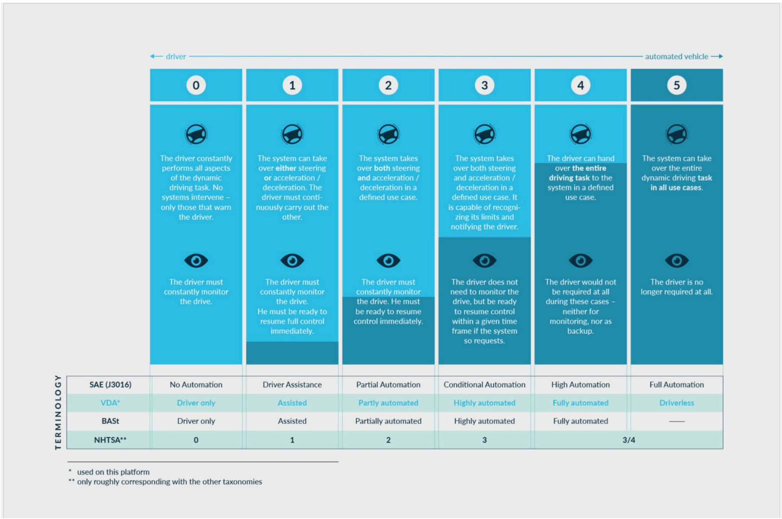
2 — **Partial Automation**: Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving task and monitor the environment at all times.

3 — **Conditional Automation**: Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice.

4 — **High Automation**: The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle.

5 — **Full Automation**: The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle.

FIGURE 6: SAE J3016 STANDARD FROM 2014. BELOW: SAE J3016 STANDARD UPDATED IN 2018. SOURCE: SOCIETY OF AUTOMOTIVE ENGINEERS. HTTPS://WWW.SAE.ORG/NEWS/2019/01/SAE-UPDATES-J3016-AUTOMATED-DRIVING-GRAPHIC



← driver — automated vehicle →

0: The driver constantly performs all aspects of the dynamic driving task. No systems intervene – only those that warn the driver. The driver must constantly monitor the drive.

1: The system can take over **either** steering **or** acceleration / deceleration. The driver must continuously carry out the other. The driver must constantly monitor the drive. He must be ready to resume full control immediately.

2: The system takes over **both** steering **and** acceleration / deceleration in a defined use case. The driver must constantly monitor the drive. He must be ready to resume control immediately.

3: The system takes over both steering and acceleration / deceleration in a defined use case. It is capable of recognizing its limits and notifying the driver. The driver does not need to monitor the drive, but be ready to resume control within a given time frame if the system so requests.

4: The driver can hand over **the entire driving task** to the system in a defined use case. The driver would not be required at all during these cases – neither for monitoring, nor as backup.

5: The system can take over the entire dynamic driving **task in all use cases.** The driver is no longer required at all.

TERMINOLOGY

	0	1	2	3	4	5
SAE (J3016)	No Automation	Driver Assistance	Partial Automation	Conditional Automation	High Automation	Full Automation
VDA*	Driver only	Assisted	Partly automated	Highly automated	Fully automated	Driverless
BASt	Driver only	Assisted	Partially automated	Highly automated	Fully automated	——
NHTSA**	0	1	2	3	3/4	

* used on this platform
** only roughly corresponding with the other taxonomies

In a research interview, a Human Factors in Engineering researcher at a US university describes the dissimilarities between human driving and software-led driving.[42] He says that AVs can be slow, ("like a little grandma driving!"), stopping if it doesn't know what to do when faced with new situations that are not covered by the rules in the learning system. Humans, on the other hand, he says, extrapolate from past experience to figure out how to address new challenges. "Autonomous vehicles tend to stop entirely in new or unfamiliar situations, which is not always very helpful" he goes on. A human driver is able to patch together sensing, perception, memory and the body to generate the appropriate response. The same researcher also tells of a research experiment he conducted. In this, he found that human drivers caught in snowstorms that completely obscured lane markings were still able to approximately maintain the required distance in their own lanes; however, AVs were at a loss because they relied on the visual information conveyed by the lane markings, and did not have any data in their databases that might allow them to compute their way out of this problem. It is precisely this inability of the AV to adapt and respond on the fly that the human has to step in to help with. This handover has become a measure called ‚disengagements', as applied in California and defined as the number of miles driven in 'autonomous' mode (with a human required by law to be present) before the human driver has to take over (Hawkins, 2020). Every year, car companies authorised to test their driverless car technology in California must submit disengagement reports: "Manufacturers must track how often their vehicles disengage from autonomous mode, whether that disengagement is the result of technology failure or situations requiring the test driver to take manual control of the vehicle to operate safely." (California Department of Motor Vehicles) However, disengagement reports are con-

[42] Interview with Brandon Schoettle, October 2017. See appendix 2 for details.

tested because they can be misleading; an AV can record a relatively low disengagement metric by testing on open, empty highways rather than in more challenging driving environments like a crowded city. Thus, autonomy is constructed in terms of a car and its computational system taking over from the human driver in an almost direct handover, or transition, from human to machine.

Environment and infrastructure

'Autonomy' also takes shape as a response to its environment and infrastructure, as something flexible and modular. During one interview, I asked an automotive engineer in California who advises software companies *when* the driverless car would be ready. He reframed and corrected my question: "you're asking the wrong question", he said, "it is not *when* it will arrive, but *where*."[43][44] What he went on to say was that the driverless car would be determined not by its form, or capabilities, but about where, meaning in what *contexts*, it would debut. His answer was that the elements of what made the car driverless were already here and were expected to be ready by 2020 (Hudda et al 2013). He told me that the driverless car could materialise anywhere with a fixed path and a stable environment where nothing out of the ordinary or irregular would happen; such as an airport, a very tightly controlled and regulated space. This could be, for example, a bus transporting passengers or airline crew to the aircraft, contexts that were likely to be fixed and stable. Thus the driverless car is something

[43] Interview with Sven Beiker, July 2018. See appendix 2 for details.

[44] A point also made by the technology commentator and writer Benedict Evans in a piece called Where Not When in 2018. It is possible this interviewee was referring to Evans' piece without expressly citing it: https://www.ben-evans.com/benedictevans/2018/3/26/steps-to-autonomy

modular, like Lego; it is an object understood in terms of its constituent technologies that can be assembled to materialise anywhere; reliant on enabling conditions. This is the the notion of AV as a big data infrastructure, as well as the robot, that needs the environment around it as fixed and unchanging if it is going to function effectively.

These enabling conditions for autonomous automobility are being established in terms of measures of policy, and technical infrastructure, called the Autonomous Vehicle Readiness Index (AVRI), that helps policymakers prepare for the eventual emergence of AVs. “And whether you believe that will take 10 years or 30, the implications are so far-reaching that policymakers need to start planning now for our AV future” proposes the management consulting firm, KPMG (2018, p. 6). The AVRI evaluates countries on their preparedness to adopt AV technology along four ‘pillars’, each ‘pillar’ comprising variables such as the amount of government investment, the legal and regulatory environment to enable testing, research and development, to the density of electric vehicle charging stations, testing environments, and 4G coverage, among many others. These variables are all represented by a score. Some variables are made up by KPMG , like ‘change readiness’, based on an entirely different set of KPMG indices. Other pillars like ‘consumer acceptance’ are based on published data from surveys, for example. And while “number of electric vehicle charging stations”, “AV-related patents” or “total investment in AV self-driving cars” are actual figures that can be counted (p. 52), others are more difficult to render as numbers. Like, “change readiness” or “efficiency of the legal system in challenging regulations” (for which the World Economic Forum has an index). Netherlands (with a score of 27.73), Singapore, the United States, and Sweden top the AVRI in 2018 and 2019, and at the bottom of the table are Brazil, Russia, Mexico, and India (with a score of 6.14).[45]

From this perspective of autonomy as environmental and infrastructural, it might be possible to argue that an AV crash happens where the *conditions are not appropriate*, because the city's infrastructure in that section of the road did not support robust connectivity to the cloud, or because the policy did not enforce safety standards for testing, or verified computer vision vendors. Thus when we ask what an AV is, we have to verify what exactly we are referring to: the robot car in which the computer takes over, or the big data infrastructure, or the environment and infrastructure that sustains and is part of this new kind of automobility. Or, perhaps, the autonomous vehicle is something more fragile and ephemeral at present, a system of systems in which the human achieves new kinds of embodied subjectivity, which I turn to next.

[45] This index comprises measures on different scales based on data collected from a variety of sources, all of which have their own methodological constraints and logics. Each pillar and the variables within are 'normalised' by the min-max method, a fairly common statistical process by which different scales are 'made identical', in a sense, so that all features of a dataset are equally important. This allows the pillar with the largest number of variable to "not dominate when arriving at the aggregate score.. Meaning that each pillar has equal weight in the overall score for each country." (p. 51).

Heteromation

A popular perception of autonomous driving is that it replaces the human driver with computation and automation, and thus is sometimes colloquially referred to as ‘robot driving’. However, automation, historically, does not *replace* the human but in fact *displaces* her to take on different tasks (Sheridan and Parasuraman, 2005). Humans are distributed across the internet as paid and unpaid micro-workers routinely supporting computer vision systems; and as drivers who must oversee the AV in auto-pilot. Computer vision in AVs is not advanced enough for driving and has emerged as a weak link in all fatal crashes so far as the cases above indicate. It is not that the AV, fitted with multiple sensors, cameras, Lidar and radar to document the environment, cannot visually sense, but that it *cannot make sense of* what it senses. Humans must annotate images so that computer vision algorithms can learn to distinguish one object from another, and then apply this when encountering new and unfamiliar images. This technology is quite central to the ‘ethics’ as described by the Moral Machine and other proposals for autonomous vehicle ethics, because computer vision must first detect what is in the environment around it in order for the computational ethics system to make an assessment of it. The labour of humans in the AV’s data infrastructure is an instance of ‘heteromation’ (Ekbia and Nardi, 2014), the value extracted from micro-work across large and small online systems to support them, and that go unrewarded or are minimally rewarded. Distinct from automation where “the machine takes centre stage”; or augmentation where “the machine comes to the rescue”, ‘heteromation’ is defined as “the machine calls for help” (Ekbia and Nardi, 2014, n.p.), and in which the human becomes legible as a “computational component”. In making human micro-work visible here, I want to introduce another frame on autonomy in terms of how humans constitute its big data infrastructure, and yet eventually remain, curiously, outside the discourse of the 'ethics of autonomous driving'.

A slew of companies hire workers in low-income countries to do this annotating and tagging (Lee, 2018). This micro-work also happens through reCAPTCHAs: internet users tag images as depicting trees, storefronts or chimneys in order to complete various online transactions (O'Malley, 2018). It is also not unusual for off-the-shelf, already-annotated datasets to be installed wholesale. However, it was found that there are substantial errors in these datasets being used to train driverless car software; correcting and updating these databases require more human work (Dwyer, 2020). But, the world is not static and off-the-shelf databases are not always current, even if they are correct. The European AV industry contracts specialist online micro-work platforms, often branding themselves as 'AI companies' and distinguishing themselves from the 'legacy generalist' platforms like Amazon Mechanical Turk or Crowdflower (Schmidt, 2019, pp. 4–5). Interestingly, all these specialist micro-work platforms brand their work as having some kind of human machine teaming, collaboration, or human oversight when there is algorithmic annotation of images, and particularly in edge cases that are difficult for computer vision to parse (Schmidt, 2019, pp. 11–12). This process abounds in ironies. Humans perform these mundane, boring, routine tasks that are integral to the critical safety that AVs are supposed to achieve. Human workers in such firms are low-wage and low-income workers. And yet, this is what is referred to as 'artificial intelligence' and 'autonomy'. This is also likely to be done on-the-fly and outside the contexts where autonomous driving is actually being tested; in other words, there is computer vision piecework done in Kenya or India where there is no public autonomous driving testing underway.

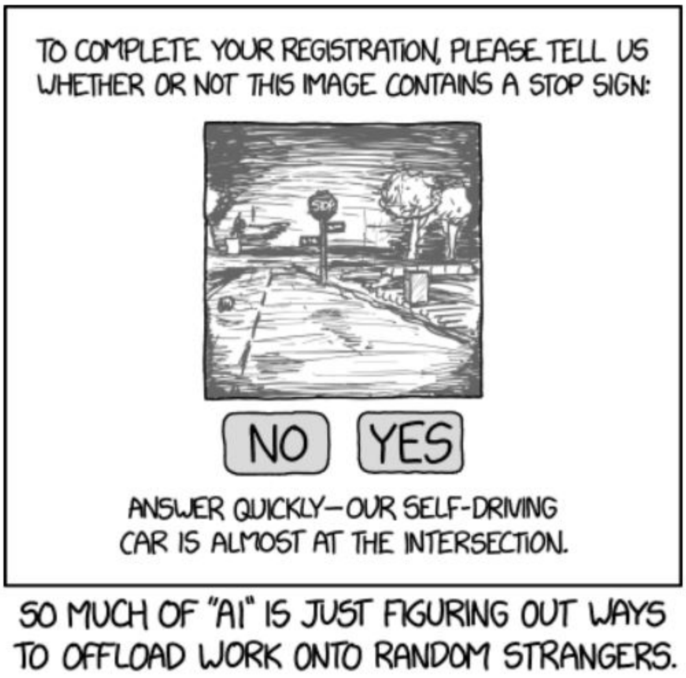

TO COMPLETE YOUR REGISTRATION, PLEASE TELL US WHETHER OR NOT THIS IMAGE CONTAINS A STOP SIGN:
STOP
NO
YES
ANSWER QUICKLY—OUR SELF-DRIVING CAR IS ALMOST AT THE INTERSECTION.
SO MUCH OF "AI" IS JUST FIGURING OUT WAYS TO OFFLOAD WORK ONTO RANDOM STRANGERS.

FIGURE 7. A COMIC FROM THE POPULAR SCIENCE AND MATHEMATICS WEBSITE, XKCD. SOURCE: HTTPS://XKCD.COM/1897/

In one interview, a transport researcher at a US university tells me about how stopgap measures for failures or gaps in data sets. He says, "with 5G, this should not be a problem...if there are doubts or errors, we can just patch in crowd workers from Pakistan or wherever to respond [make sense of the image] to whatever situation arises that the car cannot deal with."[46] He does not stop to address the assumption that there will be frictionless 5G connectivity between a car anywhere in the world and Pakistan. Another AV software maker is tapping into a 'live' source of data. Comma AI wants to build an entirely open source "superhuman driving agent" (Comma AI, 2020) but first requires less-than-superhuman agents to buy their

[46] Interview with Noah Goodall, September 2019. See appendix 2 for details.

apps and cameras, like a US$ 700 dash cam, EON.[47] This dashcam “connects to the car’s communication system to record the car’s data in sync with driving videos.” Open Pilot and Comma Two are other recent developments by this company along similar lines. This data is constantly training and updating machine learning driving models. Comma AI has already collected over 20 millions miles of driving data. Humans thus become the unrewarded yet essential seeing-eye dogs for ‘autonomous’ vehicles.

I continue to lay out different approaches to autonomy that all appear to overlap and build on each other, and the multiple ontologies of the AV as robot/AI, conventional automobile, and big data infrastructure. Next, I turn to how human embodiment remains a legacy approach that persists in the current material practice of the shaping of autonomy.

Embodiment

Cultural theorists of automobility persuasively argue the automobile as an extension of the human. A “complex hybridization of the biological body and the machinic body” (Sheller, 2004, p. 232) in which “new forms of kinship are elaborated ‘linking animate qualities to the machine’”, “not only do we feel the car but we feel *through* the car and *with* the car” (p. 228; emphasis mine). This has always been one of the most satisfying aspects of driving; but, embodiment and the ‘feel’ of the car are being eroded with the emergence of the fully autonomous vehicle as I find on a test drive (ahead). Postphenomenological approaches to technology propose that the interaction of humans with technologies constitutes a shared lifeworld that

[47] An interesting aside is that this project to build the world’s first open source driverless car is led by George Hotz, the hacker infamous for being the first person to jailbreak an iPhone.

shapes knowledge, politics, aesthetics, and social norms. Complicated new agencies emerge through the shifting subject and object positions of human and machine rendered through: practices of embodiment, hermeneutics, alterity, and 'background relations' (Ihde, 1995; Rosenberger and Verbeek, 2015). Embodiment is particularly resonant in the case of the AV; this refers to the 'taking in' of a technology device into human bodily experience, and the extension of the human back into the device, such that the technology 'disappears' and becomes notionally transparent (Rosenberger and Verbeek, 2015; Verbeek, 2011, 2017).

The register of embodiment is being deepened through approaches to encourage human connection with the evolving car, but not as drivers. Aside from online tasks, humans are encouraged to 'empathise' with the emergent machine that struggles to learn how to navigate the world. Recent business, AV Engineering, and Human Computer Interaction narratives suggest that the language of human-machine relations is changing, from 'looping' to the affective registers of 'teaming', trust, and empathy (Visser et al, 2018). There is a subtle disciplining of the human into a new approach to cars and driving. For example, robotics researchers want to match the personalities of humans to AVs to encourage humans to 'feel connected' to cars, to encourage uptake of autonomous driving perhaps. (Zhang et al, 2019) Nissan Labs (2019) are proposing 'Human Autonomy Teaming' (HAT), also developed by Human Factors research in aviation in which autonomous agents are not 'tools' but are 'team members'. (McNeese et al, 2018) 'Who Wants To Be A Driverless Car' invites people to 'empathise' with AVs in a more physical way; they have to lie down inside the frame of a motorised buggy and wear a headset that replicates the three-dimensional map view that AVs employ so as to 'understand' what they see. (Move Labs, 2019) Human operators are encouraged to be sensitive to the needs of the emergent AV. The language of teaming, trust, and empathy, speak to a kind of "body that [is itself] fragmented and disciplined to the machine" (Urry, 2004, p. 31). Post-phenomenologists typically refer to benign examples of embodiment such as reading glasses and walking sticks that work by being 'embedded' in the human body, one expanding into the other to work effectively; however in the case of the AV there is a serious edge to the 'hybridisation' of car and driver that goes beyond the body and includes psycho-affective and emotional states as well, for it can spell the difference between life and death as crashes have shown. This comes through in a macabre fashion in an interview with 'Jon', a UX designer in Silicon Valley.[48]

"In the future you're going to see rich people driving old cars", claims Jon, "like people playing vinyl records. It isn't necessarily just empty nostalgia, there is a very specific quality to the sound from a LP, which is utterly unique."

"People will use autonomous cars for ride-sharing to get around, but the rich will have cars that are manual, difficult to drive, with no power steering and so on, because the feeling of driving something like that is entirely unique, and visceral. Do you know that the vintage car market has exploded?! They're impractical but this visceral experience will be harder and harder to find, as automation becomes more and more of a commodity and the norm. It is a struggle to drive these old cars, but that's the point. There is the allure of spending time on it, fixing it, tinkering with it. It's like sensorium in high definition."

[48] Interview with Jon (name changed), June 2018. See appendix 2 for details.

Jon arrives at this prediction after telling me about his own car —"a reasonably rare Porsche" —and experiences of driving in California. Jon is a father, a senior manager in one of the world's most powerful technology companies, and one of the rich car enthusiasts he talks about. He tells me his car is from the 1970s, has manual transmission, no power steering, Cruise Control, or other automated features that a contemporary car has. The car is difficult to drive and requires "real power, real energy." He goes on to say, only half-jokingly, "You could die driving my car on the highway. …But I have an SUV to take my daughter to school of course."

Auto-pilot's aviation legacy

Jon gets it right, I find, but not quite right; embodiment and 'understanding' the car are actually important in testing or overseeing the driverless car, but not quite like in conventional driving. There is a different level of communication required, as I discovered during a test drive of a Tesla Models S, something more like aviation than automobility emerges thanks to the linchpin of autonomy: auto-pilot.[49]

It is a cold, windy, winter evening close to Philadelphia, and I am with Tuhina and her two small children at the Tesla showroom at a mall. Tuhina is a doctor, who wants to test-drive a Tesla; she is considering buying a car with Auto-pilot so as to free up time on her long commutes. It does not feel like we are driving because I cannot hear or feel the engine. With the car in Auto-pilot Tuhina finds herself not paying attention to the road. She remarks,

[49] Test drive with Dr. Tuhina Raman, March, 2018. See appendix 2 for details.

"it is so easy to forget you are driving. It is so smooth!" The Tesla representative demonstrates that Auto-pilot is a small switch that is flicked on with a beeping sound to indicate that it is engaged; a different beep indicates when the car is out of Auto-pilot mode.
"The best way to think about Auto-pilot is as a Cruise Control function. By stepping on the brakes-" "You disengage Auto-pilot." Tuhina completes the sentence for him.
"Exactly; We ask you to keep your hands on the steering wheel at all times...the steering wheel is going to move on its own so rest your hands there. Don't fight the wheel, just let the wheel guide you", says the representative. Tuhina finds that she cannot get the car to stay in Auto-pilot. She is either holding the wheel too tight thus preventing Auto-pilot from engaging because the system reads her grip as control; or, she holds the wheel lightly enough to engage Auto-pilot, but grabs it tighter when she is confused by its decision-making, thus disengaging Auto-pilot. In one instance, she wants to let the car behind overtake her, and as soon as she does, the beep indicates she has taken control from the car that was in Auto-pilot. In another instance, she finds the car in Auto-pilot overtaking another car a little faster and closer than she would have liked: "whoa, that was close" she says, visibly confused by what the car has just done. Auto-pilot is constantly beeping, signalling that it is being repetitively engaged and dis-engaged.

What seems to frustrate her even more is the gentle and persistent instruction from the Tesla representative to "let Auto-pilot do its job". "I can't get this to work, it's like you have to learn to drive all over again!" she exclaims exasperatedly.

As investigations of AV accidents show, the human, notorious for not paying attention, is freed from paying attention by automation that never loses attention; yet she pays a tragic price for inattention when something in the automated system, ie computer vision, fails to respond to a sudden change in the environment, and requires her attention to manage. In aviation and in autonomous driving, the human is in the role of monitor and manager: airplanes are now "an ensemble of technologies that process and present data to pilots, manage the relationship between pilot input and aircraft response and fly the plane automatically." (Oliver et al, 2017, p 732). Autonomous driving requires human intervention characterised by the language of handovers between human/operator and car/machine that has been developed through aviation safety practices, notably auto-pilot (Cummings, 2014, p. 7). A dyadic workflow of human and machine is not just about efficiency and productivity but about safety and accountability as well. In my field and desk research, I find that 20th century aviation engineering and safety design have been influential in shaping human-machine relations in terms of what exactly machines do better than humans, and vice versa, and what is best pursued collaboratively.

These concerns of human and machine capabilities are now transported to AVs; for example, the SAE's levels of autonomy mentioned earlier emerged from studies of human factors in automation design across industrial contexts (Sheridan, 1992; Jones, 2015: 107-112). The now-discontinued Fitts List aka 'MABA-MABA' (Machines Are Better At-Men Are Better At) is an example of a 1950s heuristic of what humans and machines were thought to perform better than the other; this emerged from the systematisation of national Air Traffic Control in the United States (Cummings, 2014). The Rasmussen skills-rules-knowledge (SRK) model applied in aviator training is a fine-grained approach to human-machine collaboration that makes a distinction between tasks as skills-based, rules-based, or knowledge-based, each of these being differentiated by how they unfold under conditions of uncertainty in the environment (Op Cit, p. 3). So skills-based tasks such as landing a plane, or parallel-parking a car, are well-suited to automation because they entail a routine, specific set of steps. But landing a plane under adverse conditions requires expertise plus intuition, and judgment sharpened through a variety of experiences; and thus is notoriously hard to formalise as requirements of an automated system (Op Cit, pp 4-6). The more formalised, specific, and certain a task is, in a particular environment, the easier it is to automate. This echoes the engineer's prediction, earlier in this chapter, that 'autonomy' emerges in relation to an enabling and supportive environment. Cummings finds "brittle" computer vision systems cannot perceive and process uncertain environmental conditions, and thus cannot become part of an automated workflow. The case of a pedestrian who wheels her bicycle across a road at a point where she should not constitutes such an unstable or uncertain environmental condition. It is thus not something that a human driver can know or preempt. Additionally, the automation of a set of tasks presents its own problems; Bainbridge finds that there are various conditions of increasing automation in which human skill decreases. (1983, pp. 775-777)

There is a heuristic based on Marc Andreesen's infamous pronouncement that "software [is] eating the [automotive] world" (Deloitte, 2018, p. 54) that compares the millions of lines of code in the first space shuttle, a Boeing 787 Dreamliner, a contemporary car, and the 'fully' autonomous vehicle. The space shuttle has the least numbers of lines of code, and the fully autonomous vehicle has the most. This is a familiar statistic because the human factors researcher I interviewed for this work, who conducted the study comparing human and computational adaptability in driving mentioned earlier, repeated the same comparison to me. Aside from the anecdotal observation that perhaps it is Deloitte that shapes the future of autonomous driving, I want to emphasise how common it is to compare the autonomous vehicle to an airplane although these are two different kinds of complex and high-end engineering. While there might be much to learn about the shaping and emergence of the AV from aviation, AV is a big data infrastructural system, and not a car that is meant for everyday use; and hence is nothing like an airplane nor its industry. What we are seeing a change in is what *driving* means, its environmental, socio-technical, and sociological conditions, not just the change in terms of who or what the *driver* is.

Of knots, loops, cascades, crumple zones

Another aviation engineering import, the 'human in the loop' (Jones, 2015, p.134; Marra and McNeil, 2012), has shaped legal accountability in robots and autonomous technologies. Both an evocative metaphor and a practical guideline, 'the human in the loop' is a safety mechanism; auto-pilot and the expectation of the human taking over control from a floundering system, is the most common example of human-in-the-loop. Meg Leta Jones however identifies problems with this conception saying that the human has *always* been part of the loop and cannot be erased or shifted out. She identifies the irony that US automation law builds on this notion of humans and machines as separate and joined by a loop, thus not acknowledging the inherently socio- technical nature of automation; and thus even as it proposes to protect human values, it actually results in less protection because it understands the two as separate (Jones, 2015, p. 81). What she is arguing is that the very notion of aviation is already a socio-technical system, one that is intrinsically human and machine. To think of aviation, or indeed autonomous driving or robotics, as simply about machines and that a human must be added into this is where the problem lies, she argues. So Jones proposes that the law—and perhaps accountability regimes more broadly—must break the loop and "tie a policy knot" instead, with the contexts of design, implementation and social relations. This becomes critical within the big data-infrastructural aspect of the AV; there are many different layers and locations of humans within multiple workflows that the rigid and repetitive dynamic of a loop neither identifies nor addresses. I return to this knot in my concluding statements. For now I want to stay with Jones' emphasis on the socio-technical nature of automated systems, that they are neither just human nor machine fitted into each other, but are a productive imbrication that cannot be easily disentangled. This poses a significant challenge for law, legal liability, and regulation.

However, it is not just the mis-recognition of the human-machine relationship alone that is at issue. To recall Bratton's blur, this dense compression of history and future requires means of accounting that acknowledge "the intersecting complexities of computational globalization, its thickened geographies, its mysterious weaving of geometries of governance and territory, seen on their own terms, not as transgressions of some other system." (Bratton, 2015, p. 14)

We do not have these means yet, which is why I reiterate that an interrogation of the entangled shaping of the ethical from that of autonomy is crucial; and current iterations on Trolley-style problems are likely to fall short.

According to the measures and language of autonomous driving, the transition from automation to autonomy will occur through automated decision-making. Such a scale of decision-making is subject to the 'cascading logics of automation', meaning that one instance of automation necessitates another; a large scale automated data collection can only be analysed through a similarly large-scale automated process, and not manually (Andrejevic, 2019, p. 8). Automation begets more automation; anything other than this is friction that will slow down the process. Cascading logics makes it possible to argue that only automated and scaled decision-making can be deployed in regulation of other large-scale, automated systems; this is a point often made in the argument for lethal 'autonomous' weapon systems (Asaro, 2019; Suchman and Weber, 2016). It is typically this kind of situation that makes it appear that the human is erased from driving. But in fact what is erased is not so much the human, but *how* humans make judgments and decisions. And driving is an area where humans are generally shown to make poor decisions. Machine decision-making is not just fast, but is also efficient and correct precisely because of the 'god trick' (Haraway, 1988, p. 581) of seeing everything from nowhere, or 'objectively' as big data technologies are thought to. Consider a metric by MobilEye[50], the esoteric-sounding, "time to reflect reality" or TTRR (Mattern, 2017, n.p) that captures this god trick. TTRR refers to the time lag that exists between pushing the "mas-

[50] MobilEye is an IBM-owned, Israeli software company that is the leading maker of LIDAR technology, the mapping and computer vision technology uniquely made for driverless cars. https://www.mobileye.com/

ter map in the cloud [to] cars in the field" (Ibid). The empirical decisions and actions made by the car are contingent on this lag. What the lag denotes is that the world between the cloud and car changes. As a result the car does not 'know' how to react and what to do when faced with an uncertain environmental condition, and would have to hand back control to the driver. To minimise TTRR and the risks associated with this transition, MobilEye and Tesla have recently developed 'road experience management' and 'fleet learning' respectively, to ensure that that the ever-changing world is constantly being mapped, made legible, and accessible to driverless cars.[51] It is possible that these systems will work, but they could also fail. They attempt to perfectly collapse territory on a map into a single ontological entity (a key observation by Hind and Gekker, ahead) Whether this will work, we cannot be sure, and yet this is the ambition of autonomy. Given this scale of automated decision-making that co-exists with the world as it is, arriving at accountability is extremely challenging. In the case of the AV crashes and computer vision, we see that it is near impossible for a human to actually intervene except at the very last moment even though, ironically, she is also involved in assessing those images, possibly, in some distributed image-annotation farm far away.

At the time of writing, it was reported that the test driver in the Arizona crash, Rafaela Vasquez, was found guilty of 'negligent homicide' resulting in the death of Elaine Hertzberg. Arizona's easing of testing norms in a bid to "lure" AV companies, and the Volvo/Uber test vehicle's failed technologies already mentioned, were found to be at fault (Levin, 2020). It

[51] Road experience management or REM is MobilEye's approach to minimising TTRR by architecting fine-grained algorithmic systems that are constantly recording and processing the world, adding them to the cloud in small packets, and layering this data with rich "semantic" details to capture local conditions https://www.mobileye.com/our-technology/rem/ Tesla's 'fleet learning' approach is to have everything that all cards are recording to be shared across the entire system. In both these approaches, we see a distinct move away from

was possible to identify that the test driver, Rafaela Vasquez, spent 34 percent of her time looking at her phone streaming a TV show; that in the three minutes before the crash, she glanced at her phone 23 times; and that she looked back at the road a second before impact (National Transport Safety Board, 2019). However neither the companies nor the state were eventually liable, only Vasquez ended up in a "moral crumple zone": humans are ultimately responsible for failures of more advanced software that are supposed to replace them (Elish, 2019). In fact, this is a well-documented fact in studies of auto-pilot in aviation as well, that humans are always ultimately responsible for more competent and efficient machines (Elish and Hwang, 2015). This irony might be compounded by the Autonomy-Safety paradox: "as the level of robot autonomy grows, the risk of accidents will increase, and it will become more and more difficult to identify who is responsible for any damage." (Matsuzaki and Lindemann, 2016, p. 502). It is possible that this will recur, and thus there is a real urgency to research, develop, and regulate going forward. At minimum, we might begin by recognising the displacements humans inhabit as workers, managers, overseers, drivers, consumers, and other relevant publics. All the crashes discussed earlier, including Tuhina's test drive experience, are evidence of the human operator/driver in the difficult role of having to be simultaneously vigilant and relaxed so as to take over at a moment's notice; and particularly in the context of the auto-pilot, the technology that makes autonomous driving appear 'real' in the sense of the car being *self-driving*. But, as these recent crashes indicate, neither computer vision nor auto-pilot are robust. Consequently, surveillance and monitoring of human drivers has become a part of the shaping of 'autonomy'. AV testing requires that a driver-facing camera be fitted to record and monitor driver behaviour, physiological states, and affect. Despite research that these technologies for monitoring and managing drivers' and passengers' states and moods, like road rage and driver fatigue, are incomplete at best, and invalid at worst (Barrett et al., 2019).[52] No doubt this surveillance data will protect car and ride-sharing companies against future liability if drivers are found to be distracted.

Conclusion

It is difficult to work backwards from the moment of a crash and identify what actually caused it; the more complex the vehicle, the more complex the cause, as well as the process of identifying it. Crash prevention and accountability in high-end engineering is a mini-industry unto itself, requiring its own models, where functioning and malfunctioning become the sites of observation, testing, maintenance, and further research. (Downer, 2007, p. 22) And nowhere is this more true than in the aviation industry. Here, identifying the cause of an accident, a "single point of culpability" (Brown, 2006, p. 378), never seems to arrive; what emerges instead in this Byzantine exercise is a deep entanglement between human actions and the perceived agency of technologies. There is a "recurrent strain between a drive to ascribe final causation to human factors and an equally powerful, countervailing drive to assign agency to technological factors" (Galison, 2000, p. 4); and so there is always inevitably "contested ambiguity" between human and machine, between what is technological protocol, and what is human judgement (Op Cit, p. 40). Human action and material agency are entwined to the point that causal chains both seem to terminate at particular, critical junctures,

[52] The artist, Paolo Cirio maintains Sociality Today (https://www.sociality.today/about/), a database that tracks patents related to affective and social manipulation. In this, 40 "affect patents" can be identified in response to the search term 'driving'. Affectiva, a Boston-based software company, is a frontrunner in developing this for autonomous driving: https://www.affectiva.com/

as well as “radiate out” towards human interactions and organisational cultures (Op Cit, p. 4) Thus accident reports are always uncertain and unstable; Galison refers to this as ‘disorder’ and ‘nightmare’, and that the work of historical writing is to attempt to keep them at bay. Galison is referring to what we learn from the socio-technical, and Latourian agency-that our appreciation of human-technology systems has to regard the two as compounded rather than merged.

This is why the Bratton-ian stack of computational governance offers this work a compelling frame. The multivalent ontologies of big data infrastructure and automobile bring us to that which is difficult to exactly name, but is strongly experienced. What happens when we have multiple scales, regulatory systems, automation histories, heteromated humans, and computational sensing, perception and analysis resulting in a crash is not quite like a conventional car crash or airplane accident. We are not dealing with well-mapped, tightly closed and coupled systems. We are in a more fractured place where the ironies of autonomy coalesce into the blur, the “oscillation between the real-but-as-yet-unnamed and the imagined-but-as-yet-not-real” (Bratton, 2015, p. 13) How might this fraught place suggest a way to think about crashes and what they mean for the shaping of autonomy?

For one, Bratton says that blurs lead to accidents. He takes Paulo Virilio’s famous axiom that the invention of a new technology is also the invention of a new kind of accident, and turns it around, saying that the opposite is also true: “*the accident also produces a new technology.*” (Bratton, 2015, p. 17; emphasis in original). Bratton is not just suggesting circularity or co-productivity, I think, but is gesturing to a mode of computational governance of the stack wherein it generates its own logics and practices to sustain itself. And, by 'generating', I do not mean in the sense of 'creating', but rather, necessitating; so the crash—real or speculative—produces technologies that attempt to *fix* the environmental and infrastructural world around the robot car so that it might function effectively. Thus, the technologies of the AV crash are increased algorithmic and automated surveillance and monitoring of human managers; a requirement for improved infrastructure for the multivalences of the AV to work, the world, as it were, for the robot car that would result in the driverless car already being here. Just as the AV's data and AI infrastructures make demands on the human mind and body to 'lean in', be empathic, and work 'as a service' (Irani and Silberman, 2013), so do many other such systems. "Humanly extended automation" is a trend in highly digitised work environments like Amazon's 'Fulfilment Centres' where humans and robotic technologies shift tasks between each other depending on their own competencies. (Delfanti and Frey, 2020) Heteromation turns humans into "computational components", just as 'humanly extended automation' relies on "living labor [to] mak[e] up for machine shortages" (21) It would appear 'autonomous' driving is less about the promise of freedom *from* the car as promised, and is perhaps *for* the car.

Another important aspect of the crash is how it exposes the many layers of *re-mediation* underway. I draw on re-mediation theory not only because the AV is a media artefact in terms of it being *both* big data, *and* a conventional automobile. Rather, re-mediation and in fact, all media studies, is consumed with the creation and maintenance of representations and realities; re-mediation is a place from which to understand the shaping of autonomy as a discursive representation: of freedom, mobility, and the authenticity and legitimacy of AI. Hence, I want to bring together aspects of the AV and AV crashes in terms of re-mediation theory to understand what we are witnessing in this. According to re-mediation theory (Bolter and Grusin, 2000), media are what perform re-mediations of *other* media; and are characterised by the jarring simultaneity of the transparency of mediation, and the actual social and material density of mediation, or hyper-media. Thus, in every experience of media we are encountering both the seamless erasure of the fact of media, as well as hyper-mediation, that creates multiple acts of representation that we come to enjoy and experience as—media. The emergent driverless car suggests multiple sites of re-mediation. Quite prominently and obviously, in the sense of new media being a recombination and transformation of older media, the autonomous vehicle is re-mediation of the traditional automobile. It proposes to erase the material immediacy of driving, to make it seem as if we are being transported without feeling mobile or having to deal with the embodied stress and challenges of being in and negotiating traffic. However it is in terms of the computational and big data infrastructures underpinning the driverless car that we find multiple, ongoing kinds of re-mediation underway. For instance, in the worlds made by and for the driverless car, there is no territory, there is only a constantly updated map that exists in the cloud, which must be 'pushed' to AVs in the 'field'. This is data-space, or the space of data, rather than what a human might understand as territory per se; this data brings space and time together in an odd *pas de deux*, with the production of the map being contingent on the Time To Reflect Reality, or TTRR as it is known to people in that industry. The lag that constitutes TTRR is that hypermedia moment, when the reality of infrastructural delays, distance, and simply time itself, prevent the production of the map space that the AV will navigate in the smooth performance of autonomy.

Sam Hind and Alex Gekker propose the wonderful new ontology, that of the *map-territory,* a flat ontological plane in which neither map nor territory subsume or precede the other, and act as and in relation to each other. Their discussion draws on the work of Michel DeLanda, and big data theorists, Rob Kitchin and Martin Dodge, to arrive at map-territory, a never-complete, always-being formed, compound entity, not unlike auto-pilot. They suggest that map-territory accompanies the auto-pilot, a replacement of the car-driver hybrid, and thus enabling navigation and driving to collapse into each other as well (2019). Hence, their key argument is that driving becomes all about data-based, map-based navigation. As I do, they arrive at a similar conclusion, that this entire infrastructure is being generated for navigation, and hence for autonomy. I arrive at this place in a slightly different way, with the multivalent ontologies of the AV, and by starting with the many dimensions of the crash as emergent therein. All of this is a complex set of practices of making conventional driving disappear; and right on the heels of this comes the hypermedia moment: the disengagement, and the crash. In fact, every single disengagement is a mini-crash in itself.[53] The human has to take over because the map does not correctly perceive its environment, and navigation/driving stalls. The moment of the disengagement, the faulty handover, and the AV crash, might be read in terms of Alexander Galloway's provocation that in the moment of a crash a computer potentially becomes 'non-media':

[53] I am grateful for this key insight by Dr. Nishant Shah, which has been very productive in organising these ideas.

> A computer might remediate text and image. But what about a computer crash? What is being remediated at that moment? It can't be text or image anymore, for they are not subject to crashes of this variety. So is a computer crash an example of non-media? (2012, p. 21)

I argue contrariwise that the ontological multivalence of the AV renders its crash into something that is *not non-media*; If the AV cannot re-mediate data into the navigation/driving during a crash, then it becomes a car that needs a human. The AV as much as it is big data is also a car, it is not just a computer. Its crash is also a moment of more than re-mediation; it is also a site of human death. It is a site of deep fracture between conventional understandings of the world and technology, and unknown relationships yet to be mapped. To conclude then by returning to the blur. The blur is also the place of the Trolley Problem and Moral Machine, the imagined-but-not-yet-real crash. The search for the end-point of culpability for crashes is perhaps how AV ethics is being shaped. This is the logic of the Moral Machine Problem: that if we can reason—logically, or algorithmically—our way out of a conundrum, we might be able to architect fewer crashes, and better driving, from the top-down. For reasons we are about to examine, I believe it does not quite turn out like that; I find that crashes that occur may be the ones authored by risk calculations, that we have no way to hold to account. I turn to this in the next chapter about machine ethics, the Trolley Problem, and the Moral Machine.

CHAPTER 5. THE ETHICS OF AUTONOMOUS DRIVING

Ethical apparatus

In tracing the intellectual history of AI, Matteo Pasquinelli notes that the science of thermodynamics was developed only after the invention of the thermal engine. Similarly, he says, Alan Turing did not come up with a test for differentiating machines from humans on the basis of a notion of intelligence; what he *did* invent was the imagination of a universal computing machine that could achieve all possible functions. (Pasquinelli, 2017, n.p) He wondered what might be required of a system that could make fine-grained distinctions between people in order to distinguish them. The formalisation and automation of specific aspects of human intelligence only came later. Similarly, it is important to study as best we can the practices of making 'autonomy' or 'ethics' because they alert us to the epistemological formations that are in emergence. In this chapter I discuss how academic and industrial research has organised itself around shaping a discourse about ethical decision-making as a part of autonomous driving; its reliance on the machine ethics branch of computational ethics; how experts from influential centres such as MIT Media Lab have advanced this thinking; and what the implications are of setting up such an incalculable problem for a computational system to solve. At one level, we might see applications of the Trolley Problem as laughable, or even slightly ridiculous. I believe this 'ridiculousness' carries some important insights into the nature of the problem that informs 'ethics'. As I find in interviews, the Trolley Problem was intended to inspire engineers to consider the possible outcomes of designing 'fully' autonomous vehicles. It has also made for captivating press about emerging AI and autonomous systems, and has fuelled the popular imagination, as understood by a scan of the internet's meme factories. It has influenced machine ethics-inspired thinking about the design requirements of 'ethical machines'. All these come to have influence in the shaping of this technology and its emergence in society. Studying these practices gives us cues and clues as to what comes next in such discursive and epistemological terms. The stakes in identifying the epistemologies-to-come in the driverless car context relate to the following: a collapse of the differences between human and machine decision-making, for the machine to pass as human in a variety of tasks, in this case, driving, and what is happening to our understanding of 'ethics' as a practice of reasoning about and living in the world? What kind of world model is set up through the provocation of a thought experiment that models a certain reality, and what might this model mean for the shaping of the world in its terms? To address these questions I first present an overview of the machine ethics and value alignment domains to situate the Trolley Problem and the Moral Machine in a specific theoretical tradition. Then, I turn to an explicit analysis of the Trolley Problem, and its discursive influence through experts, including responses from philosophers and ethicists I have interviewed. Finally, I conclude the chapter with a discussion and analysis of the Moral Machine project that appears quite similar in form to the Trolley Problem but ends up diverging significantly from it.

Machine ethics

Machine ethics is concerned with the design of *'ethical machines*', or machines whose behaviour is intentionally aligned with relevant ethical principles. (Anderson and Anderson 2006; McDermott, 2008; Cave et al., 2018) This domain is:

> concerned with giving machines ethical principles or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making … this would mean they could function autonomously … without human causal intervention after they have been designed for a substantial portion of their behaviour" (Anderson and Anderson 2011)

For philosophers and ethicists, the architecture of human ethical value systems into electronic and computational systems is a detailed and extensive area of specialisation. A similar approach, '*ethically aligned machines', ,* is adopted by the Institute for Electronics and Electrical Engineers (IEEE) in their multi-year ethically aligned design (EAD)[54] standards development project[55] specially for AI and autonomous systems. This project relates to standards develop-

[54] https://ethicsinaction.ieee.org/

[55] https://standards.ieee.org/initiatives/artificial-intelligence-systems/standards.html#p7000

ment for machines “whose behavior adequately preserves, and ideally furthers, the interests and values of the relevant stakeholders in a given context.” (Cave et al., 2018 p. 3) James Moor has a widely cited categorisation of ‘ethical agents’ into four types that makes distinctions between machine ethics and other potential approaches to ethics and design (2006; Winfield et al 2019). In this schema, ‘Ethical Impact Agents’ are “any machine that can be evaluated for its ethical consequences”, where such a ‘machine’ can be anything from a chatbot to a facial recognition system. ‘Implicit Ethical Agents’ are those machines that are designed to avoid unethical outcomes, such as an automated teller machine (ATM), an auto-pilot setting, and even a pharmacy that has a duty of care to sell only safe and licensed products. Explicit Ethical Agents are machines that can reason about ethics; and ‘Full Ethical Agents’ are machines that can make explicit moral judgments and justify them. Moor is quick to caution however that a “bright line” is understood to exist between machine ethics approaches and ‘full’ and ‘explicit’ ethical agents, like humans; we would need to be very sure that we even want machine systems to be designed as such.

Adjacent to machine ethics is “computer ethics, robot ethics, ethicALife, machine morals, cyborg ethics, computational ethics, roboethics, robot rights, and artificial morals.” (Yampolskiy, 2012, p. 389) Similarly, computational ethics draws on simulation, agent-based modelling and mathematics to provide a “descriptive as opposed to prescriptive”, means to “explore how multiple individual agents interact with each other as well as the agent society with regard to a moral dilemma, thereby providing a means of analyzing the evolution of an emergent ethical system of the agent society” (Ruvinsky 2007, p. 76-77); “[to] formalize philosophical definitions that are available in natural language and to translate concepts that can be programmed in a machine and can be understood easily while getting rid of ambiguities" (Bonnemain et al, 2018, p. 57). While Ruvinsky's description sounds like an approach to appreciating the diversity of outcomes from a situation to stimulate thinking and further planning, the latter definition appears aspirational, one in which natural language replaces computer language, which is, in other words, the ambition for AI.

There are three kinds of machine ethics approaches: bottom-up, top-down and hybrid approaches (Abney, 2011), modelled on how humans are thought to make moral judgements: a "hybrid of both (1) bottom-up mechanisms shaped by evolution and learning and (2) top-down criteria for reasoning about ethical challenges." (Lin, 2008, p. 38) These approaches loosely mirror approaches to AI programming: top-down programming that echoes symbolic AI or GOFAI, and bottom-up, connectionist alternatives. There is however another element of the connection to AI in machine ethics (aside from the convergence of an imaginary of Artificial General Intelligence as so humanoid that it is able to perform ethical reasoning): the emphasis on rules and game-playing . In his thinking about how an 'explicit ethical agent' could be architected, James Moor asks if it could "do ethics like a computer can play chess [?]. Can a machine represent ethics explicitly and then operate effectively on the basis of this knowledge? (2006, p. 19-20)

Never mind that Chess did not eventually offer much to the development of AI except to waylay it, perhaps[56] (Ensmenger, 2012), Moor clarifies that machine ethics is indeed about architecting rule sets and to identify where these rules may violate another set of ethical frameworks. To imagine an 'explicit ethical agent' as one that acts according to a set of finite rules akin to Chess, an extremely popular and well-documentedmapped game,is odd but revealing. It reveals a fissure between two dominant approaches to AI and ethical decision-making. On the one hand, there is the imagined potential for unethical outcomes framed within the narrative imaginary of Artificial General Intelligences portrayed through cinematic characters like HAL in *Space Odyssey 2001*, or Ava in *Ex Machina.* These powerful fictional portrayals mark the distinction between human and non-human decision-making and the role of ethics as distinctly human and as an aspect of human cognition. In parallel there is the development of machine 'cognition' (i.e, 'intelligence') in terms of competitive game-playing that requires the negotiation of well-documented rules. In the case of Alpha Go, discussed ahead, it was about identifying winning moves from a vast database of existing moves. These developments in parallel suggest that ethical decision-making can be modelled through the development of 'lower' cognitive exercises such as rule-following, that games offer, and that with time, and 'learning', higher cognitive functions, such as 'ethics' might be achieved. This rationale is architected on the dominance of analytical-philosophical approaches to ethical decision-making.

[56] Ensmenger's history of Chess in the evolution of AI reveals that it did not advance the study or development of human intelligence in AI technologies; all it did as to build a computer that excelled at beating Gary Kasparov, the reigning Chess champion in 1987.

The remit of machine ethics is fairly controlled and specific, and requires details of contexts to be clearly spelled out for a machine system to explicitly recognise and then solve problems In demanding attention to detail, they appear limited in scope, which is not necessarily a cause for concern. On the contrary, it is useful to have approaches to architecting values that demand detail of the world; the problem is perhaps one of scaling up these approaches to the level of the world. In the next section I discuss machine ethics approaches mapped onto applications in the driverless car/autonomous vehicle ethics context.

Top-down, bottom-up, and hybrid

Top-down approaches use specific standards or rules to guide the development of decision-making architectures, and thus test the limits of how to describe "the situation of the world" (Wallach and Allen, 2009, p. 88). Such approaches determine an architecture that directs what the path to selecting certain values should be, to consider all future outcomes, and those outcomes that could be ignored. Top-down approaches are so precisely determined that they tend to be inflexible, and the rules may not be applicable to every new instance that evolves. For example, balancing personal ethical choices with that of other people's choices is an enduring question in machine ethics, and particularly in the format of the Trolley Problem; if everyone had their own "personalised ethics setting" (Goggoll and Müller, 2017) for which way the car should turn, we might end up with a destabilisation of social morality because everyone chooses to save themselves when the brakes of a driverless car fail. However, you cannot then have a top-down 'mandatory ethics setting' because it would be near-impossible to arrive at a lowest-common denominator approach that serves *everybody*.

On the other hand, ‘bottom up’ approaches sift through individual values and responses to identify patterns of broadly held responses, and on the basis of this, ‘learn’ appropriate patterns of behaviour, and then use these patterns to architect rules. Bottom-up approaches can be an assembly of discrete subsystems each performing unique functions with distinct capabilities; or it can refer to “the emergence of values and patterns of behavior in a more holistic fashion.” (Lin, Abney and Bekey, 2008, pp. 34-35) Bottom-up approaches invoke ‘democracy’ and ‘voting’ as ideal practices that can be applied to determine outcomes from a diversity of data points. (Noothigattu et al, 2017) An example of the bottom-up approach to computing ethics is the Moral Machine Project that crowdsourced 39.6 million responses to an online game scenario involving a driverless car whose brakes had failed. (Awad et al, 2018) While this is not the basis for any actual AV ethics proposal, its visibility and popularity in shaping the discourse is undeniable, as evidenced through a citation in *Nature Machine Intelligence*, and in various talks across industry, academia, and policy settings. I discuss this project in more detail ahead in this chapter. However, scholars point out that bottom-up approaches are tenable when there is only one goal to achieve, but more challenging when there is a diversity of possible outcomes (Wallach and Allen, 2009, pp. 113-114). And this is why bottom-up approaches are nothing like voting or democracy; in the latter, there are fixed options to choose from resulting in just one choice, unlike ethics in which different positions and contexts generate multiple possible outcomes. Bottom-up approaches tend to be less effective when there are no fixed goals to work towards and the system has to determine appropriate paths. Goals act as constraints of sorts, but in the case of ethical decision-making it is the goal itself that is in question: which way should the car swerve?

In bottom-up approaches to programming decision-making, the system has to rely on a training data set to identify the structure of responses and identify the best choices and outcomes. That said however, such an approach risks learning potentially negative outcomes, or just very specific ones because of what might be in the data set. A bottom-up approach-based example project in the AI/machine ethics literature is GENETH, a 'general ethical dilemma analyzer' (Anderson and Anderson, 2014). It addresses the problem of disagreement between ethicists (or, broadly, between ethical approaches) through the combination of casuistry (a branch of applied ethics that takes a case study based approach by responding to "new ethical dilemmas by drawing conclusions based on parallels with previous cases" (p. 20), machine learning, and inductive reasoning.Thus GENETH takes online internet data as the basis for a training data set that would identify patterns in opinions and use this to architect rules for action. Dating back to 2005, the authors might get a pass for not imagining how flawed an approach this is, given what we know now about the quality and value of online opinions. For example, Microsoft's disastrous Tay bot 'learned' from being exposed to violent and hateful speech online (Gibbs, 2016). Further, only a truly omniscient computer or 'world agent' would be able to capture the diversity of preferences in the world and synthesise them. (Wallach and Allen, 2009, p. 88) 'Ethics Bots' are another bottom-up proposal, described as "front line worker bees" that scoop up personal data about an individual (this is the 'bottom-up' part), and assess the individual's needs and preferences across a range of domains, from shopping to driving to volunteering (153). The authors describe it as an AI program that will assess a person's moral preferences by examining their views by surveilling their computers! (Etzioni and Etzioni, 2016, p. 152). Significant concerns around privacy and security casually pushed aside, this proposal for 'AI assisted ethics' takes the case of bots that monitor the settings chosen by the user for their home heating and cooling system, Google Nest. By monitoring the settings on a thermostat over time, the Bot arrives at the user's preferences, which might be considered the baseline for future personal choices. For example, this might include a personal setting that favours less energy use. With a personal, data-munching, surveillant bot, the authors conclude that this approach offers more honest choices; and thus can be scaled to choices made in the driving of a car by encoding a range of personal preferences of the driver or owner over time, and thus arriving at ethical preferences. Similarly there is Ethics Net, a 'game' inspired by the historic Image Net challenge[57], a long-running engineering challenge to assemble an automated image annotation database. 'Ethics Net' proposes to generate a dataset of prosocial behaviours for AI to learn how to be ethical: A "community teaching prosocial behaviour to kinder machines" that promises, "together we can act as role models and guardians to raise kind AI by providing virtual experiences." Framing AI within the metaphor of childhood, Ethics Net is doing experiments into collating datasets that reflect "many different cultures, opinions and creeds, and which can expand in scope and nuance over time, to empower socially-aware thinking machines for generations to come."[58] They even have a browser extension available on the Google Play Store that allows internet users

[57] http://image-net.org/index With the emergence of big data and machine learning, Image Net is no longer needed in the same way; now, machine learning has improved to the point where image recognition is very good thanks to humans in the system who annotate and tag images.

[58] https://www.ethicsnet.org/

to tag kind and prosocial behaviours they come across online to contribute to a database for future AI. [59]

'When you maintain a top-down view of the world, everything seems bottom-up' goes a Mongolian proverb (in McKenna, 2019). Bottom-up approaches identify implicit values and biases and offer opportunities for flexibility and diversity. Top-down approaches are explicit and therefore can be useful when goals are uncertain. However, bottom-up approaches risk being vague and limit the achievement of a clear goal; and top-down approaches can be inflexible and over-determined, not adequately capturing the realities of a dynamic world. (Wallach and Allan, 2009, pp.117-118). The reality, however, tends towards a hybrid. Eventually, argue philosophers, morally intelligent robots need to maintain the dynamic and flexible morality of bottom-up systems capable of accommodating diverse inputs, while subjecting the evaluation of choices and actions to top-down principles (Lin, 2008, p. 38) In the context of AV ethics neither rational ethics approaches (such as those modelled on the Trolley Problem or other deontological approaches and typically 'top-down' approaches), nor machine-learning based approaches (that learn rules based on data sets of prior human behaviour and are therefore 'bottom up') are practically useful.

The Trolley Problem presents a case of "value incommensurability" because it positions Utilitarian ethics (valuing more positive outcomes, in this case, saving the five people working on the track) against deontological imperatives (identifying contextual rules for action) that are not equivalent, and cannot be compared. Hence, this requires a hybrid approach to resolve,

[59] https://chrome.google.com/webstore/detail/ethicsnet/djamiamgnjcpjhkknjddilkaibbhmhgc

such as a set of numerical rankings to compare different outcomes of crashes-of hitting a child, an animal, or a helmet-wearing cyclist, or destroying itself to save others, to arrive at an 'objective' choice (Bhargava and Wan Kim, 2018). These rankings take into consideration details such as the make and model of the car and its documented crash worthiness as a metric of its resilience; the environmental conditions; the number of people in each car or vehicle. However, oddly, this approach returns to putting a final decision back in hands of the driver, realising that a moral decision cannot be arrived at by a computational ranking of the outcomes of a speculative crash. This is where bottom-up reaches towards something more hybrid; in this case, requiring a top-down decision made by a human. These rankings could be crowdsourced by manufacturer, the authors note. Such a test could help to establish accident claims under the law because they have proceeded on the basis of a set of norms; and would allow manufacturers to offer their customers a "moral navigation system", much like a menu of Facebook's privacy settings from a drop-down list.

Another hybrid approach might combine a "rationalistic moral system" that will act to minimise the impact of an accident according to "generally agreed upon principles; with machine learning techniques to study human decisions across a range of real-world and simulated crash scenarios to develop similar values; with a natural language output, "so that its highly complex and potentially incomprehensible-to-humans logic may be understood and corrected" (Goodall 2014, p. 12). Other large scale attempts at achieving 'ground truth' for AV ethics have been attempted in a way that combines bottom-up and top-down. MobilEye, a company that makes LiDar technology, proposes Responsibility Sensitive Safety (RSS) in which they attempt to “formalise and contextualise human judgment regarding all driving situations and dilemmas” into rules according to which future AVs might be programmed (MobilEye, 2019, p. 3). Their model reverse engineers 37 of the most common types of road accidents (according to the US National Highway and Transport Safety Authority) to identify the specific mechanical, automotive, human, ergonomic, and environmental conditions of how each one occurs. This becomes the source of data that enables them to identify the ‘definition of a dangerous situation’, and based on this, generate rules that could or should be followed, such as safe distances, to prevent accidents. This work is startling in its attention to detail and how it arrives at precise but narrow rule sets. Eventually, there are limits to the formalisation of ethics into rules that can be logically processed and there is a need for a diversity of perspectives in the world of unexpected and complicated accident scenarios.

‘Value alignment’

Machine ethics is also known as ‘value alignment’ that could be considered as organised along similar lines of top-down, bottom-up, and hybrid approaches. ‘Value alignment’ refers to the desire to align machines to human values. Value alignment is defined in terms of identifying the correct reward function for an ‘AI agent’, a reward function being what a machine learning process should achieve, either by setting it from the top down (which is not easy because it means deciding what the outcome should be), or from the bottom-up (the system identifies the best possible reward function based on existing human behaviour). To get around these challenges that veer between extreme rigidity, or lack of discernment, more hybrid approaches could be adopted. 'Overlapping consensus’, Rawls’ Veil of Ignorance, and

Social Choice theory are three approaches proposed to break out of the challenges presented by top-down/bottom-up challenges (Gabriel, 2020). ‘Overlapping consensus suggests that computational systems identify overlaps in consensus between a wide variety of moral beliefs around the world. ‘Human rights’ is perhaps one of these that finds considerable appeal around the world, whereas religious values and beliefs do not. The ‘veil of ignorance’ assumes that it is possible to assign an ‘imaginary’ position to an agent in society, and choose principles for that society; this assumes that this agent can then deliberate impartially and choose principles that do not unduly favour themselves. However, surely this agent must have some mooring in a society; every one does! And how values can be arrived at without a place in society only speaks for a suspicion of situated-ness. ‘Social choice theory’ is based on US philosopher John Rawls’ work, “to help people with different values and perspectives agree upon principles of justice for a society.” (In Gabriel, 2020 n.p) This approach expects AI designers to identify whose social values have standing in society (and this might include even ‘an AI’ and other nonhuman entities); how to assign some kind of comparative measurement to assess these values so as to include or exclude them (but how would such a value even be assigned?); and finally, to aggregate these values into a single preference (Baum, 2017).

Whether this is referred to as machine ethics or value alignment, these approaches propose that ethical and value diversity must be identified, respected, and integrated into a proposed decisional framework. This could be valuable in narrow and specific circumstances, such as a constrained game scenario, or a home heating system; so, they might emerge as nuanced *in situ*, but appear unlikely to scale precisely because they demand so much detail. However, there is an *the inability to find consensus in the plurality of human values* that runs through these related approaches. How is it possible to hold so much diversity in a single framework? This is a persistent irritation and obstacle one senses in all these approaches to machine ethics and this desire to scale up and automate ethics is what this work interrogates. Across studies of the ethics of autonomous driving, the 'ethical decision' of which way a car should swerve if its brakes fail is about the moment, with little attention for the many sedimented sites, layers, and worlds, where these moments occur. In every case requiring ethical decision-making, a machine ethics perspective is crafting a statement of values, drawing on a particular data set, identifying humans and nonhumans who have 'standing' by methods of identification that may or may not be robust, such as computer vision; and arriving at conclusions. Machine ethics approaches come with risks and limitations that Cave and colleagues summarise succinctly (2018).'Ethically aligned machines' could just fail, and in critical situations this would be disastrous. Such machines may be unable to deal with plurality of values or significant disagreements between them. They also run the risk of minimising human agency in terms of the weakening of actual skills in managing complex systems. Finally, as such systems take on more decision-making, they might be awarded the status of 'moral patients' because of their increased decision-making capacities; in other words, there might be a tendency to think such 'moral agents' also have rights and protections. (pp 8-11). Next, I turn to a case of the Trolley Problem, and then the Moral Machine, which, I argue, are examples of machine ethics approaches to designing 'ethically aligned' systems.

The Trolley Problem

In fiction and in life, the intelligence of artificial intelligence (AI) has been measured by qualitative and quantitative games attuned to the scale of human cognitive capacities. However, the challenges of game-playing are eventually limited; the ultimate challenge of AI is the world itself, to exist 'in the wild' of complex human interaction (Ribes et al, 2019). Thus the driverless car project, which started as a DARPA challenge (DARPA, 2007) in a much-publicised challenge between university-based and industry robotics and AI researchers, was conducted 'in the world' (i.e.a dedicated track and not in traffic) rather than in a simulation. The latest challenge to the intelligence of the AV is its ability to reason its way out of a simply-framed but difficult conundrum, the Trolley Problem (TP), a classic thought experiment. It has never been proposed as the basis for machine ethics, although it has served as a starting point for considering the specific kinds of ethical concerns that arise from autonomous driving. The Trolley Problem is adapted from the Oxford philosopher Philippa Foot's 1967 paper about the Doctrine of Double Effect, which was meant to inspire debate about abortion. Her collaborator in that work, the philosopher Judith Jarvis Thompson, writes about it :

> Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Un-

> fortunately, Mrs. Foot has arranged that there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. Is it morally permissible for you to turn the trolley? (1985)

There are variations of the problem Jarvis Thompson and Foot develop that substitute track workmen and the trolley with: a bystander who happens to be present, sizes up the situation and could 'throw' a switch that would change the course of the trolley without actually driving it (bystander at the switch); the possibility of pushing a very large man into the path of the trolley to stop its motion and thereby save everyone (fat man). All these versions are based on the doctrine of double effect, which "refer[s] to the thesis that it is sometimes permissible to bring about by oblique intention what one may not directly intend. (1977/2002, p. 1) Here, Foot is making a case for abortion by examining the different rationales that would enable it. She deploys a variety of scenarios to reiterate the distinction between direct and oblique intention, what we *do* and what we *intend*, saying that this is the difference between "steer[ing] towards someone foreseeing that you will kill him and … to aim at his death as part of your plan." (Op Cit, p 2). Foot suggests a provocative example of risking the life of a fat man who is part of a group of explorers. He is leading the way out of a cave when he gets stuck, trapping the others behind him. "Obviously the right thing to do is to sit down and wait until the fat man grows thin" says Foot. However, in this thought experiment, floodwaters are rising within the cave. Plus, it turns out that the trapped party have a stick of dynamite with them, conveniently, that could be used to blast the fat man out of the mouth of the cave. Either they use the dynamite, or they drown. The trapped explorers could argue, she says, that they did not intend to kill the man, but that it was a merely foreseen consequence of the act of blowing him up. “We didn’t want to kill him ... only to blow him into small pieces” or even “ ... only to blast him out of the cave.” (Op Cit, p. 2). In this and other examples, Foot makes a distinction between *wilfully doing*, and *indirectly letting happen*. Again, her intention in this work was to eventually argue in favour of a woman’s right to choose abortion In that context, the similar logic would amount to saying that to save a woman who might not want to bear and rear a child does not intend to end the existence of a foetus; this latter outcome is an indirect and foreseen outcome. Foot never intended to provide answers, only to “discern some of the currents pulling us back and forth.” (Op Cit, p. 5) Hence, Foot is saying that the very process of debating, reasoning, and arriving at the answer to such a conundrum *is* ethics. How we arrive at an answer is entirely situated, and personal. And this constitutes a social morality. However, wider social norms are constantly being decided for us through algorithmic decision-making, particularly in terms of shaping cultural, social, and political values. These may not always be related to deeply sensitive issues like abortion but the point remains that while much of human experience is incalculable, our social lives have for decades been organised around calculations, often with the intention of hedging against risk, assigning value, identifying means to compare what is good for the individual against what is good for wider society. While ethics is something we might consider personal, the machine ethics approach intends to level up decision-making through automation the scale of complex social systems such as the city and driving. . Jason Millar and Patrick Lin have been active in developing the Trolley Problem at this scale. I turn to their approaches to the Trolley Problem to establish how it has become influence in shaping ethics and autonomy.

The Millar and Lin Versions

Jason Millar[60] and Patrick Lin[61] have been key figures in shaping the ethics of autonomous driving discourse in terms of Trolley Problems. Both are clear that the question of putting autonomous cars on roads is far more complex than Trolley Problems; Lin, in particular, lays out a variety of concerns that anyone putting driverless cars on the street would have to consider (2015). Millar is committed to ethics being understood in relational terms, as something that engineers have to consider in every aspect of their work. Both reiterated that their versions of the Trolley Problem are only meant to provoke discussion; yet they remain influential more widely. Here are their two versions, below.

> Steve is travelling along a single-lane mountain road in a self- driving car that is fast approaching a narrow tunnel. Just before entering the tunnel a child errantly runs into the road and trips in the center of the lane, effectively blocking the entrance to the tunnel. The car is unable to brake in time to avoid a crash. It has but two options: hit and kill the child, or swerve into the wall on either side of the tunnel, thus killing

[60] Jason Millar, an engineer-turned-philosopher, is a professor at the University of Ottawa, Canada, and a leading name in shaping the discourse of the ethics of autonomous driving. I have interviewed Dr. Millar for this research.

[61] Patrick Lin is a professor of Philosophy, and director of the Ethics and Emerging Sciences Group at California Polytechnic State University, San Luis Obispo, California. For over a decade, Dr. Lin has been studying and writing about the ethical considerations associated with the development of autonomous technologies and robots. Lin's specialisation lies in identifying the ethical dilemmas in speculative applications of AI and autonomous technologies. Recent publications on his webpage include: *Arctic 2.0: How Artificial Intelligence Can Help Deliver a Frontier*, and *DoD Needs New Policies, Ethics For Brain-Computer Links (Jacked-In Troops?)* He is widely quoted in the mainstream and tech media on the ethics of autonomous technology, and consults with the automotive industry around the world. I met Dr. Lin at an AI and Ethics conference in 2018 and subsequently had email exchanges with him about my work, and his. He introduced me to others I interviewed for this work.

> Steve . If the decision must be made in milliseconds, the computer will have to make the call. What should the car do? (Millar, 2015a p. 54)

> Imagine a car on an expressway behind an open-top truck carrying large, heavy items. One of the ropes holding these items together starts to come loose and the large, heavy items look like they may tumble onto the self-driving car directly behind the truck. In order to avoid the objects from falling onto the car and resulting in it stopping, or harming the humans inside in some way, the car could either swerve left or right to avoid the heavy items. To the left of the car is a motorcyclist not wearing a helmet, on the right is a motorcyclist wearing a helmet. Swerving left will allow the car to avoid the heavy objects but risks hitting and killing the motorcyclist not wearing the helmet. Swerving right will similarly allow the car to avoid the heavy items but risks hitting the motorcyclist wearing helmet, who has a better chance of survival because of it. However, should the motorcyclist with the helmet be penalised for being more responsible? Also, how can the chance that even a motorcyclist wearing a helmet could sustain [possibly fatal] head injuries be accounted for? (Lin, 2015)

Jason Millar proposes that this Problem should inspire engineers to think about the design of systems. His work is organised around understanding how technologies act as “moral proxies”, that is with various expert capacities that humans can delegate moral decision-making to (2015a, p. 47-48). Strongly influenced by the design theoristPeter-Paul Verbeek and science and technology studies scholar, Bruno Latour, he cautions that designers might be paternalistic in delegating decision-making to objects. He recalls the example of the seatbelt that Latour uses to lay out such paternalism; that it is a device that straps in the user without giving them much agency; and the example of a shopping cart wheel that automatically locks when it goes out of a certain range of the shopping mall. These technologies work like switches operating on a narrow set of choices and not making space for user autonomy or interaction. So, instead, Millar proposes *autonomy by design* (2015b, p. 180; emphasis in original) wherein designers can mindfully build in ways to *delegate* decision-making to artefacts that *mediate* between humans and the world.[62] Hence, he does not want to program the car so much as enable engineers and designers to 'toggle' between various levels of delegating decision-making to moral proxies while avoiding paternalism.

While Millar proposes engineering design to incorporate ethical concerns, Lin works with automotive companies to map out the different kinds of ethical outcomes they will have to deal with. Lin has had popular media presence with TED Talks and mainstream media op-eds. He is also a consultant to a wide range of automotive companies, tech companies, think tanks, and government agencies. The 'Lin Problem' has popped up with regularity in public talks and articles about AV ethics with citations going back to 2008. In fact, I argue that he has been instrumental in furthering this perspective. Dr. Lin does not disagree with my assessment that he has a lot to do with making the Trolley Problem (or the Lin Problem)quite central to our understanding of autonomous vehicles and ethics. This is confirmed by Jason

[62] Millar also goes on to collaborate with other scholars and designers to develop a toolkit for Foresight in AI Ethics by and with the Open Roboethics Initiative, a non-profit think tank in Canada. https://openroboethics.org/ ; https://openroboethics.org//ai-toolkit/ His work relates to educating designers to become more attentive to the outcomes of their design and engineering practices.

Millar, as I narrate ahead. I met Dr. Lin at an AI and Ethics conference[63], and set up an email exchange to understand more about the problem and its provenance from his perspective. I include the email transcript in Appendix 3 with a screenshot. Dr. Lin writes a response to my email about how and if I should credit him with being the person who made the Trolley Problem popular in shaping the ethics of autonomous driving discourse:

> So, I suppose it's technically true that I'm the first to explicitly connect the trolley problem to self-driving cars, but I stand on the shoulders of giants and their prior work. In any event, this is a pretty obvious connection, at least in retrospect, that I think would have been made eventually by someone else, if not by us … By the way, I had thought about different ways to run the trolley problem scenario (save 5 vs. kill 1) with robot cars, and this one seems to be the most plausible and least problematic—the human is initially in control of the car first, and the programmer must decide whether AI should intervene or not. And here are other variations of the dilemma that press on different intuitions—they're not about 5 vs. 1 but still showcase the same kind of intractable tradeoff of values at their core. *Finally, I defend the trolley problem in this piece last year. While the dilemma might be exceedingly rare, it has happened before, and anyway it misses the point to insist that the dilemmas should be a real problem before thinking about it. Thought experiments are genetically related to science experiments; and science experiments are also "fake" in that they set up artificial conditions that don't obtain in the real world...yet no one complains about that …*" (emphasis mine)

[63] The inaugural AI, Ethics and Society Conference (aiesconference.org) organised under the aegis of the Association for the Advancement of Artificial Intelligence, an academic research body founded in 1979 at Stanford University. The first AIES conference was held in New Orleans, January 31-February3, 2018.

My intention in identifying Dr. Lin and narrating our interaction is twofold: to assert the deliberate framing of ethics in 'self driving' cars, and its epistemic implications; and to introduce how the ethical apparatus also consists of individuals in distinct roles and places who are doing the work of framing future representations of AI, autonomy, and ethics. Lin's words at the end of his email, about Science and knowledge-making, prompt reflection on how 'ethics' is being pushed and pulled into different directions as a field of study, inquiry, or practice. Lin argues for a legitimisation of ethics and the Trolley Problem, in particular, as akin to 'science', arguing that they are 'genetically similar'.[64] Second, he says that Science is 'fake' and 'artificial' anyway because of its reliance on conditions that are not mirrored or existent in the real world. He is arguing that models of the world emergent in Science are *just* models, and not the world itself, and so this legitimises a hypothetical, computational and calculative model of the world such as the Trolley Problem; in other words, this is as good a model of the world as it needs to be—for the driverless car. In effect, Lin is arguing that a thought experiment about ethics conducted inside a model is somehow as valid as science done inside a lab. However, he elides the difference between ethics, or even thought experiments, as a reflective process; and science as a larger, collective, accountable, planned, and consciously orchestrated exercise. Since there was no opportunity to follow-up with these questions, I can only gauge that Lin is referring to this thought experiment as a kind of scientific i.e. public and large scale endeavour. On the matter of Science and thought experiments, — artificially, of course Lin is correct — but he has not included necessary cautions and caveats around such a

[64] Dr. Lin did not respond to further questions about these ideas; the following emails related to scheduling a meeting in California where I was to visit a few months after this exchange.

model before launching it out into a world that relies chiefly on fiction and Hollywood-inspired imaginaries of robot cars. It is also interesting that Lin would use words like “fake” rather than “situated” to refer to experimentation in Science; there is a defensive tone in this email. This might be related to his use of the word “defend” in the sentence just before: there has been criticism of his work and use of the Trolley Problem, which I also reference in this text.

The media-discursive influence of the Trolley Problem

Technology ethicists have a more nuanced understanding of the ethical risks and considerations that AI and autonomous systems pose: “meaningful human control; algorithmic opacity and hidden machine bias; widespread technological unemployment; psychological and emotional manipulation of humans by AI; and automation bias.” (Vallor and Bekey, 2020). A Trolley Problem-style approach is not in this list. Nonetheless, the Trolley Problem and the ethics of autonomous driving have achieved considerable media popularity to the extent that it was referred to as ‘the thorniest dilemma in tech’ (Gholipour, 2016) A bit like a party game, the Trolley Problem has an engaging quality; it draws the listener or viewer in to consider the difficult choice that a hypothetical autonomous system must make, a significant challenge for designers. It fulfils that agenda of marketers and PR: to get you to talk about their product. Why would people not talk about something framed in terms of “how should a car be programmed to decide who to kill?” Not only does it trigger cultural anxieties associated with AI, it also relates to cars and driving which are quite central to modern society. Over the course of this research and right up to the time of writing, the Trolley Problem has been featured in the mainstream technology press, popular TV shows[65], podcasts, XKCD comics, and TED talks, among other popular cultural discussions about technology (Doctorow, 2015; Worstall, 2014; Lin, 2013, 2015; Rahwan, 2016). A Google image search for 'Trolley Problem' results in an endless scroll of playful and sarcastic iterations; it offers a structure that is irresistible to meme-makers of the internet (Zhang, 2017). The Problem's discursive influence in shaping ethics has to do with its media popularity, and in particular with scientists, engineers, and philosophers talking to the media and producing cultural material on this topic. This is a leveraging of the media's discursive power to have influence in the world (Reed 2013, p. 203)

Jason Millar has been centrally involved in the public and mainstream media discussions about AV ethics, and verifies the explosion of popular and academic interest. In an interview he verifies how he and other leading scholars in the field (such as Patrick Lin) have been visible in mainstream media and technology press. He says: "We could get these media interview requests, and between Pat Lin and Noah Goodall and myself, we would have to field them, distribute them among each other … just to deal with the volume."[66] Millar emphasises the point further, about the role of the technology and mainstream press. He argues, 'AV ethics' in terms of the TP has been shaped by the media, rather than by scholars, although he does admit that scholars like Lin and himself were extremely visible in the mainstream media talking about the Trolley Problem and in effect, promoting it. He reiterates that the Problem is *not* meant to determine how ethics is to be defined or architected, yet it has become a touch-

[65] Notably, *The Good Place* and the legal drama, *The Good Wife*.

[66] Interview with Jason Millar, June 2020. See appendix 2 for details.

stone of the machine ethics approach in the autonomous driving context. A more extensive discussion of the media influence on particular technology trends and hype in AI/autonomous technologies would be a valuable contribution to this domain. I include this here because figures like Lin, who work extensively with industry and government, have discursive influence on the field. From consulting with German car companies, to authoring reports for the US Office of Naval Research, to identifying AI applications in the Arctic, Lin is an academic who has a significant intellectual footprint; the discursive frame emerging around new technologies and adopted by powerful social actors, then amplified by the mainstream media.

Engineering ethics

Discussing autonomous driving and ethics with philosophers and ethicists invariably started with them referring to the sudden and unexpected popularity of their discipline. I interviewed four ethicists and philosophers who all agree that the Trolley Problem is not something that can be solved or resolved, nor should it be solved at all.[67] It cannot be solved in a traditional sense, because the two underlying approaches to ethics, Deontological and Utilitarian, are not equivalent. Wong and Çanca are bemused at the popularity of this thought experiment and reiterate that its purpose is to work as an 'intuition pump' to generate themes for the full development of ideas for case study. Further, as Çanca notes, Trolley Problem-style thinking might be good at surfacing questions but the answers can only come from deeper analysis through the application of various other theoretical approaches. Çanca, a bio-ethicist who now works as an 'AI ethicist' vents annoyance at this discussion of ethics. She says:

[67] Interviews with Pak Hang Wong, February 2019; interview with Çansu Çanca, October 2018.

> I read an article by KPMG in which they say there are five future jobs in AI, and AI ethicist is one of them, that you can just re-train your staff to do ethics! They think ethics is something you learn from a pamphlet! Or that this is going to be a policing system. Yes, this is what technologists think, that ethics is there to police them.[68]

Çanca is responding also to the wave of interest in ethics education to engineers; Aside from the interest in frameworks such as the Trolley Problem, there is new attention to Philosophy and Applied Ethics for ethics education for computer science students.[69] The assumption is that more ethical technology might come about if computer scientists were better educated in ethics. This is an idea that has some weight, as Johannes Himmelreich, a philosopher working at a prominent Silicon Valley technology company told me about this, and his role at this company:[70]

> the Trolley Problem was developed and continues to be popular as a way to help computer scientists to think about values and ethics in design of software … And Big Tech companies like to have academics around … they [tech companies] like to learn stuff, and they like to be seen to be learning stuff.

[68] Email exchange with Çansu Çanca, March 2019

[69] There has been a focus on teaching ethics to individual engineers such as the Mozilla Foundation's Responsible Computer Science Challenge that awarded US$ 3.5 million to "promising approaches to embedding ethics into undergraduate computer science education, empowering graduating engineers to drive a culture shift in the tech industry and build a healthier internet" https://foundation.mozilla.org/en/initiatives/responsible-cs/ . Plus there are formal courses to integrate ethics into corporate data science and tech practice https://ethics.fast.ai/ and https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/ for example

[70] Interview with Johannes Himmelreich, June 2018. See appendix 2 for details. Himmelreich was bound by a non-disclosure agreement about his work at this company and hence could not share details of his work.

Himmelreich works with engineers on exactly this theme, of ethics, and he mentioned the work of Jason Millar as important in mainstreaming “the ethics discussion” among engineers. He agrees that the Trolley Problem was only ever introduced to engineers and the media as one kind of provocation to thinking about the outcomes of designing technology. It was never meant to become so popular. Millar, an engineer-turned-ethicist, clarifies that ethics is “something we are doing every day … we cannot localise it to computation alone.”[71] Millar himself is dedicated to working as an ethicist in an engineering department[72] translating between the two disciplines, and to take forward ethical applications in the design of technologies. As I have argued, ‘ethics’ has carried and continues to carry narrative and discursive power to shape the emergence of new fields. Concern with the unregulated role of Big Tech has brought attention to the role of engineers in shaping technologies as ethical: documentation of the evidence of algorithmic harms, particularly racial bias;[73] media stories about runaway drivelers cars likely to kill if not programmed correctly;[74] and the manipulation of social media data in election-related influence in the 2016 US Presidential election. In these cases there is a reflection of the top-down approach; that programming an algorithm to be ‘less racist’, or for the driverless car to even have a set of actions based on programmable values, is possible from the site and source of their origin. These cases are neither equivalent nor comparable but

[71] Interview with Jason Millar, June 2020. See appendix 2 for details.

[72] Millar’s official faculty profile https://uniweb.uottawa.ca/members/3760

[73] A striking example is the 2016 investigative piece by ProPublica, Machine Bias, about the application of an algorithmic risk assessment tool called COMPAS to predict recidivism rates in different incarcerated populations in the US. It was ‘found’ that Black people were assigned higher risk https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[74] See charts under the following section, ‘the media discursive influence of the Trolley Problem’

they serve as a moment to think about the discursive struggle over the shaping of technologies, who is doing it, and what this means for the advancement of these technologies in society. The recently emergent field of algorithmic bias mitigation, aka ‘FATML’ (Fairness, Accountability, and Transparency in Machine Learning) or FATE (Fairness, Accountability, Transparency, Ethics), is a shift to reframing ‘ethics’ as ‘bias’ in recognition of the wider culture of discrimination in society that permeates technical systems. There are glimmers of a machine ethics approach that in how computer scientists and computational means are centred in addressing this bias through technical means.

The assumption that the problem is at a putative ‘top’ of the food chain of computational design and development was not met with agreement by two interviewees, both experienced engineers; instead, they referred to professional culture as a site where approaches to ethics are made, and need to be re-made.

According to one interviewee, Alan Winfield, Tesla may have actually made very good cars but are testing them with a “flagrant” disregard for safety. Google are not an automotive company and are hence just “retrofitting auto-pilot.”[75] From the perspective of traditional car companies there is much to criticise in how software engineering is approaching the making of the autonomous vehicle. Winfield cites the differences in education and training of software engineers, and civil or mechanical engineers. He says that most people who develop AI are computer scientists and not engineers, and thus do not have to necessarily study the ‘boring stuff’ such as “reliable software standards documentation … or code reviews and how

[75] Interview with Alan Winfield, May 2020. See appendix 2 for details.

you set up a code review … you just learn how to code. And this is one of the reasons why there is poor quality AI in the world and AI-related disasters."[76]

"Software engineers do not have to worry about if their product kills people whereas mechanical engineers need to think about bridges falling and houses collapsing. So there is a very different way of thinking about a car", said the other interviewee, a German automotive engineer, Sven Beiker, who moved from Detroit, Michigan, the heart of the automobile industry, to Silicon Valley, the heart of the software industry. He goes on, "software companies have realised that building a car is really complicated, so they have teamed up with car companies … and that's why I have a job in Silicon Valley (laughs) " Here, he is also referring to the mergers and agreements between the research teams of automotive manufacturers and software companies that had taken place over the past few years.[77] It is possible that these differences in engineering cultures have much to do with how technologies are imagined and built. However, what is perhaps more relevant to this thesis is how a rather narrow moment of decision-making becomes the problem: 'how the car should behave in the case of an unexpected accident scenario'. What if the problem were framed differently? For instance, of what autonomous driving would mean for urban space, public infrastructure, or energy use. How would these realities change the conception of autonomous driving? With the development of this technology professionals have begun to ask such questions. Yet, the media-discursive influence of AV ethics in terms of the moment of the speculative crash generates a peculiar sensational impact.

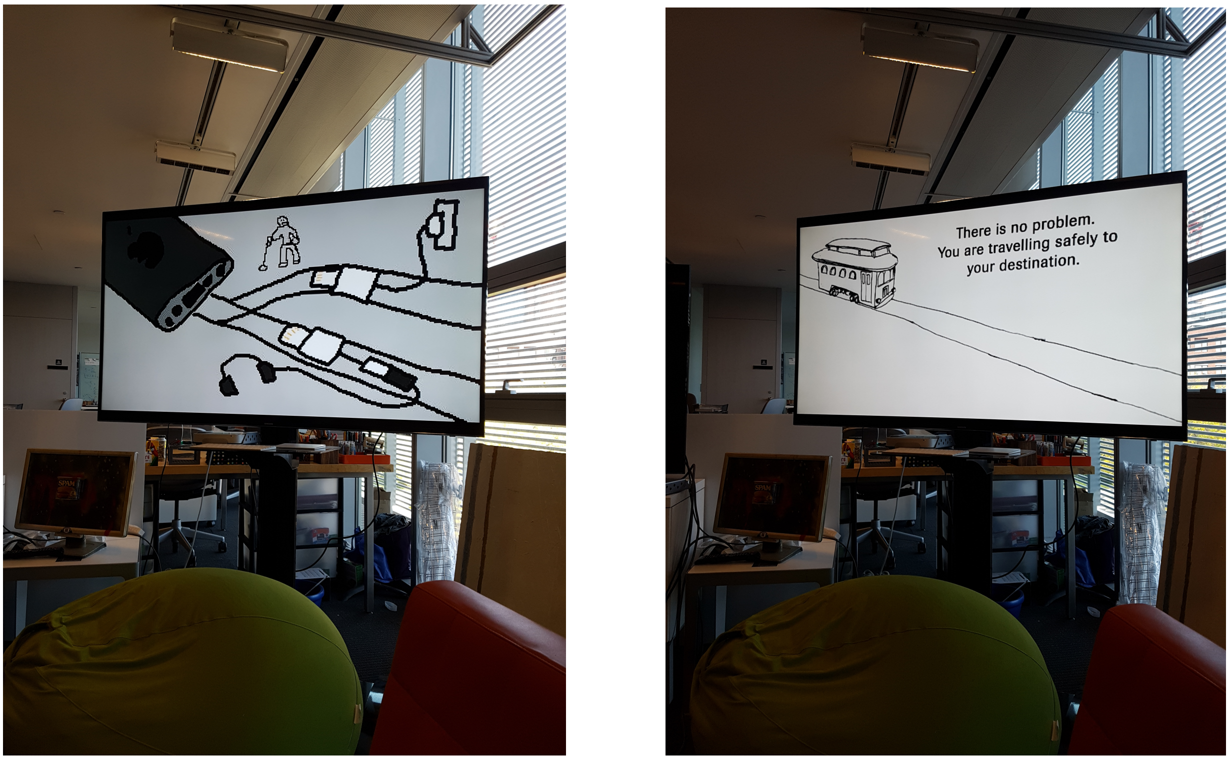
[76] Interview with Alan Winfield, May 2020. See appendix 2 for details.

[77] Interview with Sven Beiker, July 2018. See appendix 2 for details.

Criticisms of the Trolley Problem

Despite its media popularity, the Trolley Problem has attracted significant criticism (Cassani Davis, 2016; Bogost, 2018) as a significantly flawed approach to ethics (Himmelreich 2018); and that it is just one possible approach to the ethics of autonomous driving (Nyholm and Smids, 2016). Heather Roff's (2018) detailed critique makes an argument against the use of this thought experiment: that it models decision-making and morality without actually solving the problem; that it sets up discrete accountabilities, such as if it is the responsibility of the programmers, or of the machine learning system in the AV, or the human (passenger or driver) overseeing it, who are responsible for the outcome; and it does not adequately capture what other kinds of ethical concerns might exist in the development of this technology; and that it models only a particular instant or moment. Finally, Roff argues, autonomous car computation does not really function in the sort of binary way that the Trolley Problem suggests; she writes that the AV "is not making a decision at one point in time but is making sequential decisions. It is making a choice based on a set of probability distributions about what act will give it the highest reward function (or minimise the most cost) based upon prior knowledge, present observations and likely future states." I emphasise Roff's critique with my own arguments about requiring clarification in our conceptual approach to what the AV is; that it is ontologically multivalent, and hence existing and operating with diverse publics, knowledge systems, and histories. Furthermore, the approach of the socio-technical might conceptualise accountabilities differently, as not just located in the ethics education of engineers — or in machine learning. In addition, the matter of ethics has to be seen more expansively, in terms of the worlds it inhabits and shapes; for instance, to conditions of labour, and the price of accountability for those humans subjected to maintaining the infrastructures of autonomous driving. Next I turn to the Moral Machine, a large scale survey of AV ethics that appears similar to the Trolley Problem but in effect makes a significant shift away from its logics.

Moral Machine Project



There is no problem.
You are travelling safely to
your destination.

IMAGE 8: PHOTOGRAPHS TAKEN IN MIT MEDIA LAB'S SCALABLE COOPERATION GROUP OFFICE WHERE THE MORAL MACHINE PROJECT WAS DEVELOPED. CAMBRIDGE, MA, UNITED STATES OF AMERICA, OCTOBER 18, 2018. IMAGE: INDIRA GANESH.

Moral Machine[78] is an online ‘serious game’ by the (now-dissolved) Scalable Co-operation Group at the MIT Media Lab[79]. The Principal Investigator on the project, Iyad Rahwan,[80] has given many TED Talks and is a familiar figure in the public and media discussion about the ethics of autonomous driving. Moral Machine is a proposal for computing rule sets or architectures for ethics in the AV context, not unlike other bottom-up machine ethics approaches presented here, like ‘ethics bots’ (Etzioni and Etzioni, 2016), and the ‘general ethics dilemma analyzer’ (Anderson and Anderson, 2005). Widely profiled in the mainstream press and technology press and prestigious journals[81], the developers of Moral Machine say that it is:

> a platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars. We show you moral dilemmas, where a driverless car must choose the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you judge which outcome you think is more acceptable, you can then see how your responses compare with those of other people.

[78] https://www.media.mit.edu/projects/moral-machine/overview/ and https://www.media.mit.edu/projects/moral-machine/overview/

[79] The Scalable Co-operation Group at the MIT Media Lab was active from July 2015- July 2020: https://www.media.mit.edu/groups/scalable-cooperation/overview/ This group was led by Iyad Rahwan with a number of students and scholars who developed the Moral Machine project. They also developed other projects in a similar vein such as ‘MyGoodness’, a project to determine a calculation for how to donate to charities and thereby assess the ‘goodness’ of your altruism. Another project, DeepMoji, trains machine learning to understand the emotional nuances that humans convey through the use of emoji.

[80] Iyad Rahwan has a specialisation in Information Systems, and is director of the Center for Humans and Machines at the Max Planck Institute for Human Development in Berlin. Untill June 2020, he was an associate professor of Media Arts & Sciences at the Massachusetts Institute of Technology (MIT) in Cambridge, Massachusetts. In that role he was also director of the Scalable Cooperation Group at the MIT Media Lab; this Group is best known for their project, Moral Machine, that crowdsources a database of responses to how a driverless car should make a decision in the case of a Trolley Problem-like situation. I analyse this project in this work.

[81] It has been featured in 34 news outlets and scientific research publications including *Nature*: https://www.moralmachine.net/

Moral Machine is an online game in ten languages. Visitors to the website will encounter thirteen scenarios, each one featuring an autonomous vehicle with failed brakes approaching a pedestrian crossing where different human and nonhuman protagonists are crossing: cats, dogs, thieves, doctors, old people, men, women, pregnant persons, parents. In some scenarios, the AV is empty; in others it has adults and children, and in some cases just children. In some scenarios the player has to decide between saving the children in the car rather than the children crossing at a 'do not walk' sign. In another, the player has to consider choosing to save a child in the car over a pregnant woman; of men over women; dogs over cats; scientists in lab coats over thieves running away with bags of money; and, younger people over older people. In another, the site visitor might have to consider the value of the law itself: if there are children crossing at a crosswalk when the sign is red, and thus are in violation of the law[82], are their lives more or less valuable than those of the children in the driverless car?

[82] It is not expressly stated as such, but we can assume that these scenarios are situated in the United States.

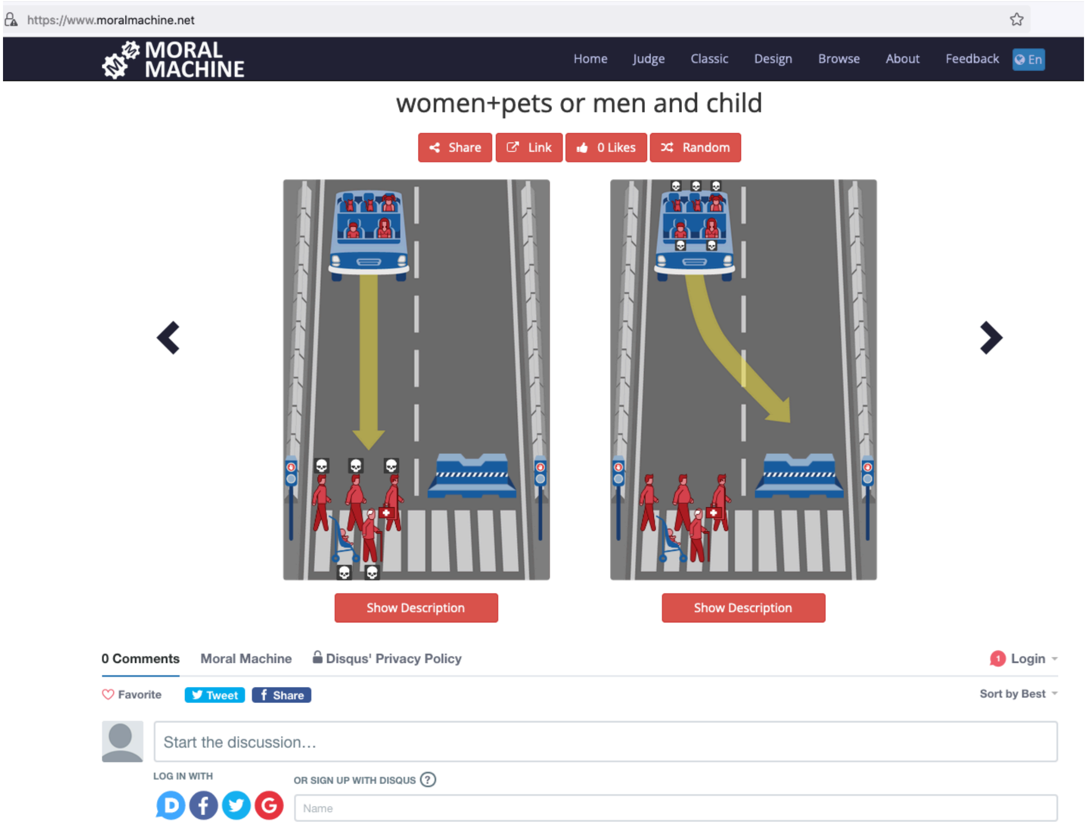

women+pets or men and child

IMAGE 9 SCREENGRAB FROM MORAL MACHINE PROJECT WEBSITE: MORALMACHINE.NET. DATE OF SCREENGRAB: JUNE 1, 2021.

The scenarios embody seven kinds of preferences in how the driverless car could act: prioritising humans over pets; passengers in the car over pedestrians; one gender over another; the young over the old; pedestrians following the rules over those who are breaking them; the fit over the less fit; and between people of different social status, like doctors over thieves.

Before this project, its developers had already conducted online surveys with close to 2000 Amazon Mechanical Turkers to a Trolley Problem style situation. The participants had to choose between saving one person or many people in the path of a driverless car with failed brakes . The authors report on the results in *Science,* showing a significant skew towards saving more lives than fewer, a Utilitarian approach. They underscore the importance of such experimental ethics tests saying, “a collective discussion about moral algorithms will need to encompass the concepts of expected risk, expected value, and blame assignment.” (Bonnefon et al., 2016, p. 1576)

In two years Moral Machine assembled 39.6 million responses from three million unique online visitors from 233 countries. The results were relatively unsurprising to the fictional quandary. They privileged the lives of the young over the old, that of humans over animals, and saving more lives over fewer. Country-specific and cultural variations in preferences emerged, such as valuing the old over the young(or the reverse) or a preference for those pedestrians not abiding by the law. The authors read these variations in terms of socio-economic and local cultural factors; that in many low-income countries, rules of the road may be more lax and that there are no strict laws about jaywalking enforced, for example (Awad et al., 2018). The authors of Moral Machine also acknowledge the limits of the responses gathered through the game: That they eventually only reflect the values of a self-selected group of people with internet access, interest, and time for a niche pursuit as a game about AV ethics. However, the results do not necessarily match the socio-demographic makeup of their countries and hence cannot be considered representative (Bigman and Grey, 2020).

The purpose of Moral Machine is not to educate, as Jason Miller proposes to do by teaching ethics to engineers, but to crowdsource a ‘ground truth’ for AV ethics, say a group of scholars involved with the project (Noothigattu et al, 2017). Other developers on that team note:

"training Deep Learning models often requires human-labeled data numbering in the millions" (Kim et al 2018 n.p), a time-consuming and resource-intensive method. Humans, however, are able to make predictions and decisions from a smaller number of "noisy and sparse examples" (Op Cit). Thus, they wanted to build:

> a model to understand how the human mind perceives and resolves moral dilemmas on the road as an important step towards building an AV with human moral values… this model would…describe[s] how the human mind processes moral dilemmas and provide[s] an interpretable process for an AI agent to arrive at a decision in a moral dilemma.

The group behind the project writing in *Nature* also say this:

> Never in the history of humanity have we allowed a machine to autonomously decide who should live and who should die, in a fraction of a second, without real-time supervision. We are going to cross that bridge any time now, and it will not happen in a distant theatre of military operations; it will happen in that most mundane aspect of our lives, everyday transportation. *Before we allow our cars to make ethical decisions, we need to have a global conversation to express our preferences to the companies that will design moral algorithms, and to the policymakers that will regulate them.* (Awad et al, 2018, p. 63; emphasis added)

There are multiple different claims and assumptions being made here, some of which inspire-confusion. First, even though this might not serve as a representative sample, some of developers argue that Moral Machine's relevance is in how it generates data that materialises a *model of the world* whereas the Trolley Problem only proposes the outlines of a problem. Second, they explicitly want to have a car make ethical decisions, and see this project as potentially influencing the development of "moral algorithms" (in the block quote above). They also want to model how the human mind processes moral dilemmas and transfer that to the autonomous vehicle. This is perhaps the most ambitious and curious aspect of their project, but perhaps in keeping with a certain kind of dream for AI, and indeed for the driverless car as embodying a high degree of complex decision-making.

Through this work I have referred to the Trolley Problem and "iterations on it" or projects that have been "inspired" by it. However, the Moral Machine is distinctly different from the Trolley Problem and its developers take pains to distinguish themselves from the Trolley Problem:

> Right now, this conversation has relied heavily on trolley dilemmas, whose lack of realism has tempted many to discard the conversation as technically irrelevant. This reaction is both legitimate and mistaken. It is legitimate because trolley dilemmas do lack realism. It is mistaken because trolley dilemmas are merely the unrealistic discrete version of a very real dilemma that emerges at a statistical level. *This statistical dilemma needs to be solved, and engineers must have a voice in this process*. (Bonnefon et al, 2019, p. 504; emphasis mine)

Here, Bonnefon and his collaborators in the Scalable Cooperation Group are making some important statements. They say that the Trolley Problem is not realistic in its construction, i.e. that it might be too narrow in the way it imagines how accidents could actually take place. They also say autonomous driving is a matter of a large-scale statistical activity, which is pertinent if we consider the driverless car crashes discussed in the previous chapter that happened because of computer vision failures. They suggest that the reality of the actual track, trolley, or numbers of workers on it is not as important as how these are processed as statistical realities. Hence, they are captured by logics of models, datasets and their standardisation and benchmarking, computer vision, optimisation functions, and similar big data processing and analytics. If we are to take Bonnefon and his collaborators at their word, and take a speculative project like Moral Machine seriously, then we might argue that it sets up a framework to generate ontologies of future road traffic violations, victims, and perpetrators based on how statistical models might 'know' about the world, and in turn create the world through this knowing. Statistics based-machine learning improves though engagement in the world, and puts out decisions that change the world. In this sense, it gains agency and effects change. However, its knowing—of the world, of what a violation or victim or perpetrator is —relies entirely on the architectures, infrastructures and cultures influencing the design the systems. The ethical apparatus that is the driverless car is quite literally engaging in developing a quantifiable model of the world based on crowdsourced responses from around the world. As much as it is computational, 'autonomy' is also deeply discursive and epistemic. So, what falls between the cracks of the past as captured by a machine learning dataset and the risks it arrives at based on these? What happens in the spaces between? I turn to these questions in my concluding thoughts.

Conclusion

This study has focused on the epistemes, epistemic modalities, instruments, measurements — in other words, a heterogenous set of knowledge practices — that establish the discursive terms under which we come to know an artefact as 'autonomous' or 'ethical'. In this conclusion, I draw out reflections on the Trolley Problem as a machine-ethics type of provocation for ethics in AI/autonomous technologies, and the Moral Machine as a recent, different instance of the same.

While the question of ethics is important, the answers offered by the Trolley Problem and Moral Machine are dissatisfying. The hubris of the driverless car ethics discourse is that we are confronted with an incomputable problem that must be computed; otherwise, we are told: *lives are at stake*! And it is up to engineers to figure out how to address this problem. If they cannot, well perhaps they need an education in ethics.

There is the question of ethics being knowing what to do; and then knowing how to architect a computational system that will know what to do even if a human cannot. The human, here, is expedient but ultimately unnecessary, for there is a limit to what humans can actually reason out, given the scale of the task. This is a discourse that is less about solving the problem and more about generating research about Aligning machine decision-making capabilities with pre-existing human value systems. The problem is identifying *which* set of human values take precedence and are applicable when decision-making technologies are imagined on a large and on a global scale. When the social context in which this decision-making is being performed is as specific and narrow as a household heating system for example, it might be possible to adopt some of these approaches, as 'ethics bots' do by taking a bottom-up approach (Etzioni and Etzioni, 2016). Similarly, and more precisely in the context of autonomous driving, MobileEye's Responsibility Sensitive Safety (RSS) attempts to derive "ground truth" for programming decision-making by reverse-engineering 37 kinds of pre-existing accidents recorded by the US' National Highway and Transport Safety Authority seem clumsy and astonishing; while a single home might create a specific set of challenges for an ethics bot, the scale of an entire city as the territory for ethics is ambitious. But this is exactly what the ethics of autonomous driving proposes. So we have to ask of the RSS approach: what are the tradeoffs in arriving at this ground truth? What falls through the cracks of the 37 most common accident scenarios? Not only is this decision-making approach highly situated, but it also reduces ethics to a specifically computable outcome.

The Moral Machine project however, marks an important shift in this thinking from a value alignment approach. Without explicitly saying so, Moral Machine appears to run a big online social psychology experiment that crowdsources values in response to a series of fictitious provocations not unlike the original Trolley Problem's original structure but with more variety. The dataset of 36.9 million responses shifts the question or the challenge of machine ethics: rather than architecting rules for action, which we imagine ethics to be, it proposes a *stat-*

istical representation of the entities around and in the car, and what they might mean in terms of various risk scenarios. However, the website still uses the words ‘machine ethics’, ‘morality’, and ‘machine intelligence’. As critics like Heather Roff have pointed out; the Trolley Problem is only a speculation and a provocation, and not even entirely appropriate to the real concerns likely to be raised by actual driverless cars (Op Cit; Nordmann, 2007, Nordmann and Rip, 2009).

The Trolley Problem as a frame for ethical reflection is like the standoff between an unstoppable force and immoveable object; algorithmic logic based on unknowable and difficult questions in the absence of knowledge is the limit of reason, a kind of ‘madness’, and irrationality (Amoore 2020, pp. 119-121). Yet this is actually what algorithms want, but only in a manner of speaking because algorithms cannot want things. Contemporary algorithmic systems “positively embrace[s] and require[s] irregularities, chance encounters” (p. 121). Contemporary algorithmic systems, and social sciences such as psychology and economics, among others, find their provenance in sciences of information organisation and planning that emerged in the second half of the twentieth century, some finding their roots in ‘cold war rationalities’ (Erickson et al., 2013). ‘Cold War rationality’ refers to a broad intellectual agenda of research and policy that emerged in the mid-Twentieth century across US universities and think tanks. Coming just after the anxiety of the Cold War, following the horrors of World War 2, and the bombing of Nagasaki and Hiroshima, the need for rational, measured, and considered decision-making in the face of complex odds, it was seen as an intellectual response to the unstable “human factor”: failed politicians, emotions, desires, and nationalisms.

What is relevant to the present discussion is to expand on Amoore's contention that algorithms embody an irrationality that is actually entirely rational, in terms of what they are architected to do. Theories of rationality were expected to identify answers that were possible, probable, and even unthinkable, because they were stripped of human limits to knowing for any reason, be that personal context or just because, well, humans are limited. In our interview, Jason Millar is deeply critical of the Moral Machine project. He says it is "more behavioural economics than ethics … or maybe psychology. It did nothing to build on existing work by ethicists and it took the conversation back (to the Trolley Problem) rather than forward."

Algorithms are architected to explicitly return all possible results to a question, and not explicitly just the 'reasonable' ones. We may not like these results and they do not always attend to our existing human ways of knowing and reasoning even though they might be constituted by human data. Hence, even a biased result, is just the algorithm doing what it does. Further, the work of algorithms is to learn from error, and to identify the very limits of unknowing.

This is how algorithms then require more data, more attention. This is one of the reasons why algorithmic decision-making is proposed as an innovation in sensitive domains such a hiring: to minimise human bias. So, we see increasingly that algorithmic decision-making enters into domains that have for long relied on what we believe to be uniquely and intuitively human; in other words, that we think of as difficult to replicate, or that is a kind of tacit knowledge: automated CV-perusal and hiring; baby-sitter vetting (Harwell, 2018); and, secretarial assistance (Lingel and Crawford, 2020). We value mathematical and quantified knowing because it isassociated with an idealised notion of objectivity and instrumental reason, the narrow fitting of means to ends (Erickson et al., 2018p. 2). However, now we know that human bias permeates algorithmic bias because algorithmic systems are inherently social and based on past data and assumptions, and also convey human values.

The Trolley Problem does not demand an answer, nor is it not solvable; but in its very structure, and by being invoked in this context, it identifies something that a human cannot answer — but that computation *could*, thus elevating what is possible with ‘AI’. Computation can produce a cold-blooded answer in a manner that a human might not have the stomach for. Eventually, it is *just* a number, just one possible answer. This is one of the rationales for Lethal Autonomous Weapon Systems (LAWS): they are not affected by contemplation of what it means to kill or die.. Ultimately, this is the harm enacted by algorithmic systems. What lies to the human then is to limit this excess. Setting up a sorting exercise to compare options of who to kill off, or let live, as the Moral Machine does is hardly a responsible approach (Amoore, 2020, pp. 111-120) (.120).

The forks in the road facing the driverless car mirror the binary options of the Trolley Problem that ask which way a car should turn in saving either one life, or five lives. ‘Forks’ are also a term commonly used to refer to algorithmic branching, or in how code is built through collective repositories. ; It refers to the expansion, and searching for more valuable outcomes toward the reward hidden in a function that extracts the deepest insights from a data set. The scale and complexity of movements required of algorithms making sense of 'the world' to arrive at the correct answer — assuming there is one — is nothing short of an almost impossible task. In the previous chapter I discussed AV crashes within the Bratton-ian 'blur', an approach to the fractured contemporary computational landscape he refers to as *The Stack.* I showed how histories of human-machine relationships and embodiment in the automation of tasks, in driving, and computer vision coalesce to form the basis of a fractured space between varying degrees of automation. There are indexes, measures, and metrics that detail what autonomy is imagined as; and through socio-technical materialities of the auto-pilot setting, and computer vision. In the moment of the crash, the breakdown of these systems and settings, and the legacy understanding of the role of the human in relation to the machine, reveal the workings of the apparatus that shapes driverless-ness. If we, for a moment, put together the model world of the Trolley Problem and Moral Machine against the real, shaky, and distributed mechanisms of global-scale computer vision systems, we might see the enormity of the computational task required for a car/software to be considered ethical, autonomous, or intelligent. In short, a computer vision system would have to correctly identify whoever is in the environment of the driverless car, and risks they might face from being subjected to the negative outcome determined by an algorithm, in other words, from being run over. How extensive would an individual's data profile need to be for an algorithm to make a correct assessment of who they are, and what their 'worth' is, and to prevent being wrongly or poorly assessed by existing algorithmic systems? Or, worse, as discussed in the previous chapter, to avoid the risk of not being recognised at all? It is not just a matter of the *reduction* of a life to something computational, or a calculated decision; rather, it is the *expansion demanded* of the algorithm (Amoore, 2020), if it is to make an adequate assessment of a human as compared to another, to truly embody the world of even one human life. It is difficult to imagine that algorithms are capable of such expansion; however, it is exactly this threshold of incapability that an algorithm and its developers might take as a challenge. (And perhaps already have.)

CHAPTER 6. CECI N'EST PAS 'ETHICS'. A CONCLUSION

Multivalent hauntologies

What is the driverless car? Is it a car or a robot? Does it exist in the past or the future? Is the human driver being replaced, or is the vehicle just being driven differently now? Given his influence in studies of language and culture, Jacques Derrida's 'hauntological' is perhaps a fitting place to begin the end of a thesis about the making of representational knowledge. Hauntology is mentioned only three times by Derrida in the text, *Spectres of Marx* (1993/2006), and yet has become a rich provocation for scholars for more than forty years. Hauntology suggests a return of the past as an apparition in the present/future. But hauntology refers to more than just this temporal drag, or 'time out of joint' as he puts it. Hauntology underlies Derrida's *différance*; the rules of language might be relatively stable, and things and concepts are known by an arbitrary assignment of words whose meanings are not fixed or stable, how they are used in speech and human communication shifts such that meanings become removed from where they started, creating an ever-widening gap between what something *is*, and the accumulations of signifiers and representations we use to refer to it. All language, he suggests, is an arbitrary set of relations in which the actual meaning of things is increasingly reliant on that which it is separate from, on that very deferral. In starting with words and language, my intention is to reiterate that this work has sought to identify the infrastructures of knowledge-making, and the making of representations through practices of distinction, amounting to discourse: A legitimate terrain of associations that enable us to speak about a thing. My intention has been to identify aporia emerging as representations gain popularity and take hold as established discursive material. Hauntology is also the acknowledgement of trace elements of one ontology in another. Multivalent ontologies of the AV suggests that each avatar carries the future possibilities and past traces of another. The robot is not here *yet* but it is, in a sense, already embodied in fictions and Hollywood. There are industries, bureaucrats, policymakers, technologists, and ethicists waiting expectantly for the driverless car to 'arrive'; some wait expectantly, others nervously. Multivalency means we are never only dealing with *just* computer vision *or* smart city visions; we are dealing with both at the same time, in deeply entangled systems that exist both in the legacy of automobility, long-held dreams of engineers, and in a speculative future. There is a haunting in the magical leap from Level 3 and 4 on the SAE scale of autonomous driving, to Level 5 of 'full automation'. The haunting is the uncertainty of what this final form will be if it ever emerges. The machine learning models that constitute various AI technologies operate on the basis of statistical correlations made on past data sets; thus, the future is contingent on inscrutable connections made between past events. Machine learning works by making distinctions and categorisations, and thus has effects in the world by identifying things as cats or dogs, rightly or wrongly. We live haunted by not knowing what will happen should our old cat or dog wander out in the path of a failed driverless car.[83] Because, ultimately, even algorithms make mistakes. What remains haunted is the possibility that we might arrive at 'ethics' in the spaces between amazement and horror; this is the haunting of human and machine, a spectrality evoked in the Trolley Problem. The spectre, I believe, is not just the anxiety evoked by AI's 'almost-human'[84] status, nor the terrible choice that an autonomous vehicle might have to make. The spectre that haunts AI and the autonomous vehicle (AV) is actually the human, any human, and the dream of an efficient human that knows what ought to be done, and then does it. The spectre is bad data, not enough data, or poorly labelled data; it is humans not being more empathetic, nor changing their behaviour to suit the needs of the car. We are not very good drivers, and start to bend the rules as soon as we can. It is impossible to predict what we might do in a Trolley Problem-type situation, or any other unexpected situation, but that impossibility is the challenge that the algorithmic system takes on. It promises a definite answer, even if we do not like it. Consider how the metaphor of the 'black box' has become a site for social research about AI, chiefly as something to be opened up and made transparent. In this view, AI and autonomous systems are industrial technologies that can and must be regulated through transparency. However, a 'black box' is also obfuscatory and threatening. The 'black box' emerges from Skinnerian Behaviourist psychology that perceived the human mind as the ultimate black box, one that could only be scientifically manipulated, which was the only proof of its existence. The more perfect human is the one making rational choices,

[83] In 2016, North Wales police faced the ire of their community for deciding to run over a dog that had run out onto the motorway to save other motorists: https://www.walesonline.co.uk/news/wales-news/police-defended-decision-deliberately-run-10940848

[84] Donna Haraway observes, "children, artificial intelligence (AI) computer programs, and nonhuman primates all here embody 'almost minds' (1989, p. 376)

and one that strives to improve herself; mentioned earlier, this is like the scholar Martha Poon's assessment of the driverless car as the perfect neoliberal driverless car, tootling along making decisions for itself. On the other hand, AI is always improving with better data, better infrastructure, and more opportunities to connect. So, with AI and AVs that get better, it seems the place of the human is to create conditions of labour to make this improvement possible. These multiple hauntings have been a challenge to bring together in this study, but not just because they are multiple.The challenge in this work has been to identify the work these hauntings *do*.

I set out to understand how and why ethics and ethical decision-making were being associated with autonomy and autonomous driving and found that a field had developed around the concern that AI technologies in future AVs might make decisions resulting in harm to humans. Autonomous driving is seen as a corrective to the problem of poor human driving that is error prone, and slow to follow rules. Computational cars, it is said, will follow the rules. However, a concern began to circulate that the AI 'inside' the driverless car might face an extreme traffic situation in which it would have to make a decision that would bring harm to human lives. This was captured by the provocative thought experiment, the Trolley Problem. This concern was being discussed in journals, round tables, podcasts, op-eds, and internet memes, and potentially only amplifying a speculative new technology rather than actually being concerned with ethics. Perhaps this has been about industry finding ways to evade regulation by demonstrating a concern for 'ethics'. Proposals for ethical machines that can be scripted to deliver decisions that align with human values have been developed. But the question becomes: *which* values, and *how* exactly will they be scripted in? This maps loosely onto an approach to AI and machine 'intelligence' in which a domain of human knowledge is captured as rules set. But the Moral Machine project makes a shift towards large-scale statistical calculations based on large and situated data sets in which decisions are made in response to models of the world around the car .. But by proposing algorithmic moral decision-making based on data models, there is always the risk that models do not adequately capture the worlds they are applied to; they are architected in terms of past worlds, and must respond efficiently to new situations that may or may not correspond to these past worlds.

However. identifying rules to architect ethical decision-making has not identified matters of actual harm and concern. Computer vision systems in existing driverless cars-in-testing have failed to adequately recognise objects around the car resulting in four fatal crashes; The crashes happened because control of the car in auto-pilot, or 'self-driving' mode at the time of computational failure was handed back to the human driver who was not paying attention. Hence, gaps open up between statistical models and the worlds they are applied to Thus autonomous driving relies on embodied and situated human knowledge and labour which are obscured. In the rest of this chapter I move between summary and synthesis, identifying how representational knowledge is being materially and socially made about what autonomous driving is, and how ethical decision-making constitutes it; and most importantly, what the implications are for the transformation of the world of human social relations, spaces, and bodies through the emergence of technologies that enable autonomous driving . I conclude with an appraisal of what kind of future we want to build with AI and autonomous technologies, and the re-conceptualisations of ethics and epistemologies we will need for this.

The material-discursive shaping of representational knowledge about ethics and autonomy

There is a mixture of concern and enthusiasm for autonomous driving from industry-aligned, industry-based, and academic researchers, scientists, ethicists, economists, and engineers; breathless and bombastic promotion by tech entrepreneurs like Elon Musk; public crashes and public testing and rollout in various cities[85];and sizeable investment of over US$80 billion as of 2017. Yet, there is no 'fully' autonomous vehicle or fully driverless car. 2020 was supposed to be its date of delivery, but there is still nothing that has emerged. Elon Musk's Tesla is under review and verification by the California Department of Motor Vehicles (DMV) for classifying their cars as "fully self-driving" (Mitchell, 2021). The UK proposes to trial autonomous driving technology in 2021 (Criddle, 2021). The ride-sharing companies, Lyft and Uber, have recently sold their self-driving technology development portfolios (Rana, 2021).

[85] Google's Waymo self-driving car has been rolled out as a ride-sharing service for the citizens of Phoenix, Arizona, in the United States.

Rene Magritte's famous painting from 1929, *La Trahison des Images (The Treachery of Images)* bears the following sentence below the image of what appears to be a pipe: 'Ceci n'est pas une Pipe' ('this is not a pipe'). When Magritte says 'this is not a pipe' he is saying, "*this* is what we call a *pipe* because we decided to use a system of language to make meaning of this object." And what exactly is *not* a pipe, he asks, the canvas, the whole painting, or the word 'this'? The entire system of signs and referents is called into question here, the cages of language made to denote things as *x* or *y*. *Not* making clear how signs come to stand for signifiers and how representations are made is representationalism (Barad, 2007, p. 360). But to challenge representationalism is more than just exposing linguistics or semantics. The shift from linguistic representations to discursive ones establishes the move from how things are represented to how they come to intervene in the world. So I have examined material, social, and cultural sites and practices of knowledge-making that I argue validate and legitimise computation plus automobility as 'fully self-driving'. These have included scales, measures, heuristics, academic research, experts being profiled in the media; as well as moments like car crashes where representationalism falls apart—literally—on its own exposing the human element in AV autonomy. I develop the *ethical apparatus* to identify what kinds of knowledge, formal or informal, scientific or cultural, are being enacted to give us the language, representations, and discursive terms with which to describe 'driverless-ness', or the 'ethics of autonomous' driving. The ethical apparatus also identifies the institutional, individual, social practices and cultural imaginaries that are shaping autonomous driving itself.

The apparatus is a complex ‘heterogenous ensemble’ as Michel Foucault defines it, of institutions, material practices, imaginaries, people, science, philosophy, technical products, industries, infrastructures, and knowledge that establishes connections and relations between things, giving discourse its form. It is a formation that responds to an “urgent need” says Foucault (1980, pp. 194-195), and in the case of autonomous driving it might be the urgency of the statistic that 94% of traffic accidents are the result of human error. A statistic that scholars remind us are also just products of situated, mostly un-contested epistemological practices of accident accountability (Braun and Rendell, 2020). The ethical apparatus invokes two meanings of the word apparatus: As a device, and as discursive power. Sometimes the device is a literal measure, or it might be an epistemic device, something that produces and leverages knowledge that brings legitimacy or validity to a concept.

Material-discursivity is about the material, cultural, and social practices enacted in a variety of contexts, in which knowledge and representations are made, creating legitimacy, validity, and the possibility to ‘intervene’ in the world. In identifying material-discursivity, I am interested in the conditions under which terms like autonomy and ethics are created and how they come to intervene and have power in the world. I investigate how these terms are created through material sites, locations and practices: statistical modelling; computer vision; socio-technical approaches to the interactions of human and machine in settings such as auto-pilot; the Trolley Problem, Moral Machine, and machine ethics; and expert power.

Taking the case of Elon Musk and Tesla, the claims being made about Tesla's full self-driving capability let me summarise how I understand representational power to emerge. It relates to the copy on the Auto-pilot website[86] that dazzles a lay person with deeply technical language, and all the infrastructure to enable that computation. It is the venture capital and wealth to bankroll the technical development of driverless cars and send rockets up into the air[87]. It is the SAE J3016 standard of autonomous driving that establishes levels of hand-off of driving from human to car; the highest level is full automation, or full self-driving. This becomes the basis for a negotiation with the California DMV to review the actual autonomy of Teslas. (How will the DMV verify that the Teslas are fully self-driving?). It is the engineer's fantasy of fully automatic safety, a dance between automobile and road. It is the belief in the power of computation and automation over the human. In short, it is the creation of an entire reality and the epistemic infrastructures to support it. The apparatus makes and sustains representations without revealing how they are made, and bringing power and organisation to discourse; this is representationalism. But to question the apparatus of autonomy and narratives of ethics entangled with it, I follow Suchman and Weber's (2016) insistence on identifying the material conditions of the emergence of these representations. Addressing representationalism means looking at opacities that are sometimes referred to as progress, innovation, efficiency, or safety. But to look at, and through, an apparatus is always challenged by the fact of also being inevitably in and captured by discourse ourselves.

[86] https://www.tesla.com/autopilotAI

[87] SpaceX is Musk's company https://www.spacex.com/

Another analytical framework that has allowed me to open up and look through and at the shaping of autonomy and the ethics of autonomous driving is the conceptualisation of the driverless car as *multivalent cultural ontologies.* I argue that the AV exists simultaneously as at least three separate forms of the machine: As the AI/robot imaginary, as a conventional automobile, and as a distributed big data infrastructural system. Recognising these co-existing, shape-shifting cultural ontologies expands the numbers of nodes from which we might examine the artefact, and also to understand that its discursive shaping is taking place across all these forms. Each form brings with it entirely different modes of knowing, histories, and materialities, which are not necessarily separate from each other either. For the automobile has been imagined as a robot or flying car since 1939, and perhaps developed as such towards this very moment. Further, what we understand as automation, or AI, or driving, or even a car, implies that politics emerges from these multiple ontologies, influencing how “problems are framed, bodies are shaped, and lives are pushed and pulled into one shape or another” (Mol, 2002, p. viii). In other words, each cultural-ontological form is instantiated by different realities that introduce a more complicated view of the publics, stakeholders, social, and political nodes that constitute this apparatus, and hence what autonomy means.

The Trolley Problem: A closed loop

Coupled with prevailing notions from science fiction about the eventual dominance of AI, and of the threat of robot technologies going rogue, the Trolley Problem generates anxiety about autonomous driving, and emphasises the need for regulation. Many have roundly criticised this framing even as its popularity has grown saying that this approach is not just unlikely but that it does not capture the reality of how accidents actually happen. In examining the representational knowledge being created about ethics and autonomy, I also argue that the device, in this case, a proposal for autonomous driving ethics, the concern is not just *computing* ethics—any number of decisions deemed ethical are being made for us constantly by machines and machines learning. The whole system is one of decisions that embody values of varying degrees and importance. But it is also that what we understand as the ethical itself is contained within the modes of knowledge-production about the world, that shape the world, through decisions enacted within computation. I argue that our conception of ethics and ethical decision-making has to spread beyond these specific moments and become distributed to different theatres of human interaction, including those possibly beyond computation.

Dr. Patrick Lin, a scholar I interviewed as part of this work, who has close ties with the auto industry, academia, transport researchers, and governments, admits that he has had considerable influence in amplifying the thought experiment. Even if it is highly unlikely to happen, Lin says, what matters is that sets up a challenge for AI and autonomous technologies. Just as science 'makes up' artificial situations to generate knowledge (meaning, through the artificial set up of experiments),here was another artificial situation to guide the development of this technology. But I believe this is convenient because it has doubled back into creating a concern where none existed; an absence of ethics has been identified as the rationale for ethics. Science and Technology scholars have identified a similar pattern in the fields of nanotechnology, and in genetic engineering, that a concern over ethics has eventually secured space for industry to proceed with its work with as little outside interference as possible. In the case of nanotechnology,‘speculative ethics’ turned ia “merely possible” future into something inevitable. And the case of the Asilomar Principles of 1975 in genetic engineering assigned ethical oversight of research to geneticists themselves. This is repeating in the case of autonomous driving ethics,, creating unique concerns where none should have existed. Eventually industries secure their own future research and development plans. While ethicists have to no doubt look ahead to the future and use dilemmas to unpack developments in science and technology, there is a risk of addressing hypothetical rather than actual developments, such as the conditions under which crashes have occurred. If anything, this only points to critical vigilance in the development of new technologies and the shaping of domains of ‘ethics’

A more important matter relates to another closed loop; Machine ethics imagines computational systems as becoming their own regulators. This necessitates second order systems to take on further monitoring, setting off loops of computational systems regulating each other. This is the dream/prediction of Cybernetics. This is what authors of the household heating and cooling ‘AI-assisted ethics bots’ propose and that this could be scaled up to driverless cars (Etzioni and Etzioni, 2016). But as I have argued, the reality of planetary scale computational systems that emerging AVs inhabit, the spaces of blur and fracture introduce complexities that demand much more, I believe, than what computation can achieve at present.

Re-inscribing human-ness

The Trolley Problem provokes a fundamental question: Why would we want to fashion the development of a new technology through the making of unthinkable choices such as e between saving the lives of one person or five people, or between sacrificing one fat man to save many thin people? In the shaping of autonomous driving, ethical decision-making has become a matter of what Gregoire Chamayou refers to as necroethics:

> While ethics is classically defined as a doctrine of living well and dying well, necroethics takes the form of a *killing* well. Necroethics holds forth on the procedures of homicide and turns them into the object of a complacent moral evaluation. (2013/2015, p. 146; emphasis in original)

Drone Theory is a powerful book about the moral and socio-political implications of the increasing use of automatised robot warfare, a kind of lethal autonomous weapons system (LAWS) towards theorising automation, autonomy, and ethics The necroethics of 'killing well' is the displacement of human morality and ethical decision-making onto the distant killing enacted by the material body of the 'subject-less' weaponised drone. Chamayou argues that once the technology (in this case, the drone) has been accepted as being a more precise targeting (killing) technology, the discussion of principles and ethics becomes diluted. The discussion then becomes one of numbers (i.e., statistics of success, cost benefit analyses), and their transparency, and accuracy, rather than the actual moral challenges posed by drone warfare (ibid)

We deploy robots to do care and cleaning work, and LAWS like robot soldiers to do our killing work. A LAWS is without moral subjectivity and only needs to be programmed to act;

it does not think or feel. We feel better about outsourcing our dirty work to a robot, and our dirty feelings too. f 'Cold war rationality' that inspired AI identified algorithmic rules to model, manage, and optimise for managing the worst outcomes of the Cold War, nuclear disaster, and other horrors of the twentieth century both real and imagined. It was a way to 'think the unthinkable' (Halpern, 2014, pp. 150-151). Now we have the precise and safer robot car which, in the case of an unexpected accident, offers to think unthinkable things for us, such as the uncomfortabledecision at the heart of the Trolley Problem. Imagine that in a hypothetical Moral Machine scenario you believed that the pregnant woman should be sacrificed to save the menopausal scientist who has just discovered the vaccine for a terribly contagious new airborne virus. You might be uncomfortable saying this in a group of people; you might feel uncomfortable making this choice even if you thought it. A robot program would not feel the same awkwardness; in fact, the work of the algorithmic system is to deliver every single possible outcome of a command without judging it. The driverless car as an AI/robot embodying an ethical decision-making program makes decisions for us, but *also* takes on the emotional, psychic, and moral labour of making difficult decisions. What for Foucault was ethics, that is, caring for the self by knowing the self, the challenging work of turning to gaze upon oneself (1994), is now what we might be relieved of in this formulation. Thus the Trolley Problem in the driverless car context is a test of humanity as it is of a machine. We have to seriously ask why we would use this as the starting point for the development of a new technology, even if it *just* a provocation. Being human is precisely about having uncomfortable and terrible feelings, but the Trolley Problem and Moral Machine propose that this might be taken on by software. Thus, I find that there is *re-inscription* of what it means to be human in terms of a “robotic imaginary”, or, “the shifting inscriptions of humanness and dehumanizing erasures evoked by robots.” (Rhee, 2018, p. 5) It is not that autonomous driving erases humans, or that the Trolley Problem is unsolvable, but that it brings serious and much-needed attention to the terms on which we construct our relationship to ourselves and our robotic others. The Trolley Problem is its own accident of morality, suggesting an exercise of applied ethics that empties humanity of itself.

‘Thinking the unthinkable’ is also a difficult part of accident scenario planing. Algorithms are useful in returning extreme and unthinkable results because that is precisely what they are intended for. In his thoughtful ethnography of the 2011 Fukushima Daiichi nuclear reactor explosion, Michael Fisch discusses a distinction arising in the accounts given by officials about why the incident occurred. He finds that the Japanese concept of ‘soteigai’ was invoked by the Tokyo Electric Power Company (TEPCO) as the cause of the disaster. “Soteigai translates loosely as referring to something that is beyond expectations. Accordingly, it is commonly understood to denote something that cannot be anticipated via existing risk management models and technologies” (2013, p.1). But Fisch pushes past this explanation showing that negligence and corporate mismanagement aside, eventually it was the closed nature of the system that led to the failure. Unlike an organic system that can evolve, a nuclear technology is a closed and tightly coupled system, this coupling produced by the nature of the rise itself. Anything irregular—in this case, a tsunami that exceeded existing data about the scale and effects of tsunamis—cannot be accounted for within the system. He concludes that soteigai was never about limits in thinking about the possible causes and contingencies of failure of the system, rather, it refers to the "*impossibility of thinking* the consequences of the nuclear crisis" (Op Cit, p 6, emphasis in original). In other words, the Fukushima Power Plant just could not allow themselves to think that a tsunami could be worse than anything they had ever known through past data. "It took some time for the reflexive realization that such an event could not have been predicted from past data alone." (Uncertain Commons Collective, 2013 ch. 2) Without venturing too much in the unique nature of nuclear technologies as complex systems, accident accountability is shaped by what we identify as the limits imposed by a technical system, and in our own thinking about causes and consequences of breakdowns and errors. In the case of Fukushima, it might have been the starting point for thinking about managing infrastructural failure by literally imagining the worst possible outcome.

Following this line of thinking about outsourcing difficult decision-making, eventually the computer will present options and make decisions. If it has to carry on computing, it has to take the next step by making a decision. It might push the button that no one is willing to; and hence, there are well-founded concerns about the automation of decision-making in a variety of circumstances. Thus, inspired by Bratton, Amoore[88], and Sprenger, we might look at decision-making as *just* decisions. Decisions as outputs are decisions by computers for computers, or by automated systems for and within that system (Sprenger, 2015, p. 21). What Sprenger means is that these decisions are not necessarily good or bad as far as computational systems go. From Bratton's perspective, decision-making is the lingua franca of planetary-scale governance; everything just seems to be computation perpetuating itself. As we know from the existing application of algorithmic technologies, harms are already taking place be-

[88] As discussed at the end of the previous chapter

cause decisions are emerging from systems that *don't feel anything* about the decisions they make. Decisions within large-scale automated systems like automated driving or automated warfare, similarly, are “mad because they can never know fully the consequences and effects of their own making” (Amoore, 2020, p. 112). Academic movements like FATML/FATE are attempting to introduce friction in the moment of madness to interrupt this ‘decisions-will-be-decisions' mode. They are asking about the consequences of architecting a computational or mathematical fix to an algorithm that is found to be biased or discriminatory precisely because decision-making does not feel things or know things. Instead, they say, the location of decision-making must be distributed elsewhere, where humans meet organisations, machines, and nature, where decisions might be more palpably struggled with, and struggled over.

Model machine

The Moral Machine project advances the ethics of autonomous driving discourse by claiming a step towards a “universal machine ethics”. The values the project identifies in their global data set of 39.6 million responses are not surprising: Save more lives, save younger lives, and save human lives; there are global variations and differences in culture and region that the researchers identify. The data set reveals a direction for what might be considered to be good and valuable as the basis for future algorithmic decision-making. However, this decision-making is not likely to be a set of rules in the strict machine ethics tradition, but proposes a shift towards statistical modelling and assessment of risk, and its management and distribution. Moral Machine’s goal is to establish ground truth; this is essential in building any computational model; a model is central to the activation of other computational agents like algorithms, optimisations, or rules for machine ethics. So, it has to arise from some where and be anchored in some reality. On the basis of this ‘ground truth’, a model can convey all the relationships in that particular world, and, deploying further statistical techniques, it can make presumptions about the future, identify moments, formations, and situations that are considered of high risk within that universe of ‘truth’. Of course, how risk is defined is itself a complex and situated practice. There is a general faith in statistical and data-based models, they have a strange power as truth; conclusions drawn from a model are believed to be infallible (Breiman, 2001, p. 202). But, all models are, also, *just* models, to paraphrase George Box’s famous dictum about all models being wrong, but that many are illuminating and useful nonetheless (1976). All models make assumptions about the worlds they are applied to, which is where a problem arises, because models are often applied to worlds that they did not originate from. Interrogating a model means we have to ask questions about the entire system underlying it, its provenance, or ‘back story’, as it were. How the models were constructed, what data was used and where they came from, and what the model was optimised for, are all questions that influence how the AV will know what kinds of make decisions to make. These variables will reveal the margins that exist between the model and the world it is being applied to, and hence between the world that *is* and is always changing, and the world as understood by a model.

Interrogating a model also means knowing how it is being applied and to what ends. Optimisation is one of those techniques that re-shapes models to work in different contexts. Optimisation is the technique that is integral to logics of smartness, proposing that everything—

every kind of relationship among human beings, their technologies, and the environments in which they live—can and should be algorithmically managed (Halpern et al., 2017 p. 119). However, it is very hard to know what a computer vision system is optimising for unless its behaviour patterns are being scrutinised carefully, or when errors occur. We know that some kinds of objects are difficult for computer vision models to identify: Like cyclists, because they are never static and do not have a defined shape like a human *or* a bicycle, but appear as an amorphous blob to a computational visual system (Fairley, 2017); darker phenotypes (Buolamwini and Gebru, 2018; Wilson et al, 2019); and, very light-coloured, large objects that are hard to distinguish from the sky. Every back story has a future. It is unfortunate when edge cases, or new cases crop up that a model cannot parse, because it is impossible to then process them at the scale and speed of fast-moving, automated, high-end engineering. It is difficult to encode the scale and pace of an unstable world into a computational system that is expected to make fine-grained decisions at a moment's notice. This is why Tesla operates with fleet learning, that is, data models about the world picked up across its entire fleet, distributed, and shared across all the vehicles.

'Model knowledge' acting in the world might be understood in terms ofwhat Antoinette Rouvroy, the Belgian legal scholar calls,'algorithmic governmentality'. This is a Foucauldian-inspired socio-technical concept that refers to the process of translating and transcribing "the physical world, its inhabitants, their trajectories, behaviours, actions, choices, preferences… into computation in order to direct the conduct of individuals and groups, to structure the possible field of action of others." (2011, p. 121). Algorithmic governmentality relies on algorithmic knowing about the world that predicts potential risks and danger in a context, and then uses that knowledge to governhuman societies, such as urban spaces. The Moral Machine and the ethics of autonomous driving presages algorithmic governmentality. This approachproduces a world legible to computational, algorithmic, machine-learning models that constitute autonomous driving/navigation, and set up a regulatory system for and of that world. Instances of algorithmic governmentalityare already in play and re-shaping the world. If something in the environment around the car is a pedestrian, a road divider, or the sky, how it is recognised by a machine, or a human, or a human overseeing a machine, is a statistical relationship between data points that make up the world around the autonomous vehicle. What that object actually *is* does not matter to the model, only what it is *perceived as* matters. This is where we have to consider how to interrogate the architecture of such models and examine their provenance. Algorithmic regulation is already a serious matter of how mis-recognition can change history itself. Facebook's partially automated flagged a famous black and white photograph by Nick Ut from Vietnam during the US invasion as 'child nudity' and hence it was considered in violation of the platform's norms on pornography. The image is of a young girl, Kim Phuc, burnt by a napalm attack on her village and running away naked. (Hansen, 2016).

Algorithmic governmentality is not just about automated decision-making but about the very transformation of reality itself. In that sense, statistical decision-making in sensitive social and political contexts are a cultural technique in the sense of *Kulturtechniken* discussed earlier. Moral Machine invites us to flip through various scenarios and decide upon the least worst way to send a fictional character to their death, or at least to manage the situation so as to minimise casualties (which does not mean *no* casualties). Who or what is more or less valuable, if the car must turn left or right, and eventually what is ethical or not ethical are instances of making of distinctions that construct particular realities. The implications of having the power to calculate what is otherwise incalculable is that the model, statistical risk calculations, or the driverless car, risks becoming imbued with the agency and power to make these kinds of decisions (Mbembe, 2019). The algorithmic sovereign becomes a decision-making apparatus of epistemic practices and actors, distributed across different spatialities and temporalities from past into the future.

Time, space, and bodies of knowledge

Hubert Dreyfus' theoretical contributions to early Artificial Intelligence include phenomenology and an effort to bring studies of embodiment and situated-ness, experience and context, to AI. Humans as 'expert systems' know the world through experience, embodiment, and in entirely local and unique ways; some of these ways of knowing can and are formalised into computational systems but most of them cannot be. Because the expert actually does not know how they know things, like how they solve particular problems or how they negotiate a world of complexity. Hence, creating rules for achieving such tasks in a comprehensive manner such that they can be executed is extremely challenging; aside from the computational resources required for a truly detailed expert system. Like the experienced human driver who builds on layers of experience to drive through a snowstorm despite not being able to see the lane markings; but translating this situated and embodied knowing into a computational formalisation is extremely difficult. A novice needs rules and has to follow them carefully. The autonomous vehicle trying to identify objects in its environment relies on human micro-work applying 'situated knowledge'. The human has to *literally see* on behalf of the machine. As discussed, human operators need to be vigilant, embodied, and alert, to take over from the car in auto-pilot that does not have the resources to make the right decision. In the emergence of autonomous driving, humans are firmly embedded in its big data infrastructures as micro-workers and as managers and overseers; their work and attention help the vehicle appear autonomous. Humans in this "cognitive assemblage" (Hayles, 2017) are monitored, managed, and ultimately re-shaped into elements of its data flow; humans are thus not erased or re-placed but displaced. Hence, as a big data infrastructure it is already always the "analog slot", or, where such a system that appears to be so fully automated is actually full of humans "making little stitches, tacking software to the social world. " (Seaver, 2018, np)

In his reading of the Foucauldian apparatus, Giorgio Agamben divides the world into living beings, and apparatuses; living beings are "incessantly captured" and are immediately subjectified within apparatuses, which in this capitalist moment are only proliferating, he notes. Thus it is from the "relentless fight" between living beings and apparatuses that the subject emerges (Agamben, 2009, p. 14-15). Agamben considers the human in helpless terms, as a "property of discourse" (Packer, 2010, pp 90-91): such as the couch potato, the phone addict who is captured by the media apparatus (Agamben, Op Cit, pp 22-23). Ironically, 'the captured', here, is quite central to the operation of the apparatus; this is the workforce of people who ensure that computer vision systems and auto-pilot settings inside emergent AVs do not fail in the negotiation of the world.

So, the machine's intelligence is also human, situated, and embodied. Intellectual shifts in the history of AI demonstrate how embodiment and situated-ness have been central to the robot's understanding of 'the world', as Rodney Brooks notes, because the brain is a limited metaphor for computational intelligence. The AV quite literally relies on embodied intelligence, but just not its own. This is where fantasy, metaphor, and imaginary count significantly; if we start with the imagination of a cockroach, as Rodney Brooks did, then we might follow a more intriguing and compelling technological development arc. He asked how it is that such small creatures are able to navigate complex new terrain; the answer is that they develop their own internal maps from experience of navigating a space, rather than a representation handed down from elsewhere. 'Insect thinking' is how the Tesla fleet learns and shares information in order to organise and harmonise the collective.

Given this, the word 'ethical' here *should* refer to the conditions of heteromated labour of human micro-workers. Meg Leta Jones says that there is a need to move from the 'human in the loop' approach to automation accountability, to 'tying a policy knot'. She identifies the irony that US automation law builds on humans and machines as separate and joined by a loop, thus not acknowledging the inherently socio- technical nature of automation; and thus, even as it proposes to protect human values, the law actually results in less protection because it understands the two as separate (Jones, 2015). What she is arguing is that automation in such a context is already a socio-technical system, one that is intrinsically human and machine. To think of the regulation of aviation, autonomous driving, or robotics as simply about machines, which a human (or a human legal system) must be added into to oversee, is, she argues, where the problem lies. Jones proposes that the law, and accountability regimes more broadly, must break the loop and 'tie a policy knot' instead, engaging contexts of technology design, implementation, and social relations. What might this look like? For one, more stringent guidelines around the development and public testing of technologies that are not robust in public (Danks, 2017). In contexts where testing takes place unregulated and in public, how are local communities assured that they will not be mis-recognised, or altogether erased by, machine vision? There are ongoing, strong movements of scholars and activists resisting being subjected to algorithmic classification, particularly in the context of facial recognition technologies. What might it mean for these solidarities to go beyond racial justice or personal privacy, and extend to the architectures of computer vision in terms of the people who work in these systems? And, new kinds of social protections and insurance for test drivers, and future drivers or owners who have only limited intervention in the AV's operations. Responsibility lies with car companies to address the practices of the computer vision industry in developing, benchmarking, and rolling out their products. Car manufacturers say they want to protect the humans inside the car, but say little about the humans within the larger system of this altered automobility that includes big data infrastructures of computer vision.[89] There has never been a more important momentfor different industries and kinds of engineers to work

[89] Recently, three German automotive manufacturers, Volkswagen, Audi, and Daimler, collectively invested in Here, a company that makes maps for autonomous urban driving. While they now own a key piece of AV driving technology, this also now increases their liability within the infrastructure of accountability for machine 'seeing'.

together as AI and autonomous systems are widely deployed. There are automated systems that humans cannot intervene in given the pace and extent of automated decision-making; and hence, cannot be held accountable when breakdowns occur. Further, if AVs are more than just cars, and are commercial data platforms, then questions of labour, data protection, and data use must become central as well. Just as gig and platform workers have organised, as have other tech workers in Silicon Valley, what kinds of protections exist for people engaged in developing AV capabilities? Innovation and policy research and advocacy could become more attentive to how multiple new publics and stakeholders are emerging in the shaping of this technology, in addition to traditional institutional actors and investors. All these networks and connections matter and must muddy the discursive construction and emergence of the AV. The irony of autonomy is that it is not about separation or doing things independently, but is enabled by bodies and minds, in connection with, and embodied in assemblages of human and non-human, social, technical, and political.

But, some transformations are subtle and go beyond influences on human subjectivity and work. The shaping of autonomous driving re-configures time and space itself. Successful autonomous driving relies on operations of frictionless exchanges between vast arrays of big data infrastructures. As I have argued here, the ‘fully driverless’ car is already here in terms of all its parts; there just is not enough data, bandwidth, or environmental support in terms of urban infrastructure to sustain the connections between the world, the car, the cloud, the fleet of other cars, and data infrastructures. So collapses and gaps occur. There is the collapse of the territory into the map, aka the ‘map-territory’ (Hind and Gekker, 2019). Just as the driverless car does not ‘see’, it only senses and makes decisions on the basis of statistical correlations in its computer vision, there is the erasure of place and territory. Everything is only the map, and driving has become navigation. And, just as there is a gap that is likely to open up between the model and the world, then there are also infrastructural gaps that threaten to jeopardise ‘autonomy’ and ethical decision-making. For example, we might consider how proposed ethical decision-making is contingent on time itself, like the Time To Reflect Reality (TTRR) metric, the time taken for the map of the world around the car to be sensed, pushed to the cloud, refreshed, and then transmitted out again to the car; or between cars in the fleet. In effect, a gap opens up between the world as it is and the world as it is known to the car or the fleet. This time lag between these different impressions of the world, all equally real and unreal, becomes a factor that proposed ethical decision-making ultimately relies on. If this is the condition in which ‘ethics’ are supposed to unfold in, then we have to ask about the politics of a game-world of maps, cars, and ‘spot the cyclist’. The ethical apparatus manifests these new relationships, knowledge, and conditions arising through adjustments to time and space that current regulatory regimes are hard pressed to comprehend, in the manner laid out by Leta Jones, above. The perspective of an apparatus weaves these dimensions together, alerting us to the transformations foreshadowed in the reorganisation of time and space to enable autonomous driving.

In the following section, I enumerate some limitations in this research, roads less travelled, and possible openings for future research.

Roads less taken

The smart city could perhaps be thought of as the ‘natural’ environment of the driverless car, the latter relying on a networked infrastructure or grid of cameras, and other environmental and surveillance sensors, clouds, mobility ‘corridors’ optimised for more efficient traffic flows, internet-of-things networks, social media. Smartness itself is a kind of apparatus, a “mandate” (Halpern, Mitchell and Geoghesian 2017); Its ideologies, institutions, and sciences create and leverage anxieties borne of real and imagined crises, and then offer us the solutions. The smart city is a public-private partnership where large, powerful corporations provide public services and utilities as a solutionist fix to problems of inefficiency, administrative silos, poor planning and lack of coordination between various agencies. The fix is a turn to an entrepreneurial mode of ‘urban governmentality’ in which the entire city is reframed as a marketplace run by logics of rational, competitive choices are made by consumer-citizens vying for services based on their budgets and existing social power (Graham, Kitchin, Mattern and Shaw, 2019, pp. 3-8). ‘Data’ becomes the epistemological basis for this new solutionism offering insights unavailable to humans and human organisational systems. In practical terms this manifests as a preponderance of digital infrastructure (provided by corporations, many of them technology corporations[90]) for automated decision-making.

[90] For example, Toronto’s now-scrapped Quayside project was a ‘collaboration’ between Alphabet and the city; Facebook provides rural electrification; and internet services through its Free Basics program. Uber and ride-sharing have changed the landscape of public transport.

Cities have always been re-designed to accommodate automobility (Seilor, 2008) Florian Cramer argues that rather than a future of sinister and omniscient AIs, we will see instead experience AI applications silently and slowly creeping up on us by re-shaping our very spaces to accommodate technologies such as AVs (Cramer, 2018, p. 39) Building smart cities in response to and alongside AVs is an ambition that has not materialised yet, except in highly fragmented or contained ways. For example, research test beds in university Transport Research Institutes have model cities that AVs are tested in.[91] Initiatives such as the Bloomberg-Aspen Institute Initiative on Cities and Autonomous Vehicles have developed a detailed 'atlas' to identify the policy and decision-making support, to cities to prepare for, and pilot, AV technologies.[92] It is quite likely that autonomous driving will emerge in restricted spaces like airports, or large campuses, gated residential estates, and as taxi services; and alongside 'micro-mobility' with e-scooters, electric bikes, prams, and motorised wheelchairs. The production of new urban spaces for the automation of driving are potential areas for future research.

The study of AI and ethics offers directions for fresh inquiry into histories of AI, or in terms of the moment of AI's emergence in the mid-Twentieth century alongside a number of other social science disciplines. This might be a way into situating decision-making itself, and approaches to intelligence, reason, and rationality across different epistemological traditions that have influenced the shaping of micro-decisions, and big decisions. For instance, bounded rationality, for which Herbert Simon received a Nobel Prize in Economics, is the study of

[91] A good example of this is M City at the University of Michigan, a city-scale test bed and development environment for driverless cars https://mcity.umich.edu/

[92] https://avsincities.bloomberg.org/global-atlas/

how people's cognitive abilities are just not up to making the best and most optimal choices, and hence make good enough ones; in other words, they 'satisfice'. The entire domain of Behavioural Economics attempts to set up the best 'choice architectures' to influence decision-making that minimises or manages risk, and are more efficient. Decisions can also be shaped and nudged, according to Behavioural Economics. The engineer-ethicist Jason Millar said to me (in an annoyed tone) that the Moral Machine is far from 'ethics' and is more like 'Behavioural Economics'. And, the critical theorist of financial systems, Martha Poon, observes that the self-driving car is imagined as the 'perfect Neoliberal subject' that trundles along making the best decisions for itself. There are some affinities emerging here, between the car imagined as independent, unfettered in making the most rational choices, and always seeking to improve itself; and the field of Behavioural Economics that studies how people make choices, and how to influence them into making better choices.[93] The influences of distinct approaches to human decision-making are entangled with

Finally, continuing with the theme of decision-making, one particular logic of AI and autonomous driving relates to time and temporality. Britt Paris addresses time as a socio-material entity in studies of policy and research projects for building a faster global internet. She tracks how time is an ideology, resource, and motivation in design practices, influencing the shaping of technology, as well as notions of the future. The net result is that modalities of time emerge that weld "human time to machine time ever more tightly, to the point that this

[93] Nudge theory is notorious for its applications in fields as diverse as marketing, to public policy to influence human behaviour to make particular kinds of desired choices. The 2008 book by Richard Thaler and Cass Sunstein *Nudge: Improving Decisions About Health, Wealth, and Happiness*, shot this behavioural economics theory to mainstream attention.

human-technical assemblage decreases humans' capacity to interact with one another by making that interaction appear inefficient, impractical, or aesthetically undesirable." (2021, p. 18) The rationale for autonomous driving is automation is faster and more accurate and hence likely to more respond better in unexpected crash situations, should they even arise. For machine driving is sold as being more efficient than error-prone humans, and assumes that accidents would not happen at all. However, an accurate or efficient response is contingent on temporality; the best response must also be delivered on time, and the best response is also that which is required in and of a moment in time. Moreover, the driverless car is an imaginary of a future moment in time when the urban environment is a 'smart city'. These visions and versions of the future with driverless cars also work to render human decision-making as limited and inefficient; however, as this work as argued, driverless cars are some way from not needing human intervention in order to be efficient.

What is ethics?

This research has identified the material and discursive practices that are shaping autonomy and ethics in the emergence of the driverless car by giving us the terms by which these technologies are understood as being autonomous or ethical. This ethical apparatus has thus shaped what we believe the *ethical* is in relation to AI, algorithmic systems, and autonomous machines more broadly. Based on my analysis, I assemble a set of reflections about ethics in AI and autonomous systems.

There are different kinds of approaches to ethics There are many different approaches to professional and computer ethics already in circulation. Technology ethicists locate ethics in institutional codes, professional cultures, norms, social practices and individual decision-making (cf Ananny 2016, p. 96). Topics under 'applied technology ethics' are broad anddiverse covering, for instance, digital surveillance that maps and tracks our facess and voices, to computing power puts a strain on energy supplies, and amplifies carbon emissions (Vallor et al, 2018, p. 4-5). The recent cases of 'ethics-washing' by Big Tech (Wagner, 2018) indicate, correctly, that discourses of ethics are also about technology companies accruing more power to themselves, to Silicon Valley, and to technologists embedded in government and industry who come up with ethics statements in the first place. This is a self-regulatory move rather than a focus on ethics at all (Ochigame, 2019). The word 'ethics' requires more thoughtful reflection and ownership. But it also risks becoming a bureaucratic exercise that finds no place for meaningful engagement in organisations (Metcalf et al,2020). Machine ethics approaches, or approaches to ethics in the Moral Machine occupy a different relationship to 'technology ethics' as described here. The former might be drawn more precisely in relation to Philosophical approaches to Ethics, such as Utilitarianism, Deontology, or Virtue Ethics; whereas technology ethics are more applied versions of any of these three broad streams and with inputs from other fields of inquiry.

Ethics is really about industrial accountability. The cases of recent AV crashes implicate statistical models inside computer vision systems. In each of these cases, the computer vision systems in the AVs did not identify something in the environment correctly. There has never been a more critical moment for different sections of engineering industries to be in closer conversation. As the automotive and mechanical engineers I have interviewed note, the worlds of software and automative engineering and manufacture tend to be separate in how their industries integrate education about safety and design. There has to be greater commitment within technology industries to turn away from framing ethics as self-regulation, or as something to be encoded into computational systems alone, and begin to examine the wider ethical and social implications of new technologies. . There is an accountability to the public where testing takes place publicly and with few guardrails; and accountability has to permeate multiple industries and actors. In the case of the Arizona crash, it was the test driver who was eventually held to account, not the company or the state government that created a testing environment with poor guardrails.

Ethics of care. Autonomous driving technologies can learn from the context of LAWS where an Ethics of Care approach places people first and focuses on the duties and responsibilities to them. A "Models of Threat approach sees people as statistics, and treats the individuals on a list as threats, whether they have done anything or not, and regardless of whether they are victims or perpetrators–thereby undermining their humanity." (Asaro, p. 2018, 7) Ethics of care are perhaps unsurprising to find in the warfare context, because we see it as an obvious moment of violence. But why is it that care does not permeate urban mobility or smart cities? The emphasis on a specific moment of the crash necessitates ethical approaches that centre decision-making in that moment as a key factor. which, while important, stays silent on the architecture of that moment. What other approaches to ethics could be brought to address the problem of the driverless car as an imaginary and as infrastructure, generate? How would the ethical challenges of this technology be re-framed?

The limits of engineering ethics in individuals and in technical artefacts.. The intention of adapting the Trolley Problem to autonomous driving was to provoke and inspire engineers to think about the outcomes of what they might build, and how to prepare for a future with autonomous driving. However, this has spiralled out of control into a largely uncritical and dramatic media narrative. At the same time, interviews with ethicists and engineers indicate that educating engineers in ethics is not going to be enough for individuals cannot effect change without structural and organisational change.. The practical suggestions made by Bender, Gebru et al. (2020) in their paper, 'Stochastic Parrots', are along the lines of nudging the culture and practices of big tech towards greater equity. Confronting Google's large scale natural language processing models that extract a significant environmental cost, and perpetuate "hegemonic world views", Bender et al., propose practical recommendations that typify ethical technology as a relational, socio-technical process, emphasising *how* the work is done, from *within* the organisation, rather than through external laws or regulation. They assess the downstream effects of technologies; ask if these technologies are beneficial to already-marginal communities around the world; encourage consistent reflection on the values through 'value sensitive design' processes; perform post-mortems and 'pre-mortems' of business products as they get released into the world; and document how datasets are selected and how they inform model-building; among others.[94] In other words, 'ethics' becomes a

[94] This paper and its authors have been in a very public conflict with Google. The second author, Timnit Gebru, used to be the co-lead of Google's Ethical AI Lab. She was fired because of her work on this paper. It is not difficult to see why Google would object to this paper: a star researcher and leader of the Ethical AI team is arguing, in effect, against the status quo, and against the raison d'être of Big Tech in general; the paper is questioning the business strategies and ambitions associated with scale.

more widespread and collective process within an organisation that builds technology rather than integrating ethics into individuals or technical artefacts.

If driverless cars are the future, then which future are we talking about? The socio-technical imaginary of the driverless car makes assumptions about what a better future might be. Something that started as a DARPA challenge has become a huge industrial development project that shows no signs of scaling back despite recent crashes; and while making some outlandish claims along the way. The imagination of an AI future in a smart city with driverless cars has implications for citizens and governments. Yet, there is little by way of public conversation about what kinds of technology futures people want or imagine.[95] Returning to the Moral Machine, its logics of flipping through options to arrive at a ground-truth of moral values is a bit like a game of chance played on a very large scale, suggesting a play of probabilities. Moral Machine's developers propose risk as the epistemological framework for data analysis, and risk is a strong theme that emerges in their writing. 'Risk' and 'risk studies' exceed the scope of this work. However, I conclude that the Moral Machine project is a kind of 'firmative speculation' which "turn[s] uncertainty into (external, calculable, knowable) risk," and "cast[s] futurity primarily in terms of technological progress, economic growth, and a prolongation of the status quo." (Uncertain Commons Collective, 2013, ch. 2) And as such, the Moral Machine proposes to transform the ethical into the optimal, with algorithmic decision-making to spread out and manage risk. The word 'firmative' comes from the German

[95] However, one smart city development has met with stiff public resistance to the point where it folded up. The Alphabet-owned Quayside project, a smart city initiative to 're-develop' Toronto's economically depressed lakefront property, was scrapped in 2019-2020 because of organised public resistance https://medium.com/sidewalk-talk/why-were-no-longer-pursuing-the-quayside-project-and-what-s-next-for-sidewalk-labs-9a61de3fee3a

firma for the ‘agency’ that, since the 1700s, has enumerated, managed, and contained risk and uncertainty for various businesses and ventures. The Uncertain Commons Collective asks how our relationship with the future might become *affirmative* instead; they identify futures produced through possibilities, play, creativity, living in common, collaboration, intuition, and most critically, the embrace of uncertainty (Uncertain Commons Collective, 2013, ch. 1). In the context of this study, I believe this means bringing the ethical into the realm of the *in-computable* rather than one of *uncertainty management.* To explain further, I turn to a concluding set of thoughts about reconceptualising ethics in terms of partiality and refusal.

A partial ethics, an ethics of refusal

In the machine ethics approach, computation itself is the site of ethical decision-making. There are parallels between this and a tension in the sub-domain of Fairness, Accountability and Transparency in Machine Learning (FATML; associated with the FAccT conference discussed earlier). Here, social scientists, lawyers, computer scientists and humanists argue that bias cannot be mitigated out by computational means, but that computation itself is a site for politics and ethics too. On the other, there is a belief that fairness, and un-biased outcomes can be programmed into machine learning. Carly Kind refers to this as the ‘second wave’ of AI ethics that is led by computer scientists and focuses on “technical fixes” (Kind, 2020) This is not unlike how AV ethics imagines the role of computation in generating ethical decisions. In architecting machine systems to putatively regulate themselves, and on that basis to organise the social, we are deploying practices of information compression that reduce complexity while actually requiring a great deal of it; and make correlations and on that basis make predictions about future outcomes. If we want to architect ethical values into such systems then those very ethical values will also be subject to the logics of computation inside AI and algorithmic technologies: compression, a loss of complexity, approximations, correlations, and predictions. What emerges are not 'ethics' but decisions made by computational ways of knowing the world in 'partial' ways (Amoore, 2020). This partiality comes from decisions made in how systems are architected, informed, and trained in response to particular notions of the world. This partiality might be way to re-conceptualise ethics, to acknowledge the facts of mediation of knowledge and reality through AI and algorithmic systems rather than reject or halt them. In other words, we might develop practices, methods, and epistemologies that acknowledge and reveal how the partialities and situated knowledge from these systems influence our thinking about the ethical. We have to acknowledge knowledge-making as processes of constant interaction and mediation between humans and technologies (Verbeek, 2011; Ihde, 1995) and hence scale our valuing of these systems accordingly in how they can act, and what supports they need to work effectively.

The ethics of autonomous driving discourse wants to eventually identify autonomous machines as "explicit ethical agents" that can reason about ethics, or as "full Ethical Agents" that can make explicit moral judgments and justify them (Moor, 2006). In this sense, AI and algorithmic systems are understood as being able to *generate* particular values and outcomes through computational decision-making. I have argued that we need to critically scrutinise *how* AI and algorithmic systems generate create value-laden positions in the emergence of autonomous driving. So, they are generating their own unique ways of knowing, representing, and filtering the world back. My concern is less about the fact of mediation of social life through computation, and more *how* computation perceives and makes sense of the social world, and then re-orders it in its own terms, the means we have to make sense of this re-ordering, and develop appropriate mechanisms of regulation and oversight. While machinic knowing and shaping the world are usually not legible to nor can be regulated by existing conventional human systems of the law, we can of course bring regulation and law to the humans and institutions building these technologies. Data-driven machine learning decisions are knowledge based on making distinctions, categorisations, and hierarchies that have unique origin stories.

When computer scientists, Joy Buolamwini and Timnit Gebrufound that facial recognition systems built by Microsoft, IBM, and Face++ were less competent in identifying darker and more female phenotypes, they went to these companies to share their results. The companies said they would try to improve their systems to be more accurate and to perform better.[96] The inaccuracies emerged because the datasets that inhabit the model of these systems, and that were used to train machine learning, were not optimising for darker and more female phenotypes, and were originally trained on a dataset populated by lighter and more male phenotypes.[97] However, many scholars and activists have pointed out that people of colour have always been identified, surveilled, and marginalised by digital infrastructures, so why would they want these systems to improve in their recognition of Black and Brown faces? (Stark,

[96] https://www.media.mit.edu/projects/gender-shades/faq/#faq-what-did-microsoft-say-about-this-work and https://www.media.mit.edu/projects/gender-shades/faq/#faq-what-did-ibm-say-about-this-work

[97] Results are organised as the ‘Gender Shades’ project where bias is defined as “practical differences in gender classification error rates between groups.” http://gendershades.org/overview.html

2019) In fact, the history of contemporary information technologies might also be told in terms of sorting, classifying, and distinguishing people, and associating hierarchical values with these distinctions, perpetuating violence along gendered, racialised, and embodied lines. Further, race and gender are representational forms interpellated by informational categories to distinguish between bodies, many of them over a hundred years old, and emerging in violent contexts of colonisation. Reducing humans to measurements of their bodies as the basis for scientific or data-based administration has paved the way for deep seated racism (Sekula, 2006; Browne, 2015; Benjamin, 2019; Rosenthal, 2019). And hence these distinctions that we call 'race' or 'gender' are being re-produced by algorithmic systems. Current data sets continue to amplify and shape race in these representational terms. This goes hand in hand with ongoing systemic racialised violence and discrimination. Given what we know about faulty computer vision systems in driverless cars, the outcomes of not being properly recognised are as precarious as being recognised. Further, it is impossible to ensure that a product made by IBM or Microsoft is not used by the police or federal surveillance community. The desire to correct algorithms by building more fair, feminist, race-sensitive or unbiased data sets does not necessarily change the outputs based on discrimination and distinction. Just as Deep Blue did not advance AI and just became a computer very good at beating Gary Kasparov, a more feminist or fair data set only gets better at identifying race based on facial features and phenotypes more accurately.

It is also re-shaping reality. A model that makes mistakes in a predictive policing algorithm is wrongly identifying a gang, or someone as part of a criminal gang based on partial data, one that cannot detect patterns from noise, or cannot detect any patterns at all, could translate into a wrongful arrest, or continue to haunt someone because they appeared as 'suspicious' within the model. The model is thus creating "data derivatives", not what *has* happened, but what *could* (Amoore, 2011); A speculation, a risk calculation. Further, these predictive policing models are based on biased crime data from already highly surveilled communities, creating closed feedback loops where the model replays its past indiscriminately, with no sense of history (Stop LAPD Spying Coalition, 2020) Such prediction is also awarding power to data-driven ways of governing and organising society in these complex areas of social and political life.

Ramon Amaro, a leading philosopher of technology at University College London,[98] changes the narrative towards one of refusal. He asks, if the intention is the *transformation* of the historic, embodied, and material realities of Black life, then what might be a place *beyond* the human-machine dialectic that is a binary one of either being appropriately legible to and in machine systems, or not at all? He asks what an "aspirational black life" might be that strives to "gain a right of refusal to representation" through these systems that are narrow in their construction of representation (Amaro, 2019). In other words, Amaro, wants to locate the transformation of his reality outside of [big] technology entirely. This does not mean a rejection of mediation through computation necessarily, or 'going off the grid'. Rather, he is saying that correcting his 'in-computability' in algorithmic terms, or seeking recognition on the epistemic terms set out by computation are not the starting point for an ethics or for trans-

[98] Dr. Ramon Amaro's writing, research and practice emerge at the intersections of Black Study, psychopathology, digital culture, and the critique of computation reason. https://www.ucl.ac.uk/art-history/dr-ramon-amaro

formation. Thus, reconceptualising ethics through AI technologies means identifying practices and locations in the deliberative and distributed places of human-computational mediation, where autonomy means connection, rather than independence. Drawing on this, an ethics of autonomous driving would not originate in the questions generated by the imagination of a speculative car crash formulated to be legible to a computer. Autonomous driving might originate in the ecologies of mobility, requirements of human transit, considerations of urban space, considerations of energy consumption and sustainability, the requirements of marginalised and differently abled people. An ethics would follow from the politics introducing complex sociotechnical machine systems in human environments. A driverless car could navigate itself somewhere, but only humans can tell it where to go.

REFERENCES

Ackerman E (2016) Self-driving cars were just around the corner—in 1960. *IEEE Spectrum: Technology, Engineering, and Science News*. https://spectrum.ieee.org/selfdriving-cars-were-just-around-the-cornerin-1960

Agamben, G. (2009). *What is an apparatus? And other essays.* D. Kishik and S. Pedatella (Trans) Stanford.

Agre, P. E. (1997). *Computation and human experience.* Cambridge University Press.

AI Debate Games https://debate-game.openai.com and https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/

Alvarez León LF (2019) Eyes on the road: Surveillance logics in the autonomous vehicle economy. *Surveillance & Society* 17(1/2): 198-204. https://ojs.library.queensu.ca/index.php/surveillance-and-society/index

Amaro, R. (14 February 2019). As if, *E-flux architecture* https://www.e-flux.com/architecture/becoming-digital/248073/as-if/

Amoore, L.A. (2011). Data derivatives: on the emergence of a security risk calculus for our times, *Theory, Culture & Society*, 28 (6), pp. 24-43.http://dx.doi.org/10.1177/0263276411417430

Amoore, L.A & Piotukh, V. (2016). (Eds) Introduction: *Algorithmic Life: Calculative Devices In The Age Of Big Data.* London and New York: Routledge.

Amoore L.A (2018). Doubt and the Algorithm: On the Partial Accounts of Machine Learning. *Theory, Culture & Society Special Issue: Transversal Posthumanities* 0(0) 1–23: https://10.1177/0263276419851846

Amoore, L.A (2020). *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham and London: Duke University.

Ananny, M. (2016). Toward an Ethics of Algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values* 41 (1) (2016): 93-117.

Ananny, M and Crawford, K. (2016) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* (2016): 1-17

Anderson, M. and Anderson, S.L (2007). Machine Ethics: Creating an ethical intelligent agent. *AI Magazine,* 28 (4). pp 15-26, https://www.aaai.org/ojs/index.php/aimagazine/issue/view/176

Anderson, M., and Anderson, S.L. (Eds) (2011) *Machine ethics*. Cambridge University Press.

Andrejevic, M. (2019). Automating surveillance. *Surveillance & Society,* 17(1/2), 7-13.

Angwin, J, Larson, J., Mattu, S. and Kirchner, L. (2016, May 23). Machine Bias. *ProPublica*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Asaro P.M. (2019). AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine*, 38(2), 40-53.

Awad E, D'Souza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon J-F, & Rahwan I (2018). The moral machine experiment. *Nature* 563(7729): 59–64, https://doi.org/10.1038/s41586-018-0637-6

Bainbridge, L. (1983). The ironies of automation. *Automatica* (6): 775- 779.

Ballard, J.G. (1973). *Crash*. Jonathan Cape.

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning.* Durham and London: Duke University Press.

Barocas, S., Hood, S., Ziewitz, M. (2013). Governing Algorithms: A Provocation Piece. https://ssrn.com/abstract=2245322

Barrett LF, Adolphs R, Marsella S, Martinez AM, Pollak SD. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements [published correction appears in Psychol Sci Public Interest. 2019 Dec;20(3):165-166]. *Psychol Sci Public Interest.* 2019;20(1):1-68. https://doi:10.1177/1529100619832930

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI & Society*, *35*(1), 165-176. https://10.1007/s00146-017-0760-1

Beiker, S. (2016). Deployment Scenarios for Vehicles with Higher-Order Automation (pp. 193–211). https://doi.org/10.1007/978-3-662-48847-8_10

Belton, O., & Dillon, S. (2021). Futures of autonomous flight: Using a collaborative storytelling game to assess anticipatory assumptions, *Futures*, 128, 102688. https://doi.org/10.1016/j.futures. 2020.102688

Bengler, K., Dietmayer, K., Färber, B., Maurer, M., Stiller, C., Winner, H (2014) : Three Decades of Driver Assistance Systems Review and Future Perspectives. *IEEE Intelligent Transportation*

Systems Magazine, vol. 6, no. 4, Winter 2014, pp. 6-22.

Benjamin, R. (2019) *Race against technology: Abolitionist tools for the new Jim Code.* Polity.

Bertoncello, M. & Wee, D. (2015). Ten ways autonomous driving could redefine the automotive world. McKinsey & Company www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-could-redefine-the-automotive-world

BBC Newsnight (2015, September 21) The Trolley Problem and Ethics of Driverless cars. BBC Newsnight https://www.youtube.com/watch?v=FypPSJfCRFk

Bhargava, V. and Kim, T.W (2018). Autonomous Vehicles and Moral Uncertainty in Patrick Lin, Keith Abney, and Ryan Jenkins (Eds), *Roboethics 2.0: From Autonomous Cars to Artificial Intelligence.* London: Oxford UP https://10.1093/oso/9780190652951.003.0001

Bigman, Y. E., & Gray, K. (2020). Life and death decisions of autonomous vehicles. *Nature*, 579(7797), E1–E2. https://doi.org/10.1038/s41586-020-1987-4

Bijker, W. E., & Law, J.(1994) (Eds.) *Shaping technology/building society: Studies in sociotechnical change.* MIT.

Bissell, D, Birtchnell, T, Elliott, A, Hsu, EL (2020). Autonomous automobilities: The social impacts of driverless vehicles. *Current Sociology*, 68(1), 116–134. https://doi.org/10.1177/0011392118816743

BMW Blog (2015) http://www.bmwblog.com/2015/12/31/bmw-to-show-autonomous-concept-in-2016/

Bogost, I (2018, March 30) Enough with the Trolley Problem, *The Atlantic,* https://www.theatlantic.com/technology/archive/2018/03/got-99-problems-but-a-trolley-aint-one/556805/

Bonnefon, J.-F., Shariff, A., Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573. https://doi.org/10.1126/science.aaf2654

Bonnefon, J-F, Shariff, A, Rahwan, I. (2019) The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars. *Proceedings of the IEEE,* Vol. 107, No. 3, March 2019

Bonnemains, V., Saurel, C. & Tessier, C. (2018).Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology,* 20:41–58 https://link.springer.com/article/10.1007/s10676-018-9444-x

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science*, 29(7), 1084–1093. https://doi.org/10.1177/0956797617752640

Boudette NE (2016, September 9). Autopilot Cited in Death of Chinese Tesla Driver. *The New York Times.* https://www.nytimes.com/2016/09/15/business/fatal-tesla-crash-in-china-involved-auto-pilot-government-tv-says.html

Bowker, G. C., & Star, S.L. (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT.

Box, G. E. P. (1976), Science and statistics. *Journal of the American Statistical Association*, 71 (356): 791–799, https://doi:10.1080/01621459.1976.10480949

Boyd, R (1993). Metaphor and theory change: What is 'metaphor' a metaphor for? In A Ortony (Ed) *Metaphor and thought*, 2nd edn. Cambridge: Cambridge University, pp. 481-532.

Braidotti, R. (2019). A theoretical framework for the critical posthumanities. *Theory, culture & soci ety*, 36(6), 31-61.

Bratton. (2016). *The Stack* (1st ed.). MIT Press.
https://doi.org/10.7551/mitpress/9780262029575.001.0001

Braun, R., & Randell, R. (2020). Futuramas of the present: The "driver problem" in the autonomous vehicle sociotechnical imaginary. *Humanities and Social Sciences Communications*, *7*(1), 163. https://doi.org/10.1057/s41599-020-00655-z

Brennan-Marquez, K, Susser, D., & Levy, K (2019) Strange Loops: Apparent versus Actual Human Involvement in Automated Decision-Making. 34 *Berkeley Technology Law Journal,* 745–771 (2019) https://ssrn.com/abstract=3462901

Bridle, J. (2016). Cloud Index http://cloudindx.com/history

Brooks, R. (1991, September 13). New approaches to robotics. *Science*, 13 Sep 1991, Vol. 253, Issue 5025, pp. 1227-1232, https://10.1126/science.253.5025.1227

Brooks, R. (2012). Is the brain a good model for machine intelligence? (2012). *Nature*, 482(7386), 462–463, https://doi.org/10.1038/482462a

Brown, A (2006) Accidents, engineering, and history at NASA 1967-2003 in Steven J. Dick and R. D. Launius (Eds) *Critical issues in the history of spaceflight.* Office of external relations, History division, United States National Aeronautics and Space Administration (NASA), Washington DC pp. 377-402.

Browne, S. (2015). *Dark matters: On the surveillance of blackness,* Durham*:* Duke University.

Bruder, J. (2020). *Cognitive Code: Post-anthropocentric intelligence and the infrastructural brain.* Montreal: McGill-Queen's University.

Buolamwini J, & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commer cial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81:77-91 http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Calo, R. (2015) Robotics and the lessons of cyberlaw. *Calif. L. Rev*. 513 (2015)

Cassani Davis, L. (2015, October 9) Would you pull the trolley switch? Does it matter?, *The Atlantic*, https://www.theatlantic.com/technology/archive/2015/10/trolley-problem-history-psychology-morality-driverless-cars/409732/

Castañeda, C., & Suchman, L. (2014). Robot visions. *Social Studies of Science*, 44(3), 315-341.

Cave, S. Nyrup, R., Vold, K., & Weller, A. (2019). Motivations and Risks of Machine Ethics. *Proceedings of the IEEE*, 107(3), 562–574. https://doi.org/10.1109/JPROC.2018.2865996

Cavoli C, Phillips B, Cohen T, Jones P (2017). Social and behavioral questions associated with automated vehicles: A literature review. UCL Transport Institute, London. https://www.ucl.ac.uk/transport-institute/pdfs/social-and-behavioural-literature-review.pdf

Chamayou, G. (2013/2015) *Drone theory*. Janet Lloyd (Trans). Penguin Random House.

Cheney-Lippold J (2019). Accidents happen. *Social Research: An International Quarterly,* (86)2 Summer 2019, p 513-535.

Chun, W. H. K. (2011). *Programmed visions: Software and memory*. MIT.

Chun, W. H. K. (2012). Race and/as Technology, or How to do Things to Race In L. Nakamura and P. Chow-White (Eds), *Race after the Internet.* New York and London: Routledge. pp. 38-60.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral considera tion. *Ethics and information technology,* 12(3), 209-221.

Comma AI (2020) Towards a superhuman driving agent. *Medium,* https://comma-ai.medium.com/towards-a-superhuman-driving-agent-1f7391e2e8ec

Costanza-Schock, S (2020). Design Justice, A.I., and Escape from the Matrix of Domination. *MIT Journal of Design Studies.* https://jods.mitpress.mit.edu/pub/costanza-chock/release/4

Cramer, F. (2018). Crapularity Hermeneutics: Interpretation as the Blind Spot of Analytics, Artificial Intelligence, and Other Algorithmic Producers of the Postapocalyptic Present. In C. Apprich, WHK Chun, F Cramer (Eds) P*attern Discrimination*. Meson pp. 23–58.

Crawford, K. (2016, June 25) Opinion. Artificial Intelligence's White Guy Problem. *The New York Times.* https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. (2019). AI Now 2019 Report. New York: AI Now Institute, 2019, https://ainowinstitute.org/AI_Now_2019_Report.html.

Crawford, K. & Joler, V. (2018). Anatomy of AI: The Amazon Echo as an anatomical map of human labor, data and planetary resources. AI Now Institute and Share Lab, https://anatomyof.ai

Criddle, C. (2021, April 28).“Self-driving” cars to be allowed on UK roads this year. *BBC News,* https://www.bbc.com/news/technology-56906145

Cummings, M.L., Mastracchio, C., Thornburg, K.M., Mkrtchyan, A (2013) Boredom and distraction in multiple unmanned vehicle supervisory control. *Interacting with Computers,* (25)1:34-47

Cummings M.L. (2014) Man vs. Machine or Man + Machine? *IEEE Intelligent Systems,* 29(5): 62-69

Cummings, M.L., & J. C Ryan, “Who Is in Charge? Promises and Pitfalls of Driverless Cars.” TR News, (May-June 2014) 292, p. 25-30. https://hal.pratt.duke.edu/sites/hal.pratt.duke.edu/files/u7TR%20news%20Cummings%20MAR14.pdf

Cummings, M.L. (2017a) Informing Autonomous System Design Through the Lens of Skill-, Rule-, and Knowledge-Based Behaviors. *Journal of Cognitive Engineering and Decision Making*, 12(1), 58–61. https://doi.org/10.1177/1555343417736461

Cummings, M.L. (2017b) Artificial intelligence and the future of warfare. Chatham House.International Security Department and US and the Americas Programme. https://www.chathamhouse.org/2017/01/artificial-intelligence-and-future-warfare

Danks, D., & London, A. J. (2017). Regulating autonomous systems: Beyond standards. *IEEE Intelli gent Systems*, 32(1), 88-91.

Dant, T. (2004) The Driver-Car. T*heory, Culture & Society Special Issue on Automobilities,* 21(4/5): 61–79. https://doi.org/10.1177/0263276404046061

Darling, K. (2015). "Who's Johnny?" Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. *Social Science Research Network.* https://doi.org/10.2139/ssrn.2588669

Darpa Urban Challenge (2007) https://archive.darpa.mil/grandchallenge/overview.html

Dastin, J. (2018, October 21) Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Delfanti, A. & Frey, B. (2020). Humanly extended automation or the future of work seen through Amazon patents. *Science, Technology, & Human Values*, 1-28, https://10.1177/0162243920943665

Deloitte Touche Tomahtsu (2018) Autonomous Driving Moonshot Project with Quantum Leap from Hardware to Software & AI Focus. https://www2.deloitte.com/content/dam/Deloitte/de/Documents/consumer-industrial-products/POV_Autonomous-Driving_Deloitte.pdf

Derrida, J. (1993/2006). *Specters of Marx: The state of the debt, the work of mourning and the new international.* Routledge. ISBN 9780415389570

Dixon L (2019). Autonowashing: The Greenwashing of Vehicle Automation.

https://10.13140/RG.2.2.19836.69761

Dixon-Roman, E. (2016). Algo-Ritmo: More-Than-Human Performative Acts and the Racializing Assemblages of Algorithmic Architectures. *Cultural Studies ↔ Critical Methodologies,* 2016, Vol.16(5) 482–490. https://10.1177/1532708616655769

Doctorow, C. (2015, December 23) The problem with self-driving cars: who controls the code? *The Guardian.* https://www.theguardian.com/technology/2015/dec/23/the-problem-with-self-driving-cars-who-controls-the-code

Dodge, M. & Kitchin, R. (2006). Code, vehicles and governmentality: The automatic production of driving spaces (NIRSA) Working Paper Series. No.29 [Monograph]. NIRSA - National Institute for Regional and Spatial analysis. http://www.nuim.ie/nirsa/research/documents/WPS29.pdf

Doebbe, R. &Ames, M. (2019, February 9) Up Next For FAT*: From Ethical Values To Ethical Pratices. *Medium* https://medium.com/@roeldobbe/up-next-for-fat-from-ethical-values-to-ethical-practices-ebbed9f6adee

Dogan, E., Chatila, R., Chauvier, S., Evans, K., Hadjixenophontos, P., & J Perrin (2016) Ethics in the design of automated vehicles: the AVEthics project. https://www.researchgate.net/publication/306908478

Douglas, M. (1990). Risk as a Forensic Resource. *Daedalus*, 119(4), 1-16.

Dourish, P., & Bell, G. (2014). “Resistance is futile”: Reading science fiction alongside ubiquitous computing. *Personal and Ubiquitous Computing*, 18(4), 769–778. https://link.springer.com/article/10.1007/s00779-013-0678-7

Downer, J. (2007). When the chick hits the fan: representativeness and reproducibility in technologi cal tests. *Social Studies of Science*, 37(1), 7-26.

Downer, J. (2011). "737-Cabriolet": the limits of knowledge and the sociology of inevitable failure. *American Journal of Sociology*, 117(3), 725-762.

Dwyer B (2020, February 11) Self-driving car dataset missing labels for hundreds of pedestrians. *Roboflow Blog* https://blog.roboflow.com/self-driving-car-dataset-missing-pedestrians/

Easterling. (2014). *Extrastatecraft*. Verso.

Ekbia, H. & Nardi, B. (2014) Heteromation and its (dis)contents: The invisible division of labor be tween humans and machines. *First Monday,* 19(6) https://doi.org/10.5210/fm.v19i6.5331

Ekbia, H, Nardi, B. (2018). From form to content. *Cultural Anthropology.* (33)3: 360-367. ISSN 0886-7356

Elish M.C. & Hwang, T. (2015). Praise the machine! Punish the human! The contradictory history of accountability in automated aviation. Comparative Studies in Intelligent Systems – Working Paper #1 Intelligence and Autonomy Initiative1. Data & Society.
https://www.datasociety.net/pubs/ia/Elish-Hwang_AccountabilityAutomatedAviation.pdf

Elish, M.C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science,Technology, and Society,* 5 (2019), 40-60, https://10.17351/ests2019.260

Ensmenger, N. (2012). Is Chess the drosophila of artificial intelligence? A social history of an algorithm. *Social Studies of Science*, Vol. 42, No. 1 (February 2012), pp. 5-30.

Ernst, C., Schröter, J., and Sudmann,A (2019) AI and the Imagination to Overcome Differ ence, in *Spheres: The Spectre of AI*, Issue #5. Leuphana University.

Escobar, A., Hess, D., Licha, I., Sibley, W., Strathern, M., & Sutz, J. (1994). Welcome to Cyberia: Notes on the Anthropology of Cyberculture [and comments and reply]. *Current anthropology,* 35(3), 211-231.

Etzioni, A. and Etzioni, O. (2016) AI assisted ethics. *Ethics of Information Technology,* 18:149–156, https://10.1007/s10676-016-9400-6

Fairley, P. (2017, January 31) The Self-Driving Car's Bicycle Problem. *IEEE* https://www.spectrum.ieee.org/cars-that-think/transportation/self-driving/the-selfdriving-cars-bicycle-problem

Fisch, M. (2013). Meditations on the Unthinkable (soteigai). In E.G Solomon (Ed) *The Space of Disaster*. Resling Publishing.

Foot, P. (1977/2002) The Problem of Abortion and the Doctrine of the Double Effect In Virtues and Vices and Other Essays in Moral Philosophy. *Oxford Review* No. 5 (1967)

Foucault M (1972/2002). *The archaeology of knowledg*e. AM Sheridan Smith (Trans). London and New York: Routledge.

Foucault, M. (1980). *Power/Knowledge: Selected interviews and other writings, 1972-1977*. G Colin (Ed). New York: Harvester.

Foucault, M. (1994). Ethics, subjectivity, and truth. In P Rabinow (Ed) *The essential works of Foucault 1954-1984, Vol 1* New York: The New Press.

Friedman, B. & Nissenbaum, H. (1997) (Eds). *Human Values and the Design of Computer Technology*, Cambridge: Cambridge University.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines,* 30(3), 411-437. https://arxiv.org/abs/2001.09768

Galison, P. (1996). Computer simulation and the trading zone. In P Galison and D.J. Stump (Eds) *The Disunity of Science: Boundaries, Contexts, and Power,* Stanford: Stanford UP, pp. 118–57.

Galison, P. (2000). An Accident of History. In P. Galison and A. Roland (Eds) *Atmospheric Flight in the Twentieth Century*. Springer Science and Business Media, pp. 3-43.

Galloway A.R. (2012). *The interface effect*. Polity.

Ganesh, M.I. (2016). Entering the Factory. *Cyborgology,* https://thesocietypages.org/cyborgology2016/10/28/entering-the-factory/

Ganesh, M.I. (2018). The center for humane tech doesn't want your attention. *Cyborgology,* https://thesocietypages.org/cyborgology/2018/02/09/the-center-for-humane-technology-doesnt-want-your-attention/

Ganesh, M.I. (2020). The ironies of autonomy. *Nature Humanit Soc Sci Commun* 7, 157 (2020). https://doi.org/10.1057/s41599-020-00646-0

Gerla, M., Lee, E-K., Pau, G., Lee, U (2014) Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds. *IEEE World Forum on Internet of Things (WF-IoT). IEEE*, 2014, pp. 241–246

Gibbs, S. (2016, March 30) Microsoft's racist chatbot returns with drug-smoking Twitter meltdown. *The Guardian,* https://www.theguardian.com/technology/2016/mar/30/microsoft-racist-sexist-chatbot-twitter-drugs

Gilman, N. & Ganesh, M.I. (2020) Making sense of the unknown: AI's metaphors. In The Rockefeller Foundation (Ed), *AI+1: Shaping our integrated future* New York, The Rockefeller Foundation.

Gitelman, L & Jackson, V. (2013). Introduction. In L Gitelman (Ed) *Raw data is an oxymoron*. MIT, pp. 1-14.

Goggoll, J. and Müller, J.F. (2017). Autonomous cars: In favour of a mandatory ethics setting. *Science and Engineering Ethics*. 23:681–700, https://10.1007/s11948-016-9806-x

Goodall, N. (2014) Ethical Decision Making During Automated Vehicle Crashes. Transportation Re search Record: Journal of the Transportation Research Board. 2424. 58-65. 10.3141/2424-07. p 5

Goodall, N. (2016) Can you program ethics into a self-driving car? *IEEE Spectrum*, June 2016, https://10.1109/MSPEC.2016.7473149

Google. Self Driving Car Project. Monthly Report. Aug. 2015. https://www.driverlesstransportation.com/media/google.com/en/selfdrivingcar/files/reports/report-0815.pdf

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. https://doi.org/10.24251/HICSS.2019.258

Gunkel, D. J. (2018). *Robot rights*. MIT.

Halpern, O. (2014) *Beautiful data: A history of vision and reason since 1945.* Duke University.

Halpern, O, Mitchell, R. Geoghesian (2017). The Smartness Mandate: Notes toward a Critique, *Grey Room*, no. 68 (Summer 2017): 106–129, https://doi:10.1162/GREYa00221

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 501–512. https://doi.org/10.1145/3351095.3372826

Hansen, E. E. (2016, September 8). Dear Mark. I am wring this to inform you that I shall not comply with your requirement to remove this picture. *Aftenposten*: https://www.aftenposten.no/meninger/kommentar/i/G892Q/dear-mark-i-am-writing-this-to-inform-you-that-i-shall-not-comply-with-your-requirement-to-remove-this-picture

Haraway, D. (1988) Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*. Fall 1988; 14, 3.

Haraway, D. (1989) *Primate Visions: Gender, Race and Nature in the World of Modern Science.* New York: Routledge.

Hassabis, D (2016). Alpha Go: Using Machine Learning to Master the Ancient Game of Go. Google Blog. www.googleblog.blogspot.de/2016/01/alphago-machine-learning-game-go.html

Harwell, D. (2018, November 23). Wanted: The ‘perfect babysitter.’ Must pass AI scan for respect and attitude. *The Washington Post* https://www.washingtonpost.com/technology/2018/11/16/wanted-perfect-babysitter-must-pass-ai-scan-respect-attitude/?noredirect=on&amp;utm_term=.efea589e7bb4

Hayles, N. K. (2005) Computing The Human. *Theory, Culture & Society,* Vol. 22(1) pp 131–151, https://10.1177/0263276405048438

Hayles, N. K. (2017). *Unthought: The power of the cognitive nonconscious.* University of Chicago.

Hawkins, A. J. (2020, February 26). Everyone hates California’s self-driving car reports. *The Verge.* https://www.theverge.com/2020/2/26/21142685/california-dmv-self-driving-car-disengagement-report-data

Himmelreich, J. (2018). Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. *Ethical Theory and Moral Practice*, 21(3), 669–684. https://link.springer.com/article/10.1007/s10677-018-9896-4

Hind, S & Gekker, A (2019) On Autopilot: Towards a Flat Ontology of Vehicular Navigation in C. Lukinbeal, L. Sharp, E.Sommerlad, & A. Escher *Media’s Mapping Impulse.* Franz Steiner Verlag pp 141-160, https://elibrary.steiner-verlag.de/book/99.105010/9783515124256

Hoffmann, A.L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse, *Information, Communication & Society*, 22:7, 900-915.

Hood , C. (2007) Intellectual obsolescence and intellectual makeovers: Reflections on the tools of government after two decades. *Governance*. 2007 Jan;20(1):127-44.

Hu, T. H. (2015). *A Prehistory of the Cloud*. MIT.

Hudda, R., Kelly, C., Long, G., Luo, J., Pandit, A., Phillips, D., Sheet, L. and I Sidhu (2013) Self-driving cars, Coleman Fung College of Engineering, University of California, Berkeley. Fung Technical Report No. 2013.05.29.https://www.coursehero.com/file/35443675/self-driving-carspdf/ (*Document removed from Berkeley online repository)*

Hurlbut, J. B. (2015). Remembering the future: science, law, and the legacy of Asilomar. In S. Jasanoff and S-H Kim (Eds) *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*, pp 126-51.

Hutson M.,(2018). Artificial intelligence could identify gang crimes—And ignite an ethical firestorm. *Science,* AAAS. https://www.science.org/content/article/artificial-intelligence-could-identify-gang-crimes-and-ignite-ethical-firestorm

Ihde D (1995). *Postphenomenology: Essays in the postmodern context*. Northwestern University Press.

Ingold, D. & Soper, S (2016, April 21) Amazon doesn't consider the race of its customers. Should it? *Bloomberg,* https://www.bloomberg.com/graphics/2016-amazon-same-day/

Irani L.C., Silberman, M.S. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. Proceedings of Conference of Human Factors in Computing Systems CHI 2013, Apr 28-May 2, 2013. Available at https://escholarship.org/uc/item/10c125z3

Irani, L. (2015). Difference and dependence among digital workers: the case of Amazon Mechanical Turk. *South Atlantic Quarterly* 114(1): 225–234.

Ito, M., & Okabe, D. (2005). Technosocial situations: Emergent structurings of mobile email use. *Personal, portable, pedestrian: Mobile phones in Japanese life*, 20(6), 257-273.

Jasanoff, S. (2015) Future imperfect: Science, technology and the imaginations of modernity in S. Jasanoff and S-H. Kim (Eds) *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power.* University of Chicago. pp. 2-33

Jobin A, Ienca, M. & Vayena, E. (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, D (2011) Software Agents, Anticipatory Ethics, and Accountability in G.E. Marchant et al. (eds.), The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight, *International Library of Ethics, Law and Technolog*y 7, Springer Science+Business Media B.V. pp 61-77. http:// 10.1007/978-94-007-1356-7_5

Kahn, J. (2018, August 16). To get ready for robot driving, some want to reprogram pedestrians. *Bloomberg,* https://www.bloomberg.com/news/articles/2018-08-16/to-get-ready-for-robot-driving-some-want-to-reprogram-pedestrians?leadSource=uverify%20wall

Karp, P., & Knaus, C. (2018, April 4). Centrelink robo-debt program accused of enforcing “illegal” debts. *The Guardian,* https://www.theguardian.com/australia-news/2018/apr/04/centrelink-robo-debt-program-accused-of-enforcing-illegal-debts

Karppi, T. (2018). “The Computer Said So”: On the Ethics, Effectiveness, and Cultural Techniques of Predictive Policing. *Social Media + Society*, April-June 2018: 1–9.

Karppi, T., Böhlen, M., & Granata, Y. (2018). Killer Robots as cultural techniques. *International Journal of Cultural Studies*, 21(2), 107–123. https://doi.org/10.1177/1367877916671425

Katz, Y. (2020) *Artificial whiteness: Politics and ideology in Artificial Intelligence*. New York: Columbia University.

Katzenbach, C. & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review*, 8(4). https://doi.org/10.14763/2019.4.1424

Keller, R. (2011) The Sociology of Knowledge Approach to Discourse (SKAD). *Hum Stud* 34:43–65, https://10.1007/s10746-011-9175-z

Kerry, C.F. & Karsten, J. (2017). Gauging investment in self-driving cars. Brookings. 16 October 2017 https://www.brookings.edu/research/gauging-investment-in-self-driving-cars/

Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1-22.

Kim, R., Kleinman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S.,Tenenbaum, J., and Rahwan, I. (2018). A Computational Model of Commonsense Moral Decision Making. MIT, Cambridge. Paper presented at the Artificial Intelligence, Ethics and Society Conference, New Orleans, LA. Jan 31-Feb 2018. http://www.aies-conference.com/2018/accepted-papers/

Kind, C. (2020) The term ‘ethical AI’ is finally starting to mean something. *Venturebeat*, 23 August 2020. https://venturebeat.com/ai/the-term-ethical-ai-is-finally-starting-to-mean-something/

Kirby, D. (2010). The future is now: Diegetic prototypes and the role of popular films in generating real-world technological development. *Social Studies of Science,* 40(1), 41-70.

Kitchin, R. (2014). Thinking critically about and researching algorithms. Programmable City Working Paper 5. http://eprints.maynoothuniversity.ie/5715

Kitchin, R., & Lauriault, T (2014) Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work. The Programmable City Working Paper 2; pre-print version of chapter to be published in Eckert, J., Shears, A. and Thatcher, J. (eds) Geoweb and Big Data. University of Nebraska Press. Forthcoming, https://ssrn.com/abstract=2474112

Kitchin, R. & Lauriault, T. (2018) Digital data and data infrastructures. In J. Ash, R. Kitchin, and A. Leszczynski, (Eds) *Digital Geographies.* London: Sage. pp. 83-94.

Knight, W (2017, September 20) Finally, a Driverless Car with Some Common Sense. *MIT Technology Review,* https://www.technologyreview.com/2017/09/20/149046/finally-a-driverless-car-with-some-common-sense/

Knorr Cetina, K.D (1994). Primitive classification and postmodernity: Towards a sociological notion of fiction. *Theory, Culture & Society*, 11(3), 1-22.

Knorr Cetina, K. D (2007). Culture in global knowledge societies: Knowledge cultures and epistemic cultures. *Interdisciplinary Science Reviews*, 32(4), 361–375. https://doi.org/10.1179/030801807X163571

Kraemer, F., van Overveld, K. & Peterson, M. (2011) Is there an ethics of algorithms? *Ethics of Information Technology,* 13:251–260.

Kröger, F. (2015) Automated driving in its social, historical and cultural contexts. In M.J. Maurer, C. Gerdes, B. Lenz, H.Winner H (Eds) *Autonomous driving: Technical, legal and social aspects*. Heidelberg and Berlin: Springer, pp 41-68. https://doi.org/10.1007/978-3-662-48847-8

LaFrance, A. (2016, March 1) What should we call self-driving cars? *The Atlantic*, www.theatlantic.com/technology/archive/2016/03/what-should-we-call-self-driving-cars/471711/

Latour, B (1992) Where are the missing masses? In W.E. Bijker and J. Law (Eds) *Shaping Technology/Building Society: Studies in Sociotechnical Change*. MIT Press, pp. 225–258.

Leonardi, P (2010) From Road to Lab to Math: The Co-evolution of Technological,Regulatory,and Organizational Innovations forAutomotive Crash Testing, *Social Studies of Science,* 40/2; 243–274.

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute*. https://zenodo.org/records/3240529

Lewis, S. (2019, April 25). The Racial Bias Built Into Photography. *The New York Times,* https://www.nytimes.com/2019/04/25/lens/sarah-lewis-racial-bias-photography.html

Jones M.L. (2015). The ironies of automation law: Tying policy knots with fair automation practices principles. *Vanderbilt Journal of Entertainment & Technology Law*, 18, 77.

Levin, S. (2020, September 16). Safety Driver Charged in 2018 Incident Where Self-Driving Uber Car Killed a Woman. *The Guardian* https://www.theguardian.com/us-news/2020/sep/16/uber-self-driving-car-death-safety-driver-charged

Lin, P. (2013, October 8). The ethics of autonomous cars, *The Atlantic,* https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/

Lin, P. (2015a). Why Ethics Matters for Autonomous Cars. in M. J. Maurer, C. Gerdes, B.Lenz and H.Winner (eds) *Autonomous Driving: Technical, Legal and Social Aspects.* Springer. pp 69-85.

Lin, P. (2015b). The Ethical Dilemmas of Self Driving Cars. TED Talk. https://ed.ted.com/lessons/the-ethical-dilemma-of-self-driving-cars-patrick-lin

Lingel, J., & Crawford, K. (2020). Alexa, Tell Me about Your Mother, *Catalyst: Feminism, Theory, Technoscience*, 6(1), Article 1. https://doi.org/10.28968/cftt.v6i1.29949

Luetge, C. The German Ethics Code for Automated and Connected Driving. *Philos. Technol.* (2017) 30:547. https://doi.org/10.1007/s13347-017-0284-0

Lutz, C. &Tamo, A. (2015) RoboCode-Ethicists – Privacy-friendly robots, an ethical responsibility of engineers? WebSci '15, June 28 - July 01, 2015, Oxford, United Kingdom. https://dl.acm.org/doi/10.1145/2786451.2786465

MacKenzie, D., & Wajcman, J. (1999). *The social shaping of technology*. Open university press. http://eprints.lse.ac.uk/28638/

Marcus, G. (2012, November 24) Moral machines. *The New Yorker.* https://www.newyorker.com/news/news-desk/moral-machines

Marra, W.C. & McNeil S.K. (2012). Understanding “The Loop”: Regulating the Next Generation of War Machines, 36:3 *Harvard Journal of Law and Public Policy,* 1139.

Marvin, C. (1988). *When old technologies were new: Thinking about electric communication in the late nineteenth century.* Oxford University.

Matsuzaki, H. & Lindemann, G. (2016). The autonomy-safety-paradox of service robotics in Europe and Japan: A comparative analysis. *AI & Society*, 31(4), 501–517. https://doi.org/10.1007/s00146-015-0630-7

Mattern, S. (2017) Mapping’s intelligent agents. *Places Journal* https://placesjournal.org/article/mappings-intelligent-agents/

Maurer M.J., Gerdes, C., Lenz, B., Winner, H. (2015) (Eds) *Autonomous driving: Technical, legal and social aspects.* Heidelberg and Berlin: Springer, https://doi.org/10.1007/978-3-662-48847-8

Mbembe, A (2019, September 5) Thoughts on the planetary: An interview with Achille Mbembe. *New Frame,* https://www.newframe.com/thoughts-on-the-planetary-an-interview-with-achille-mbembe

McCarthy, J. (1988) Mathematical Logic in Artificial Intelligence, *Daedalus* Vol. 117, No. 1, *Artificial Intelligence* (Winter, 1988), pp. 297-311

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In M. Kaufmann (Ed) Readings in artificial intelligence (pp. 431-450). http://www-formal.stanford.edu/jmc/

McDermott, D. (2008) Why Ethics is a High Hurdle for AI. In: North American Conference on Computers and Philosophy (NACAP 2008), Bloomington, Indiana (July 2008), http://cs-www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf

McGuigan, J. (2013). Mobile Privatisation and the Neoliberal Self. *Key Words: A Journal of Cultural Materialism,* (11), 75-89.

McLuhan, M. (1964/1994) *Understanding media: The extensions of man*. MIT Press.

McNeese, N.J, Demir, M., Cooke, N.J., Myers, C. (2018) Teaming with a synthetic teammate: insights into human-autonomy teaming. *Hum. Fact*. 60, 262–273. https://10.1177/0018720817743223

Metcalf, J., Moss, E., boyd, d., (2019) Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly,* Volume 82, Issue 2, Summer, 2019, pp. 449-476. https://muse.jhu.edu/article/732185

Millar, J. (2015a) Technology as Moral Proxy: Autonomy and Paternalism by Design, *IEEE Technology and Society Magazine*, vol. 34, no. 2, pp. 47-55, June 2015,

Millar, J. (2015b). Technological Moral Proxies and the Ethical Limits of Automating Decision-Mak ing In Robotics and Artificial Intelligence (Doctoral dissertation). https://qspace.library.queensu.ca/server/api/core/bitstreams/4cb9c561-1fdb-46c3-a39b-ebc77a960d00/content

Mitchell, R (2021, May 17). DMV probing whether Tesla violates state regulations with self-driving claims. *Los Angeles Times,* https://www.latimes.com/business/story/2021-05-17/dmv-tesla-california-fsd-autopilot-safety

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), https://doi.org/10.1177/2053951716679679

Mladenović M.N, Lehtinen. S, Soh. E, & Martens, K. (2019). Emerging Urban Mobility Technologies through the Lens of Everyday Urban Aesthetics: Case of Self-Driving Vehicle*. Essays in Philosophy,* 20(2), 146-170.

Mobileye (2019). Implementing the RSS model on NHTSA pre-crash scenarios https://www.mobileye.com/responsibility-sensitive-safety/

Mobility Round Table (2020, November 8) Mobility Roundtable online seminar with Frances Berman, Tim Dawkins, and Linnet Taylor. https://youtu.be/snlXZfwYATU

Mol, A (2002) *The body multiple*. Duke University.

Moor, J. H. (2006) The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21(4) (2006): 18–21.

Move Labs https://move-lab.space/ (Project discontinued as of September 2020)

Mozilla Foundation, 2018. https://foundation.mozilla.org/en/initiatives/responsible-cs/

Natale, S. and Ballatore, A. (2017) Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence*: *The International Journal of Research into New Media Technologies*: 1–16 . https://doi.org/10.1177/1354856517715164

National Transportation Safety Board (2017) Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida. May 7, 2016. Ac cident Report. NTSB/HAR-17/02 PB2017-102600. https://data.ntsb.gov/Docket/Forms/search-docket

National Transportation Safety Board (2018) Preliminary report highway: HWY18MH010. https://www.ntsb.gov/investigations/AccidentReports/Pages/HWY18MH010-prelim.aspx

National Transportation Safety Board (2019) Public meeting. Collision between vehicle controlled by developmental automated driving system and pedestrian. November 19, 2019. Available https://www.ntsb.gov/news/press-releases/Pages/NR20191119c.aspx

Neyland, D. (2016) Bearing account-able witness to the ethical algorithmic system. *Science, Technology, & Human Values*, 41(1), pp.50-76.

Nissan Labs (2019) Human Autonomy Teaming. https://twitter.com/sladner/status/1193576799772520448?s=03

Nissenbaum, H. (2005) *Values* in Technical *Design, Introduction to the Encyclopedia of Science Technology and Ethics* New York: MacMillan. pp: Ixvi-lxx.

Noble, S. (2017) *Algorithms of oppression: How search engines reinforce racism*. NYU.

Noothigattu, R., D'Souza, S., Gaikwad, S.S., Awad, E., Rahwan, I. Ravikumar, P., Procaccia. A.D. (2017). A Voting-Based System for Ethical Decision Making. https://arXiv1709.06692v1cs.AI

Nordmann, A. (2007) If and Then: A Critique of Speculative NanoEthics, *Nanoethics* 1:31–46 https://10.1007/s11569-007-0007-6

Nordmann, A., & Rip, A. (2009). Mind the gap revisited. *Nature nanotechnology*, 4(5), 273-274.

Norton P (2011) *Fighting traffic: The dawn of the motor age in the American City*. MIT.

Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: An Ap plied Trolley Problem? *Ethical Theory and Moral Practice*, 19(5), 1275–1289. https://link.springer.com/article/10.1007/s10677-016-9745-2

Offert, F. & Bell, P. (2020). Perceptual bias and technical metapictures: Critical machine vision as a humanities challenge. *AI & Society,* https://doi.org/10.1007/s00146-020-01058z

Oliver, N., Calvard, T. and Potočnik, K (2017). Cognition, Technology, and Organizational Limits: Lessons from the Air France 447 Disaster. Organization Science 28:4, pp 729-743 https://doi.org/10.1287/orsc.2017.1138

Olson, E. (2019, February 28). The Moore's Law for Self-Driving Vehicles. *Medium,* https://medium.com/may-mobility/the-moores-law-for-self-driving-vehicles-b78b8861e184

O'Malley J (2018) Captcha if you can: How you've been training AI for years without realizing it. *Tech Radar,* https://www.techradar.com/news/captcha-if-you-can-how-youve-been-training-ai-for-years-without-realising-it

Oswald, K.F &Packer, J. (2013) Flow and mobile media. *Communication matters: Materialist approaches to media, mobility and networks*, 276-287.

Packer, J. (2010) What is an Archive? An Apparatus Model for Communications and Media History. *The Communication Review*, 13:1, 88-104, https://10.1080/10714420903558720

Packer, J. &Oswald, K.F. (2010) From Windscreen to Widescreen: Screening Technologies and Mobile Communication, *The Communication Review*, 13:4, 309-339, http://dx.doi.org/10.1080/10714421.2010.525478

Parikka, J. (2010). *Insect media: An archaeology of animals and technology* (Vol. 11). University of Minnesota.

Paris, B. S. (2021). Time constructs: Design ideology and a future internet. *Time & Society*, *30*(1), 126-149.

Parvin, N. & Pollock, A. (2020) Unintended by Design: On the Political Uses of "Unintended Consequences", *Engaging Science, Technology, and Society* 6 (2020), 320-327, https://10.17351/ests2020.497

Pasquinelli, M. (2017) Machines that Morph Logic. *Glass Bead*. https://www.glass-bead.org/article/machines-that-morph-logic/

Pasquinelli, M and Joler, V (2020) The Nooscope Manifested: Artificial Intelligence as Instrument of Knowledge Extractivism, KIM HfG Karlsruhe and Share Lab. https://nooscope.ai

Pias, C. 2011. On the epistemology of computer simulation. *Zeitschrift für Medien und Kulturforschung*.2011(1), 29-54.

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science*, *14*(3), pp. 399-441.

Poovey, M. (2003). Can Numbers Ensure Honesty? Unrealistic Expectations and the U.S. Accounting Scandal. Public lecture, International Congress of Mathematicians, Beijing. August 22, 2002.

Porter,T.M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life.* Princeton University.

Powles, J and Nissenbaum H (2018, December 7). The seductive diversion of 'solving' bias in artificial intelligence. *Medium*. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

Radiolab (2017 September 26) Driverless Dilemma. Podcast. http://www.radiolab.org/story/driverless-dilemma/

Rahwan, I (2016) The Social Dilemma of Driverless Cars. TEDx Cambridge. https://www.youtube.com/watch?v=nhCh1pBsS80

Rahwan, I. (2018) Society-in-the-loop: programming the algorithmic social contract. *Ethics Inf Technol* (2018) 20:5–14, https://10.1007/s10676-017-9430-8

Rajan, S. C. (2006). Automobility and the Liberal Disposition. *The Sociological Review*, *54*(1_suppl), 113–129. https://doi.org/10.1111/j.1467-954X.2006.00640.x

Rana, P. (2021, April 26). Lyft Agrees to Sell Autonomous-Driving Unit to Toyota for $550 Million. *Wall Street Journal*. https://www.wsj.com/articles/lyft-agrees-to-sell-autonomous-driving-unit-to-toyota-for-550-million-11619467500

Reed, I.A. (2013) Power: Relational, Discursive, and Performative Dimensions. *Sociological Theory* 31(3) 193–218, https://10.1177/0735275113501792

Désséguier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society,* 7(2), 2053951720942541. https://doi.org10.1177/2053951720942541

Ribes, D., Hoffman, A. S., Slota, S. C., & Bowker, G. C. (2019). The logic of domains. *Social Studies of Science.* https://doi.org/10.1177/0306312719849709

Roberge, J., Senneville, M., & Morin, K. (2020). How to translate artificial intelligence? Myths and justifications in public discourse. *Big Data & Society*, 7(1), https://doi.org/10.1177/2053951720919968

Roff, H. (2018, December 7)The folly of trolleys: Ethical challenges and autonomous vehicles. Brookings. https://brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/

Rosenberger, R. & Verbeek P-P (2015). Postphenomenological Investigations: Essays on Human–Technology Relations. Lexington Books.

Roth, L (2009) Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity. *Canadian Journal of Communication*, Vol 34 (2009) 111-136.

Roussi, A. (2020). Resisting the rise of facial recognition. *Nature*, 587(7834), 350–353. https://www.nature.com/articles/d41586-020-03188-2

Rouvroy, A. (2011). Technology, virtuality and utopia: Governmentality in an age of autonomic computing. In M Hildebrandt and A Rouvroy (Eds) *Law, human agency and autonomic comput ing: The philosophy of law meets the philosophy of technology* Routledge. pp. 119-140.

Schoettle B, Sivak M (2015). Potential impact of self-driving vehicles on household vehicle demand and usage (Report 2015-3). Ann Arbor: University of Michigan Transportation Research Insti tute.

Schuppli, S. (2014) Deadly Algorithms: Can Legal Codes hold Software accountable for Code that Kills? *Radical Philosophy* Issue 187 UK, pp 2-8.

Seaver, N. (2017)"Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems. *Big Data and Society,* 4, no. 2 https://doi.org/10.1177/2053951717738104

Seaver, N. (2018) What Should an Anthropology of Algorithms Do? *Cultural Anthropology* 33 (3): 375-85, https://doi.org/10.14506/ca33.3.04

Seilor, C. (2008) *Republic of drivers: A cultural history of automobility in America*. University of Chicago.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fair ness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59-68).

Sekula A (1986) The body and the archive. *October*. 1986 Dec 1;39:3-64.

Sharkey, N., & Suchman, L. (2013). Wishful mnemonics and autonomous killing machines. In

Proceedings of the AISB (Vol. 136, pp. 14-22).

Sheller M. (2004) Automotive Emotions: Feeling the Car. *Theory, Culture & Society* . 21(4/5): 221-242. https://doi.org/10.1177/0263276404046060

Shepardson D (2018) Tesla says crashed vehicle had been on autopilot prior to accident. *Reuters*. https://www.reuters.com/article/us-tesla-crash/tesla-says-crashed-vehicle-had-been-on-autopilot-prior-to-accident-idUSKBN1H7023

Sheridan, T.B., Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics* 1(1): 89–129. https://doi.org/10.1518/155723405783703082

Shilton, K. (2012) Values Levers: Building Ethics into Design, *Science, Technology, & Human Values* 38(3) 374-397 https://doi.org/10.1177/0162243912436985

Siegert, B. (2013). Cultural techniques: Or the end of the intellectual postwar era in German media theory. *Theory, Culture & Society*, 30, 48–65 p. 57

Siegert, B (2015) C*ultural Techniques Grids, Filters, Doors, and Other Articulations of the Real.* G Winthrop Young (Trans). Fordham University.

Simon, J. (2016). Distributed Epistemic Responsibility in a Hyperconnected Era in L Floridi (Ed) *The Onlife Manifesto: Being Human in a Hyperconnected Era*. Springer International,

Sivak, M., & Schoettle, B. (2015). Road Safety with Self-Driving Vehicles: General Limitations and Road Sharing with Conventional Vehicles. University of Michigan Transportation Research Institute, 2015. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/111735/103187.pdf?sequence=1&isAllowed=y

Smolensky, P. (2012) Symbolic functions from neural computation. *Philosophical Transactions of the Royal Society* A 370: 3543–3569.

Society of Automotive Engineers (2014) Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems J3016_20140 https://www.sae.org/standards/content/j3016_201401/

Society of Automotive Engineers (2018) Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles J3016_201806 https://www.sae.org/standards/content/j3016_201806/

Sprenger, F. (2015) *The politics of micro-decisions: Edward Snowden, net neutrality, and the architectures of the internet.* Meson

Star, S.L (1999) The Ethnography of Infrastructure, *American Behavioral Scientist* 1999; 43; 377

Stark, L. (2019). Facial recognition is the plutonium of AI. *XRDS* 25, 3 (Spring 2019), 50–55. https://doi.org/10.1145/3313129

Stayton, E. & Stilgoe, J. (2020). It's time to rethink levels of automation for self-driving vehicles. http://dx.doi.org/10.2139/ssrn.3579386

Stewart, J. (2017, May 10) Mapped: The top 263 companies racing toward autonomous cars. *Wired.* https://www.wired.com/2017/05/mapped-top-263-companies-racing-toward-autonomous-cars/

Stilgoe, J. (2017). Seeing like a Tesla. How can we anticipate self-driving worlds? *Glocalism: Journal Of Culture, Politics And Innovation.* 2017 (3):1-20. https://doi.org/10.12893/gjcpi.2017.3.2

Stilgoe, J. (2019) *Who's driving innovation? New technologies and the collaborative state.* Switzerland: Palgrave MacMillan. https://doi.org/10.1007/978-3-030-32320-2

Stop LAPD Spying Coalition and Free Radicals (2020) The algorithmic ecology: An abolitionist tool for organizing against algorithms, *Medium.* https://stoplapdspying.medium.com/the-algorithmic-ecology-an-abolitionist-tool-for-organizing-against-algorithms-14fcbd0e64d0

Suchman, L. (2015). Situational awareness: Deadly bioconvergence at the boundaries of bodies and machines. *Media Tropes,* 5(1), 1-24.

Suchman L. & Weber J. (2016) Human-machine autonomies. In: N. Bhuta, S. Beck, R. Geis, H-Y Liu, C. Kreis (Eds) *Autonomous weapons systems*. Cambridge University Press pp: 75-102.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, 11(3), 10-29.

Thrift, N. (2004). Driving in the City. *Theory, Culture & Society*, *21*(4–5), 41–59. https://doi.org/10.1177/0263276404046060

Thomson, J.J (1985). The Trolley Problem. *Yale Law Journal*. Vol. 94, No. 6 (May, 1985), pp. 1395-1415

Transmediale 2020 Adversarial Hacking in the Age of AI: Call for Proposals. https://2020.transmediale.de/content/adversarial-hacking-in-the-age-of-ai-call-for-proposals

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: key problems and solutions. *AI & Society*, 1-16.

Uncertain Commons Collective (2013) Speculate this! Duke University. https://wtf.tw/ref/uncertain_commons_speculate_this.pdf.

Urry. J (2004). The ‘system’ of automobility. *Theory, Culture & Society* (21): 25–39. https://doi.org/10.1177/0263276404046059

Vallor, S. & Bekey, G. (2017) AI and the Ethics of Self-learning Robots in P. Lin, K. Abney, R. Jenkins, (Eds) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford UP pp 338 - 355.

Vallor, S., Green, B., Raicu, I. (2018). Ethics in Technology Practice. The Markkula Center for Applied Ethics at Santa Clara University. https://www.scu.edu/ethics/

Vaughan, D (1997) *The Challenger Launch Decision: Risky Technology, Culture and Deviance at NASA.* University of Chicago.

Verbeek, P-P (2011) *Moralizing Technology: Understanding and designing the morality of things* . University of Chicago.

Verbeek P-P (2017) in van den Hoven J Miller, S., & Pogge, T. (Eds.). (2017). *Designing in ethics.* Cambridge University, Cambridge Core. https://doi.org/10.1017/9780511844317

Visser, E.J. de, Pak, R. & Shaw, T.H. (2018). From ‘automation’ to ‘autonomy’: The importance of trust repair in human–machine interaction. *Ergonomics*, 61(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

Wagner, B. (2018). Ethics as an escape from regulation: From ethics-washing to ethics-shopping. In: E. Bayamlioglu, I. Baraliuc, L. Janssens(Eds) *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen.* Amsterdam: Amsterdam University, pp. 84–89.

Wajcman, J. (2010). Feminist theories of technology. *Cambridge Journal of Economics*, 34(1), 143-152.

Wallach, W., and Allen, C. (2009). *Moral Machines: Teaching robots right from wrong*. Oxford University Press.

West, D.M., Travis, L.E. (1991) From society to landscape: Alternative metaphors for artificial intelligence. *AI Magazine* (12)2: 71 Association for the Advancement of Artificial Intelligence (AAAI)

Williams, R. (1974/2003) *Television: Technology and cultural form.* Routledge.

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive Inequity in Object Detection. http://arxiv.org/abs/1902.11097

Winfield, A.F., Michael, K., Pitt, J., Evers, V. (2019) Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems [Scanning the Issue], *Proceedings of the IEEE*, vol. 107, no. 3, pp. 509-517, March 2019, https://10.1109/JPROC.2019.2900622

Winograd, T. (1990/2000) Thinking machines: Can there be? Are we? In R Chrisley (Ed) *Artificial Intelligence: Critical Concepts*. Routledge, pp 432-458

World Economic Forum (2019, August 12) Look no hands: self-driving vehicles' public trust problem https://www.weforum.org/agenda/2019/08/self-driving-vehicles-public-trust/

Worstall T (2014, June 18). When should your driverless car from Google be allowed to kill you? *Forbes*. https://www.forbes.com/sites/timworstall/2014/06/18/when-should-your-driverless-car-from-google-be-allowed-to-kill-you/?sh=4ba236c0fa5b

Yurtsever E, Lambert J, Carballo A, Takeda K (2020) A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access* (8): 58443-58469, https://ieeexplore.ieee.org/document/9046805

Zer-Aviv, M. (2016, January 8) If everything is a network, nothing is a network. *Visualising Informa tion for Advocacy Blog* 8 January 2016. https://visualisingadvocacy.org/node/739.html

Zhang, L. (2017) Behind the Absurd Popularity of Trolley Problem Memes. *HuffPost*. https://huffpost.com/entry/behind-the-absurd-popular_b_10247650

Zhang, Q., Esterwood. C., Yang, X.J., Robert, Jr L.P. (2019) An Automated Vehicle (AV) like Me? The Impact of Personality Similarities and Differences between Humans and AVs. http://arxiv.org/abs/1909.11766.

Ziewitz, M. (2017). A not quite random walk: Experimenting with the ethnomethods of the algorithm. *Big Data & Society*, 4(2), 2053951717738105. https://doi.org/10.1177/2053951717738105

Zuboff, S (2015) Big Other: Surveillance Capitalism and the Prospects of an Information Civilization *Journal of Information Technology* (2015) 30, 75–89, https://doi:10.1057/jit.2015.5

APPENDICES

Appendix 1 Schedule of research activities 2016-2020

Date	Research Activities
June 2016	Visit to BMW factory, Leipzig, Gemany
November 2016	Workshop at Here.com in Berlin, Here.com is a technology company that makes maps for driverless cars.
October 2017	Interviews: University of Michigan, Ann Arbor,
January 2018	AI and Ethics Conference, New Orleans, Louisiana, United States
March 2018	Tesla test drive, Philadelphia, PA, United States.
June -July 2018	Interviews: Silicon Valley, California
October 2018	Interviews: Cambridge, Masschusetts
November 2018	Digital Research Salon: Daimler AG. Berlin.
May and June 2020	Online interviews

Appendix 2: List of interviewees with dates

No.	Name of interviewee	Affiliation and role	Interview location	Date of interview	Pseudonym	Consent details
1	Brandon Schoettle	University of Michigan Transport Research Institute; Human Factors Specialist	Ann Arbor, Michigan, United States	27/10/2017	Jason	Given via email to use real name.
2	Andrew Selbst	Data and Society Fellow; Lawyer; Postdoc	Online	23/03/18	Not cited directly in text	Given verbally and in writing
3	Jim McPherson	Lawyer	San Francisco, California, United States	27/06/18	Not cited directly in text	Given verbally to use real name
4	Jeff Helzner	Mathematician & philosopher; decision scientist, AIG Insurance	New York City, New York, United States	31/10/17	Not cited directly in text	Given verbally to use real name
5	Name requested to be redacted	UX Designer	Mountain View, California, United States	28/06/18	Jon	**Consent not given to use real name**
6	Sven Beiker	Stanford University affiliated with CARS	Online	10/07/18	Sven Beiker	Given verbally to use real name
7	Cansu Canca	AI Ethics Lab, Harvard University	Online	18/02/19	Cansu Canca	Given verbally to use real name
8	Pak-Hang Wong	Hamburg University, Dept of Informatics, Ethics in IT research group	Via telephone	01/02/19	Pak-Hang Wong	Given verbally to use real name

No.	Name of interviewee	Affiliation and role	Interview location	Date of interview	Pseudonym	Consent details
9	Tuhina Raman	Medical doctor; personal friend and Tesla owner	Philadelphia, Pennsylvania, United States	02/03/2018	Tuhina	Given verbally to use real name
10	Johannes Helmreich	Philosopher, Apple University	Apple HQ, California, United States	24/06/ 2018	Johannes Helmreich	Given verbally to use real name
11	Patrick Lin	Professor, California Polytechnic, San Luis Obispo, California, United States	Via email	2018	Patrick Lin	Given verbally to use real name
12	Name requested to be redacted	Federal Ministry of Transport and Digital Infrastructure,	Berlin, Germany	09/08/18	Not cited directly in text	**Consent not given to use real name**
13	Noah Goodall	Virginia Transport Research Institute, Charlottesville, VA	Skype	30/9/19	Noah Goodall	Given verbally to use real name
14	Francien Dechesne	Leiden University	Online	24/4/20	Francien Dechesne	Given verbally to use real name
15	Zeerak Waseem	Sheffield University, UK	Online	24/4/20	Not cited directly in text	Given verbally to use real name
16	Jason Miller	University of Ottawa	Online	12/6/20	Jason Miller	Given verbally to use real name
17	Alan Winfield	Bristol University	Online	15/5/20	Not cited by name	Given verbally to use real name

No.	Name of interviewee	Affiliation and role	Interview location	Date of interview	Pseudonym	Consent details
18	Pradeep Gopi	Nissan Motors	Online	21/3/19	Not cited directly in text	Given verbally to use real name
19	Stefan Podgrabinski	Computer scientist, Ryerson University, Toronto, Canada. Retired	Online	17/8/20	Not cited directly in text	Given verbally to use real name
20	Judith Simon	Professor, Hamburg University, Informatics Dept	Hamburg University	18/5/18	Not cited directly in text	Given verbally to use real name

Appendix 3 Email exchange with Dr. Patrick Lin

Full email of email exchange with Patrick Lin; text accompanying screenshot from inbox indira.ganesh@stud.leuphana.de (below)

On 26/04/2018 21:50, Patrick Lin wrote:
> Quick reply, Maya, as I head out to my meetings and classes:
>
> A reporter recently asked me about the origins of the trolley problem for AV ethics. Here's what I said--feel free to use as needed:
>
> "Great question, since I actually had looked into this not long ago, too:
>
> To the best of my knowledge, Colin Allen and Wendell Wallach were the first (in print, at least) to suggest that the trolley problem could be an actual problem in robotics, but their example was with driverless trains. See chapter 1 of their 2008 book: https://www.amazon.com/Moral-Machines-Teaching-Robots-Right/dp/0199737975 (use "Look Inside" feature). By the way, it's still an excellent and seminal book on the topic; my recent Robot Ethics 2.0 book has several chapters about robot cars, but their Moral Machines book was really the first of its kind.
>
> In 2012, Gary Marcus wrote this short piece for The New Yorker that seems to riff off this idea (Gary knows Colin and Wendell, as do I), but Gary does connect something like the trolley problem (but not exactly) to autonomous cars: https://www.newyorker.com/news/news-desk/moral-machines Unfortunately, Gary didn't invoke the trolley problem directly or say "trolley", and the article contains no links to sources or inspirations. It's also not a 5 vs. 1 scenario. I believe I linked to Gary's article, or a related one that talked about it, in that Wired article you found, since I try to be careful about attribution, too.
>
> So, I suppose it's technically true that I'm the first to explicitly connect the trolley problem to self-driving cars, but I stand on the shoulders of giants and their prior work. In any event, this is a pretty obvious connection, at least in retrospect, that I think would have been made eventually by someone else, if not by us. And if you believe this writer from Time, his research seemed to show that I had "thought about the ethics of driverless cars more than anyone": http://time.com/2837472/driverless-cars-ethics-morality/ (I wouldn't say that myself, but I didn't know anyone else at the time focusing on this area, and so I can't think of anyone else who'd have a better claim to being first in this.)
>
> By the way, I had thought about different ways to run the trolley
> problem scenario (save 5 vs. kill 1) with robot cars, and this one
> seems to be the most plausible and least problematic—the human is

> initially in control of the car first, and the programmer must decide
> whether AI should intervene or not:
> https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-a
> d
> justable-ethics-settings/
>
> And here are other variations of the dilemma that press on different
> intuitions—they're not about 5 vs. 1 but still showcase the same kind
> of intractable tradeoff of values at their core:
> https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be
> - programmed-to-hit-you/ (animated short here:
> https://ed.ted.com/lessons/the-ethical-dilemma-of-self-driving-cars-p
> a
> trick-lin)
>
> Finally, I defend the trolley problem in this piece last year. While the dilemma might be exceedingly rare, it has happened before, and anyway it misses the point to insist that the dilemmas should be a real problem before thinking about it. Thought experiments are genetically related to science experiments; and science experiments are also "fake" in that they set up artificial conditions that don't obtain in the real world...yet no one complains about that: https://www.forbes.com/sites/patricklin/2017/04/03/robot-cars-and-fake-ethical-dilemmas/ ..."
>
>
> -----Original Message-----
> From: Maya Indira Ganesh [mailto:lg069809@stud.leuphana.de]
> Sent: Thursday, April 26, 2018 12:32 PM
> To: Patrick Lin <palin@calpoly.edu>
> Subject: Re: Request for an interview
>
> Dr Lin,
>
> It's really impressive that you remember! Yes, you did interrupt that
> conversation - and thank you ;-))
>
> I'd be happy to do a Skype call with your colleague when I'm in California, so please do put me in touch.
>
> And when you're done with Iceland, maybe I will get back in touch with more detailed questions, for now though, there is really just one:
>
> who came up with the Trolley Problem for AV ethics? Like, where did it really start, who first said 'this is what we should use as a framework'

> ? I think I have also read, and quote, a lot of what you've written about this, but am looking for the genealogy of the TP in the AV ethics context specifically. And why? Have there been other frameworks or suggestions for other kinds of thought experiments? (I'm following the CS and Math literature that loves developing solutions based on the TP).
> My dissertation starts with this question of why the Trolley / Helmet / Lin Problem, and picks up from there looking at other ways in which we test machine learning. And if there are some leads I should go look up, please do send them my way.
>
> Thank you so much, I appreciate your help with this.
> Best wishes, and enjoy Iceland, I hear it is really amazing.
>
> Best,
> Maya
>
>
> On 26/04/2018 21:02, Patrick Lin wrote:
>> Hi Maya -
>>
>> Good to hear from you. (I had interrupted your convo with Daniel
>> Estrada at AIES, was that right?)
>>
>> When you're in California, I'll be in Iceland for a project from mid-June to mid-July. But I can connect you to my colleagues here who also work closely with me on AV ethics. We're 4-6 hours by car from any of the cities you list, so a virtual meeting would likely be needed.
>>
>> My schedule is way overbooked until after the Iceland project, but I
>> might be able to field a couple questions by email; or perhaps our
>> previous publications/interviews can answer your q's:
>> http://ethics.calpoly.edu/robots.htm
>>
>> Best of luck with your dissertation!
>>
>> Patrick Lin, Ph.D.
>> Professor, Philosophy Dept.
>> Director, Ethics + Emerging Sciences Group California Polytechnic
>> State University
>> 1 Grand Avenue
>> San Luis Obispo, CA 93407
>> [e] palin@calpoly.edu
>> [w] http://ethics.calpoly.edu
>>
>>

>>
>> -----Original Message-----
>> From: Maya Indira Ganesh [mailto:lg069809@stud.leuphana.de]
>> Sent: Thursday, April 26, 2018 5:19 AM
>> To: Patrick Lin <palin@calpoly.edu>
>> Subject: Request for an interview
>>
>> Dear Dr. Lin,
>>
>> We met very briefly at AIES in New Orleans a few months ago where you gave a closing keynote. You probably don't remember, but I'm doing a PhD in Germany and do some work with automotive companies here.
>>
>> I am going to be in California for a few weeks from mid June- early July, and I'm writing to ask if you're free for an interview about your work, and autonomy and ethics in particular. I will be in Irvine, LA, SF, Palo Alto and San Jose. Will you be in any of these places? If this is not possible, then I hope it's okay if I just ask you some specific questions via email or a call?
>>
>> Thanks a lot,
>>
>> Maya Indira Ganesh
>> --
>> Doctoral candidate, Leuphana University die Fakültat
>> KulturWissenschaften/MECS
>> T: @mayameme
>> W: bodyofwork.in
>

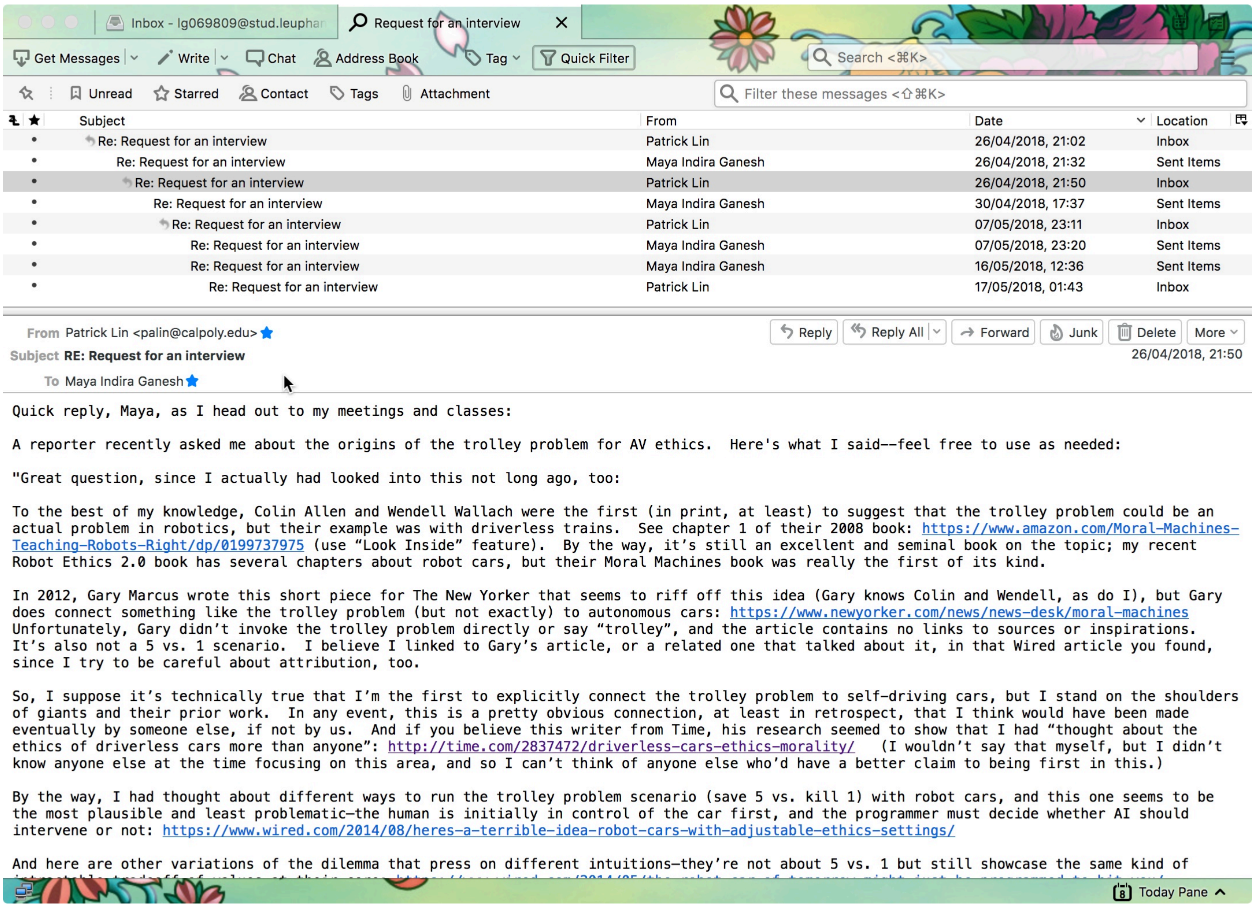
[1] https://www.amazon.com/Moral-Machines-Teaching-Robots-Right/dp/0199737975

[2] https://www.newyorker.com/news/news-desk/moral-machines

[3] http://time.com/2837472/driverless-cars-ethics-morality/

[4] https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/

[5] https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/



Subject	From	Date	Location
Re: Request for an interview	Patrick Lin	26/04/2018, 21:02	Inbox
Re: Request for an interview	Maya Indira Ganesh	26/04/2018, 21:32	Sent Items
Re: Request for an interview	Patrick Lin	26/04/2018, 21:50	Inbox
Re: Request for an interview	Maya Indira Ganesh	30/04/2018, 17:37	Sent Items
Re: Request for an interview	Patrick Lin	07/05/2018, 23:11	Inbox
Re: Request for an interview	Maya Indira Ganesh	07/05/2018, 23:20	Sent Items
Re: Request for an interview	Maya Indira Ganesh	16/05/2018, 12:36	Sent Items
Re: Request for an interview	Patrick Lin	17/05/2018, 01:43	Inbox

From Patrick Lin <palin@calpoly.edu>
Subject **RE: Request for an interview**
To Maya Indira Ganesh
26/04/2018, 21:50

Quick reply, Maya, as I head out to my meetings and classes:

A reporter recently asked me about the origins of the trolley problem for AV ethics. Here's what I said--feel free to use as needed:

"Great question, since I actually had looked into this not long ago, too:

To the best of my knowledge, Colin Allen and Wendell Wallach were the first (in print, at least) to suggest that the trolley problem could be an actual problem in robotics, but their example was with driverless trains. See chapter 1 of their 2008 book: https://www.amazon.com/Moral-Machines-Teaching-Robots-Right/dp/0199737975 (use "Look Inside" feature). By the way, it's still an excellent and seminal book on the topic; my recent Robot Ethics 2.0 book has several chapters about robot cars, but their Moral Machines book was really the first of its kind.

In 2012, Gary Marcus wrote this short piece for The New Yorker that seems to riff off this idea (Gary knows Colin and Wendell, as do I), but Gary does connect something like the trolley problem (but not exactly) to autonomous cars: https://www.newyorker.com/news/news-desk/moral-machines Unfortunately, Gary didn't invoke the trolley problem directly or say "trolley", and the article contains no links to sources or inspirations. It's also not a 5 vs. 1 scenario. I believe I linked to Gary's article, or a related one that talked about it, in that Wired article you found, since I try to be careful about attribution, too.

So, I suppose it's technically true that I'm the first to explicitly connect the trolley problem to self-driving cars, but I stand on the shoulders of giants and their prior work. In any event, this is a pretty obvious connection, at least in retrospect, that I think would have been made eventually by someone else, if not by us. And if you believe this writer from Time, his research seemed to show that I had "thought about the ethics of driverless cars more than anyone": http://time.com/2837472/driverless-cars-ethics-morality/   (I wouldn't say that myself, but I didn't know anyone else at the time focusing on this area, and so I can't think of anyone else who'd have a better claim to being first in this.)

By the way, I had thought about different ways to run the trolley problem scenario (save 5 vs. kill 1) with robot cars, and this one seems to be the most plausible and least problematic—the human is initially in control of the car first, and the programmer must decide whether AI should intervene or not: https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/

And here are other variations of the dilemma that press on different intuitions—they're not about 5 vs. 1 but still showcase the same kind of

[6] https://ed.ted.com/lessons/the-ethical-dilemma-of-self-driving-cars-patrick-lin

[7] https://www.forbes.com/sites/patricklin/2017/04/03/robot-cars-and-fake-ethical-dilemmas/

Submitted on: June 5, 2021

Thesis defence on: March 17, 2022

Main supervisor and reviewer: Apl. Prof. Dr. Rer. Nat. Martin Warnke.

Second reviewer: Professor Dr. Nishant Shah

Third reviewer: Professor Dr. Claus Pias.