Machine learning, social learning and the governance of self-driving cars

Jack Stilgoe

Department of Science and Technology Studies, University College London, UK

Abstract

Self-driving cars, a quintessentially 'smart' technology, are not born smart. The algorithms that control their movements are learning as the technology emerges. Self-driving cars represent a high-stakes test of the powers of machine learning, as well as a test case for social learning in technology governance. Society is learning about the technology while the technology learns about society. Understanding and governing the politics of this technology means asking 'Who is learning, what are they learning and how are they learning?' Focusing on the successes and failures of social learning around the much-publicized crash of a Tesla Model S in 2016, I argue that trajectories and rhetorics of machine learning in transport pose a substantial governance challenge. 'Self-driving' or 'autonomous' cars are misnamed. As with other technologies, they are shaped by assumptions about social needs, solvable problems, and economic opportunities. Governing these technologies in the public interest means improving social learning by constructively engaging with the contingencies of machine learning.

Keywords

self-driving cars, autonomous vehicles, responsible innovation, machine learning, social learning

Correspondence

Jack Stilgoe, Department of Science and Technology Studies, University College London, Gower Street, London WC1E 6BT, UK
Email: j.stilgoe@ucl.ac.uk

Introduction

In late 2016, the car company Tesla announced that the new generation of its Model S would include 'full self-driving hardware' (Tesla, 2016a). This would be a technology capable of realizing a long-held dream of automotive autonomy, with the requisite sensors and processing power to drive 'all the way from LA to New York' without

human input by the end of 2017, according to Elon Musk, the company's CEO (quoted in Etherington, 2016). However, this quintessentially 'smart' technology has not been born smart. The brain of this self-driving-car-in-the-making is still not fully formed. The algorithms that its creators hope will allow it to soon handle any eventuality are being continually updated with new data. It is a car that is learning to drive.

My curiosity about this particular model has a morbid side. In May 2016, a Tesla Model S was involved in what could be considered the world's first self-driving car fatality. In the middle of a sunny afternoon, on a divided highway near Williston, Florida, Joshua Brown, an early adopter and Tesla enthusiast, died at the wheel of his car. The car failed to see a white truck that was crossing his path. While in 'Autopilot' mode, Brown's car hit the trailer at 74 mph (in a 65 mph zone). The crash only came to light in late June 2016, when Tesla (2016b) published a blog post, headlined 'A tragic loss', that described Autopilot as being 'in a public beta phase'.

The self-driving car is a technology that is already with us as well as a work-in-progress, laden with promise for what it might become. It is an important vehicle for the development and application of machine learning. With machine learning, as with other emerging technologies, society has not yet worked out the terms of responsibility, the distribution of liability, the thresholds of acceptable safety or the lines dividing recklessness from negligence. The emergence of self-driving cars is therefore also a test of social learning, which can be defined as the way in which society and its institutions make sense of novelty.

In this article, I analyse the public debate about self-driving car innovation, considering competing definitions of problems, solutions and concerns. I begin by considering the emerging politics of machine learning and the relative

problematization of algorithmic outcomes and processes. I describe how the application of deep learning – a mode of artificial intelligence in which software learns its own rules to solve tasks – has revived the possibility of self-driving cars, making the engineering challenge, as well as social problems from road safety to sustainability, appear 'solvable'. In this light, the terminology of 'self-driving cars' and 'autonomous vehicles' and the promises surrounding these terms appear misleading. New companies perform their versions of idealized self-driving futures while incumbent car manufacturers work out how to respond, each seeking to redefine technological novelty in their interests. When these nascent technologies go wrong, the scale of the gap between promise and reality starts to become clear.

My examination of the Tesla crash and its aftermath draws on official investigations as well as informal online exchanges among users engaged in a process of haphazard social learning. An analysis of what was learnt and what was ignored after the crash allows for the anticipation of governance challenges ahead. Innovators' insistence that the answer lies in continued autonomy and algorithmic enhancement leads to a rejection of new forms of governance and represents a substantial privatization of learning. This jeopardizes both public trust and the long-term potential of technologies that could be hugely beneficial. Existing governance approaches, including the responses from US regulators to the Tesla crash, suggest some misguided assumptions, but also some cause for optimism. In my conclusion, I point to some governance options that seek to prioritize social learning, focusing in particular on the sharing of data.

Being 'in beta'

The suggestion that technologies are social experiments has become a commonplace critique. If experiments are understood, following Rheinberger (1997), as systems for the organized production of surprises, then an approach to technology as a social experiment should prompt us to ask what sorts of uncertainties are considered pertinent by different actors and how they respond (Gross, 2010; Stilgoe, 2016). In other words, we should pay close attention to who learns what and how.

Technological accidents can wrest control of the social experiment away from the technologists, laying bare the rules and assumptions that shape black boxes and exposing the uncertainties that are so easy to ignore when things work well (Wynne, 1988). Except in cases of total cover-up, accidents can force public reframing and institutional reflection.[1]

However, while accidents may be an instructive form of 'informal technology assessment' (Rip, 1986), the interpretation of their lessons is, as with all learning, socially filtered. It is hard to teach institutions things that they do not want to learn. The vagaries and blind spots of social learning have been well described by writers in the social sciences (see Parson and Clarke, 1995) and elsewhere. As TS Eliot has the doomed archbishop Thomas Becket say in 'Murder in the cathedral',

> We do not know very much of the future
> Except that from generation to generation
> The same things happen again and again
> Men learn little from others' experience.

As I describe below, the recognition of such human failings has helped spur efforts to rationalize learning in machines. This exacerbates the tendency to blame the things that come to be labelled 'unintended consequences' (Jasanoff, 2016), on 'human

[1] While social scientists often talk about technological 'accidents', it is worth noting that others, including crash investigators, shun the term because it suggests that nobody is to blame.

error' (Shorrock, 2013) or 'user error' (Wynne, 1988)). In systems where machines and humans interact, users often become the 'moral crumple zone' (Elish, 2016) for technological failure.

Social scientists' studies of technological accidents point to a more constructive alternative view. Reframing errors as 'system-induced' (Wiener 1977) rather than merely a function of human frailty allows for genuine learning. Perrow (1984) emphasizes that humans should be regarded as a resource, not a problem, for technological safety. His counterexample to what he regards as inherently dangerous nuclear power stations is air travel. Despite the mindboggling socio-technical complexity of aeroplanes, Perrow (1984: 127) argues that 'no other high-risk system is so well-positioned to effectively pursue safety as a goal', because of an emphasis on social learning.

Innovation is inherently unruly (Wynne, 1988). Rather than just following a set of pre-established social rules, such as those to do with safety, technology is a set of practices that generate new rules. Questions of safety are therefore hard to pin down, because they tend to morph into questions such as 'safe enough for what?' Discussions that start with 'risk' end up being about the purposes of a particular technology and its alternatives (Rayner and Cantor, 1987).

Work in science and technology studies (STS) has gone some way towards explaining the social constitution of technologies such as nuclear power, genetically modified foods, nanotechnology and geoengineering (Kearnes et al. 2006; Szerszynski et al. 2013; Winner 1980). But there has been less attention to the dynamics through which such constitutions emerge and alternative modes of governance might suggest improvements. Rather than seeing technologies as constitutionally static, we can instead imagine them as processes of learning.

There are multiple ways in which social learning is relevant to the governance of emerging technologies. The first, originating in educational psychology (e.g., Bandura, 1988), starts from the observation that people can and do learn in groups – with and from others. This literature sees social learning as the product of interactions between the cognitive activities of learners, the social behaviours they see modelled and the wider environment. The normative possibilities of social learning lie in understanding the multidimensionality of issues and experimenting with new approaches to achieve social change (e.g., Friedman and Abonyi, 1976; Reed et al., 2010). Systems that are good at social learning are seen as more resilient, interpreting shocks and crises as opportunities (Berkes and Turner, 2006). The substantial literature on social learning for sustainability blends two approaches to social learning: the first focussing on how people learn socially, the second asking how societies learn (Parson and Clark, 1995).

Theories of social learning in governance emphasize the design of systems and institutions that learn and improve over time (Fischer, 2000; Rayner, 2004; Wynne, 1992). These frameworks have an analytical basis in theories of politics as a form of learning (Hall, 1993), with policymaking seen not just as the playing out of stakeholder interests, but as 'puzzling together' (Hoppe, 2011), in the tradition of Dewey (1916).

The conception of governance as social learning is particularly pertinent to new technologies, which typically emerge in what Hajer (2003) calls an 'institutional void', surrounded by uncertainties not just about the effects of technology (Collingridge, 1980) but also the object of governance itself (Owen, 2014). Conventional technology assessment fixates on the products of innovation, and in particular its hazards. Frameworks of responsible innovation aim to also engage

upstream with the processes and purposes of innovation, recognizing that, in addition to questions of risk, public concerns also relate to how and why innovation happens (Stilgoe et al., 2013). Recent efforts to expand practices of public deliberation on new technologies can be seen as a form of learning with and about the social context of new technologies, with the additional aim of galvanizing collective action (Fischer, 2000; Webler et al., 1995).

Many of the most profound limits to responsiveness in innovation relate to inadequacies in social learning. The historians' lament that those ignorant of history are doomed to repeat it is particularly apposite because innovation is imagined as a project of novelty. As Rayner (2004) has argued, presumptions of novelty act against learning. It becomes easy for innovators to argue that 'this time it's different' and that past pathologies of technological development, such as widening inequality, ethical dilemmas, unjustified hype, novel risks and other unintended consequences will not be repeated. Institutions often lack the reflexive capacity to take on board and respond either to the views of others (Wynne, 1993) or to early warnings of danger (EEA, 2001). It is easier for institutions to imagine deficits of public knowledge, public trust or regulation (Rayner, 2004) than to question the uncertainties and contingencies of their own commitments. At the heart of Paquet's (2005, p. 315) critique of 'solutionism', in which issues are 'interpreted as puzzles to which there is a solution rather than problems to which there may be a response', is a concern that '[t]here is no place for social learning… the development of rough and ready arrangements around which collaboration and negotiations might be built'.

Given the uncertainties of innovation and an institutionalized tendency towards hubris (Jasanoff, 2004), a focus on social learning offers a way to understand and democratize the means and ends of the social experiment of technological change.

Both senses of social learning – how people learn socially and how societies learn – apply. The latter is more obviously relevant to the governance of new technologies. But the former becomes particularly interesting when we consider the 'social' within machine learning.

The sociology of machine learning

The recent and rapid resurgence of machine learning poses particular challenges for governance based on social learning. Machine learning systems such as those in control of nascent self-driving cars offer a literal example of Wynne's (1988) characterization of technology as rule-making as well as rule-following behaviour. The algorithmic architecture of self-driving car programming begins with 'if-then-else' rules, which define actions under certain conditions. The car senses and classifies the world around it before probabilistically making choices based upon what it has learnt.

However, driving is too complicated to fit a predict-and-provide approach. As engineers have come to recognize the breadth of possible situations that might need to be defined by formal rules and then engraved in algorithms, they have turned to machine learning using deep neural networks. Here, the aim of the game is to work out the rules. The machine is trained by extracting patterns from vast datasets, a process referred to as 'rule learning', 'rule induction' or 'rule extraction'. For engineers, the gains in efficiency are clear. Lipson and Kurman (2016: 8) claim: 'The fact that deep-learning software 'learns' by looking at the world gives it [a] major advantage: it's not rule bound.' Sebastian Thrun, formerly Google's self-driving car lead, goes further, arguing that '[t]he data can make better rules' (Vanderbilt, 2012). One engineering analysis (Moore and Lu, 2011) argues that the rules written by

machines from available data would be less arbitrary than the legal 'rules of the road' (see also Both, 2016).

For governance, this unruliness is a cause for concern. The tension between rule-following and rule-making in machine learning has started to be explored by literature on the politics of algorithms (Burrell, 2016). As it has become clear that citizens' lives are increasingly shaped and cajoled by the influence of algorithms and artificial intelligence, social scientists, legal scholars and philosophers have developed critical approaches (Mittelstadt et al., 2016). Many of these focus on the question of opacity. Algorithms are 'black boxes' (Pasquale, 2015) but, unlike many others, they may be extremely hard, if not impossible, to prise open with sociological or historical tools.

Burrell (2016) sees three modes of algorithmic impenetrability. First, algorithms are a source of competitive advantage and therefore likely to be proprietary. For this reason, access to code and the data that enables its learning will, according to Pasquale (2015), become a growing point of contention between companies and others seeking to understand their actions. Secondly, as algorithms become more specialized, more complex and composed by multiple authors with different perspectives, even their creators may no longer be able to understand them (Burell, 2016). As computer entrepreneur-turned-academic Kevin Slavin puts it: 'We're writing things we can no longer read' (quoted in Neyland, 2016). For those on the outside seeking to hold algorithms to account, the challenge of legibility is even greater. The third way in which algorithms become obscure is in their application, creating complexity as they make use of large datasets. Understanding and dealing with this means going beyond conventional calls for algorithmic transparency to scrutinize real-world uses (Annany and Crawford, 2016; Burell, 2016).

Those seeking to govern algorithms have much to learn from past examples of emerging technologies. Over the past three decades, a fault line has emerged in the debate over agricultural biotechnology regulation. One approach, favoured by US regulators, seeks to evaluate an innovation's products – traits, risks, benefits and other outcomes. The assumption is that technologies, like sausages and laws, should be judged on the quality of the product, not the process that created it. The other approach, more common in Europe, focuses on the novelty of the process (Jasanoff, 1995; Kuzma, 2016). Those advocating a more precautionary approach to regulation have argued that a product-based approach, by emphasizing the 'substantial equivalence' between an innovation and its predecessors, overlooks uncertainties that may emerge as problematic (Millstone et al., 1999). The presumption of substantial equivalence is also likely to generate controversies about labelling when groups disagree about the nature and implications of technological novelty.

For machine learning, there are process considerations relating to the often-opaque generation of outcomes by digital systems, as well as to the processes through which new technologies are created, such as the gathering of data and the assumptions made within algorithms. The identification of tasks and modes of machine learning (for example, reinforcement learning, which adopts a trial, error and reward approach to optimization, described in one leading paper (Mnih et al., 2015) as 'deeply rooted in psychological and neuroscientific perspectives on animal behaviour') is inescapably social.

Deep learning systems are seen by their creators as means of engaging with an uncertain world that is impossible to capture with a set of formal rules. However, in developing rules, such systems may create new social uncertainties. In gaining the ability to recognize and make decisions about unfamiliar information, they lose the

ability to account for their actions. Algorithmic outcomes may therefore be inscrutable, their decisions being the computing equivalent of a hunch. The *what* is prioritized over the *why* or the *how*. This has led some to call for a 'right to explanation' (Goodman and Flaxman, 2016) in cases where the 'production of prediction' (Mackenzie, 2015) has profound consequences for people's lives.[1] For the machine learning community, the challenge is seen as one of 'interpretability' (Vellido et al., 2012). For governance, an additional question is whether standards of interpretability might differ between companies, regulators, users, and citizens (Edwards and Veale, 2017). An explanation that is considered adequate by engineers may not satisfy a sceptical NGO, for example. And, if datasets are large and multidimensional, simple explanations may be impossible. The separation of outcomes from processes if interpretability is trumped by efficiency forms a substantial barrier to social learning.

The politics of interpretability have already revealed themselves with the type of image recognition techniques at work in self-driving car systems. When Google's algorithms were misidentifying images of people as dogs and, with predictably greater controversy, black people as gorillas, the company's own engineers could not pinpoint where the problem lay (Annany and Crawford, 2016). The machines were learning in ways that their creators could not understand. Innovators express the surprise they feel when they see one of their creations learning something for itself. One Netflix executive described the insights generated by their algorithms being like a 'ghost in the machine' (Finn, 2017). After years of unsupervised machine learning, researchers are only now beginning to understand how a deep neural network goes about identifying an image (Lipson and Kurman, 2016; Nvidia, 2017).

One early governance concern with self-driving cars has related to the ethics of algorithms making life or death decisions in the event of crashes and the responsibilities that this might place upon designers and manufacturers (Bonnefon et al., 2016; Crawford and Calo, 2016). Deep learning could outsource such decisions to the machine and make it impossible to account for them. As such, it could become a form of codified irresponsibility, a convenient way for companies to avoid both liability and learning.

For robust governance, systems need to be understood and scrutinized in multiple ways from multiple perspectives (Annany and Crawford, 2016). Allowing learning to be defined with reference only to a particular algorithm's task restricts the potential for contestation according to different purposes. The imagined purposes of technologies and their justification are tightly linked with the definition of particular problems (Morozov, 2013). As an optimizing strategy, machine learning demands the tight identification of tasks. Problems need to be well defined such that they can be 'solved' and the efficiency of the solution evaluated. For cars, this means considering how what engineers call the 'driving task' has been articulated, cut into various sub-tasks and reconstructed as 'solvable'.

Self-driving as a 'solvable problem'

Much of the early history of automated cars, downplayed in more recent rhetoric, involved plans for the integrated innovation of vehicles and highways. From the 1950s up to the end of the century, it was assumed that, in order to get self-driving cars to work, they would need to communicate with similarly intelligent highways (Wetmore, 2003).[2]

At the same time, car manufacturers were adding automated systems such as cruise control and airbags to improve safety and comfort (Vinsel, forthcoming). Adaptive cruise control, lane-departure warnings, collision warnings, rear view cameras and other aids came to be classified as Advanced Driver Assistance Systems. Tesla's Autopilot, installed in almost 100,000 cars as of January 2017, might be seen as a straightforward next step. However, a narrative of incrementalism would overlook the role of machine learning, which is seen by self-driving enthusiasts as the tool that will upend conventional transport.

Newer companies with self-driving car investments such as Tesla, Google, Delphi, Nvidia, Mobileye, Nutonomy, Otto and Cruise Automation, most of whom have emerged from the vicinity of Silicon Valley, would trace the histories of their transport work back to the 'Grand Challenge' robot driving competitions staged by the Defense Advanced Research Projects Agency (DARPA). At the first event, in the Mojave Desert in 2004, the best of the cars managed only 7 miles of the 150-mile course. The next year, five vehicles completed the course. The 2007 version was staged in an urban environment. The six cars that finished were seen as an announcement to the world that innovation was happening apace. The 2005 and 2007 events brought attention to engineers from Stanford and Carnegie Mellon universities, who would go on to populate some of the leading private-sector self-driving car teams.

The advancing robotics revealed by the DARPA challenge benefitted from and spurred on advances in computing. Hardware originally designed for videogame graphics was found to be extremely good at the rapid parallel processing required for machine learning. From 2008 onwards, these Graphics Processing Units (GPUs) enabled dramatic improvements in complex computer models known as deep neural

networks. As of 2016, these neural networks' greatest feats have been in digital image recognition, where it is claimed their abilities have surpassed humans' (Krizhevsky et al., 2012), and voice recognition (Lecun et al., 2015). Training these networks became possible with the accumulation of massive, labelled datasets. For self-driving cars, the gathering of data from both mapping projects and real-world driving has become central to the development of vehicles' processing capabilities.

Autonomous driving is now imagined as possible despite the unpredictability of the open road. Innovators insist that self-driving cars are now a 'solvable problem' that, according to Musk, speaking in 2015, is 'almost … a solved problem. We know exactly what to do and we will be there in a few years' (quoted in Scoltock, 2015). The CEO of Nvidia, a pioneer of GPUs for machine learning and a Tesla hardware partner, told a Consumer Electronics Show audience that 'We can realize this vision [of self-driving cars] right now'. On the screen behind him were the words 'AI [artificial intelligence] is the solution to self-driving' (Recode, 2017).

The gamification of driving

The reduction of the driving task to a solvable machine learning problem follows a set of high-profile achievements in artificial intelligence. AI has long held a maxim known as the Moravec paradox, in honour of the robotics pioneer Hans Moravec: '[T]he hard problems are easy and the easy problems are hard' (Pinker, 1995: 192). For early AI, things that humans could do without thinking, such as walking, seemed almost impossible to emulate, while cognitively hard tasks like chess proved remarkably easy for computers.

Driving does not require a genius. Indeed, proponents of self-driving cars are fond of pointing out the intellectual deficits of human drivers. The quest is therefore to tear apart Moravec's paradox, to turn a task that many humans find easy into one

that computers can solve, with enough data and processing power. In this way, driving has been turned into a form of machine learning game.

Rather than attempting to replicate the thought processes of human drivers, the approach has been one of brute force, with the key resource being data. With the turn to 'big data', machine learning has seen what Lipson and Kurman call 'the transition from data-poor "clever" algorithms to data-rich "simple" ones' (Lipson and Kurman, 2015: 219). The complexity of driving data is far greater than in chess, but the sources of learning – billions of hours of real and simulated driving – are greater too. As I explain below, self-driving cars learn as fleets rather than as individual robots.

One attempt (though flawed (Borup et al., 2006)) to capture and challenge the levels of promise surrounding emerging technologies has been developed by Gartner, a consultancy firm. Their 2015 'hype cycle' had autonomous vehicles (AVs) at its apex, on the cusp of the 'trough of disillusionment'. AVs were, according to Gartner, 5-10 years away. The 2016 version had AVs beginning their slide towards reality, and now being more than ten years away. A cynic might highlight the contradiction: that the technology's arrival is getting ever more distant. Wetmore (2003) points out that self-driving vehicles were 'only 20 years away' for more than 60 years in the 20th Century. But the more important concern is that there is no clear, uncontested line demarcating the presence of a technology in society. Chris Urmson, then leader of Google's self-driving car project, said in March 2016:

> How quickly can we get this into people's hands? If you read the papers, you see maybe it's three years, maybe it's thirty years. And I am here to tell you that honestly, it's a bit of both. (in Gomes, 2016).

Companies with commitments to self-driving cars and investment horizons of only a few years have little choice but to exaggerate the speed and downplay the friction of technological change. They must navigate what Rayner (2004) calls the 'novelty

trap', advertising a technology as ground-breaking and transformative while seeking to persuade regulators that it is no more than an incremental step beyond existing approaches. The perfect self-driving futures currently on offer are further away than their proponents imagine. And yet, in the formal and informal experiments currently underway, self-driving cars can be said to be on the streets already. With Musk's encouragement, Tesla owners and journalists have described Autopilot's benefits as vast (Musk, 2016a; xkwizit, 2016; xrayvsn, 2016). One technology reporter wrote that Tesla's 'self-steering was suddenly, overnight, via a software update, a giant leap toward full autonomy' (Bradley, 2016). At the same time, the company's response to regulatory scrutiny is to suggest that their innovations are mere 'baby steps' (Frankel, 2016). The rhetorical management of the definition of increments and revolutions in innovation is a central project of contemporary innovation (Borup et al., 2006). The gap between baby steps and giant leaps, between concept cars and transport systems, is filled with promise and speculation.

The global unevenness of road surfaces, built environments, other road users, weather conditions, regulatory regimes and driving cultures means that self-driving cars can, without paradox, be both complicated and straightforward, implausible and probable, distant and just-around-the-corner. The development of artificial intelligence is not a line from easy to hard. If machine learning is 'the part of artificial intelligence that actually works' (Kosner, 2013), we must question the contexts in which it can be said to 'work'.[3] Despite the existence of a car with '[f]ull self-driving hardware' (Tesla, 2016a), it may still be impossibly hard to chart a route through the messy transitions, mixtures and missteps that may be encountered en route to a self-driving world.

In addition to systems such as Tesla's Autopilot, the public performance of inevitability has taken the form of a number of high-profile tests. Unlike the DARPA challenges, which were genuine public experiments, subsequent efforts had closer control of their uncertainties (see Collins, 1988). In his testimony to a US Senate Committee, Glen de Vos (2016) from Delphi described how,

> 'in April of 2015, Delphi completed the first automated vehicle cross-country drive … a 15-state, 3,400-mile journey from San Francisco to New York City with a car that, 99 per cent of the time, was driving without human input.'[4]

Google claimed to have achieved 'the world's first truly self-driving ride on public roads at the end of 2015' (Dolgov, 2017). Other companies have released videos that show off their cars' autonomy, navigating a city's streets and finding parking spaces while the human in the driving seat merely watches.

Even if a car manages to drive itself 99% of the time, there is reason to treat the final 1% with extreme caution. Journalists and public observers have leapt upon a series of very public bumps and scrapes as self-driving cars have started to be tested in cities (Muoio, 2016; Reynolds, 2016). Such uncontrolled encounters are more publicly useful than any number of advertisements of perfection, which, as roboticists have pointed out, are 'doomed to succeed' (Brooks and Mataric, 1993).

Tesla has claimed that, even without self-driving, their cars are generating data that will persuade regulators of their increased safety. But regulators legitimately focus on the uncertainties of technological performance. A technology that works well right up to the point that it doesn't, particularly when that point demands the attention of a user who has lost concentration, represents a substantial regulatory problem. By the time a technology switches off, its user has probably also switched off. As one Toyota engineer has argued 'none of us in the automobile or IT industries

are close to achieving true … autonomy …. In some ways, the worst case is a car that will need driver intervention once every 200,000 miles' (Toyota USA, 2017).

For self-driving car engineers, this issue is imagined as the 'handoff problem', a feature of human-computer interaction well known to ergonomists that is now being relearnt. An effect of autopilots on aeroplanes is that they can reduce a pilot's capacity to take control when the situation demands, either because of a long-term atrophying of skills through lack of practice or a short-term disorientation through lack of attention. Long periods of autonomy fundamentally change the role of the human from operator ('in the loop') to supervisor ('on the loop') (Cummings and Thornburg, 2011; Perrow, 1984). Studies of simulated handoffs in cars, when automated systems demand engagement from distracted humans, suggest that it can take as long as 40 seconds for humans to regain full control of a car (Morgan et al., 2016).

In the real world, humans may be unaware of technological failure happening. In some recorded cases with Tesla's Autopilot, the explanation for a malfunction has been clear, such as when a large moth blocked a Model S's radar sensor on a highway (Redebo, 2016). In other cases, however, bemused and trusting users have been unable to explain why Autopilot didn't work as they had expected. Many of these glitches have proven inconsequential. However, they have pointed to a gap between the promise and the reality of a technology that is failing to live up to its name.

'Autonomous vehicles' and 'self-driving cars' as misnomers

The terms 'self-driving cars', 'autonomous vehicles' and 'driverless cars' have been used almost interchangeably in public discourse (Cohen et al., 2017). The differences in nuance implied by these terms should not distract us from a larger concern, which

is with the rhetoric of autonomous technology. Technology, however, is never self-driving (Bijker et al., 1987; Winner, 1977). Claims that technology has a will of its own (e.g. Kelly, 2010), typically disguise a political agenda that is libertarian and deregulatory.

Just as self-driving cars cannot be self-driven, so autonomous vehicles can never be truly autonomous (Stayton, 2015). Self-driving cars are driven by social processes of goal-selection, machine-making, governance, use and their encounters with the world around them. The word 'autonomous' belies not just the human involvement in the cars' creation, but also the connectivity that enables their operation. As Bradshaw and colleagues (2013) point out, 'autonomous system' is a misnomer, if not an oxymoron. 'Autonomy' only happens when tasks and real-world contexts are sufficiently constrained (Bradshaw et al., 2013). 'Autonomous vehicles' are not self-contained, they are not self-sufficient and they are not self-taught. Unlike a conventional car, a self-driving car can function only as part of a fleet. Tesla's cars are bought as individual objects, but commit their owners to sharing data with the company in a process called 'fleet learning'. 'The whole Tesla fleet operates as a network. When one car learns something, they all learn it', with each Autopilot user as an 'expert trainer for how the autopilot should work' (Musk, quoted in Fehrenbacher, 2015). Together, these users generate millions of miles of data every day (Musk, 2016b). Each car's neural network learns how to drive not just from its own experience, but also from the accumulated experience of its thousands of sister vehicles.

> 'Whenever a self-driving car makes a mistake, automatically all the other cars know about it, including future unborn cars … The ability of cars to develop artificial intelligence is so much greater than the ability of people to keep up with them' (Thrun, quoted in Lane, 2016).

For engineers, fleet learning has the dual advantage of massively increasing the speed with which cars can understand the world, while avoiding the imperfections of human learning that are captured in archbishop Becket's lament. Indeed, some researchers have sought to escape the constraint of real-world trial-and-error. Reinforcement learning in simulated environments, including games such as Grand Theft Auto, is seen as one way of accelerating self-driving car training (Filipowicz et al., 2017; You et al., 2017).

For some companies, the interconnectedness of self-driving cars is central to the business model. While Tesla owners have been experimenting with self-driving largely on open highways, ride-sharing taxi companies such as Uber and Lyft have been early investors and experimenters in urban self-driving. In such environments, the efficiency gains from vehicle-to-vehicle and vehicle-to-infrastructure communication become clearer and the tension between interdependence and autonomy becomes more visible (Yoshida, 2017). The benefits of autonomous vehicles may therefore be inversely proportional to their autonomy from one another. In this respect, the stimulus of DARPA may have set self-driving in a counter-productive direction given that the needs of military vehicles, operating in unstructured environments, are very different from conventional ones, which are, in most places, embedded in well-organized highway systems (Schladover, 2009).

Technologies are imagined as solutions to particular problems, and we can examine the construction of those problems in order to interrogate the explicit and implicit purposes of the technology. But technologies also create problems of their own. As Latour and Venn (2002) have discussed, technologies are not merely means to ends, a way of getting from A to B. They are detours, taking us to futures that may be unintended and are impossible to fully calculate in advance. The ends enabled by

technological innovation are typically hard to anticipate and hard to challenge, in part because of innovators' lack of reflexivity about uncertainties and contingencies (Guston, 2014). Accidents therefore provide a constructive opportunity for reframing.

Responding to the first self-driving death

On 30 June 2016, Tesla published a blog post announcing and responding to the crash that had happened six weeks earlier. After explaining that the circumstances of this tragedy were exceptional, following a pattern familiar from previous technological failures (Wynne, 1988), the company's response went on to explain that Autopilot was still a technology 'in beta' and that responsibility for safe driving remained the driver's alone. Musk later told reporters,

> Perfect safety is really an impossible goal. It's really about improving the probability of safety – that's the only thing possible (quoted in Lambert, 2016).

Musk's responses represent an attempt at renegotiating the contract between carmakers, drivers and the market. Conventional carmakers have largely been limited to innovating and testing in private before releasing their products into the wild. Crash-testing, for example, has been a vital part of automotive innovation, but it has been imagined as a private activity, conducted in laboratories or computer simulations (Leonardi, 2010). A production car typically needs to be crashed more than twenty times in tests in order to satisfy regulators. During the early life of the Tesla Model S, the company emphasized that the hardware was world-beating. As well as acceleration that beat most petrol-driven cars, the company trumpeted 'the best safety rating of any car ever tested' (Tesla, 2013). Nevertheless, the Tesla is not sold as a finished article. It is seen as a technology capable of development, a framework for

the accommodation of software that can be continually upgraded. The company argued that, with existing hardware, their software target was a 90% reduction in crashes by using Autopilot (Musk, 2017a). The May 2016 crash and the response therefore represented a public trial less of the car's body than its brain.

The first official report of the May 2016 crash, from the Florida police, put the blame squarely on the truck driver for failing to yield the right of way (Traffic Crash Records, 2016). However, the circumstances of the crash were seen as sufficiently novel to warrant investigations by the National Transportation Safety Board (NTSB) as well as the National Highway Traffic Safety Administration (NHTSA). The NTSB is tasked with identifying the probable cause of every air accident in the US, as well as some highway crashes. However, the novelties of the Tesla collision limited the Board's ability to report quickly on events. Their preliminary report was matter-of-fact. It relates that, at 4:40pm on a clear, dry day, a large truck carrying blueberries crossed US Highway 27A in front of the Tesla, which failed to stop. The Tesla passed under the truck, sheering off the car's roof. The collision cut power to the wheels and the car then coasted off the road for 297 feet before hitting and breaking a pole, turning sideways and coming to a stop. Brown was pronounced dead at the scene. The truck was barely damaged.

The NTSB (2016) noted that the Tesla Model S is a fervent data generator. The car's records revealed that the car was travelling at 74 mph when it hit the truck, 9 mph above the speed limit. The car was also able to tell the NTSB that Autopilot was active at the time of the crash.

In May 2017, the NTSB released its full docket of reports on the crash. From further data released by Tesla, the investigators learnt that Joshua Brown's 40-minute journey consisted of two-and-a-half minutes of conventional driving followed by 37

and a half minutes on Autopilot, during which his hands were off the steering wheel for 37 minutes (NTSB, 2017a). He touched the wheel eight times in response to warnings from the car. The longest time between touches was six minutes. In the minutes before the crash, neither Brown nor Autopilot ever applied the brakes. The NTSB calculated that Brown would have had at least ten seconds to react had he seen the truck. The only witness to speak to the NTSB said that the crash looked like,

> A white cloud, like just a big white explosion … and the car came out from under that trailer and it was bouncing … I didn't even know … it was a Tesla until the highway patrol lady interviewed me two weeks later …. She said it's a Tesla and it has Autopilot, and I didn't know they had that in those cars. (NTSB, 2017b)

This window into the functioning of a car that its owner regarded as capable of self-driving is important, but it remains misty. The car is replete with sensors, but these offer no insight into what the car thought it saw nor how it reached its decisions. The car's brain remained largely off-limits to investigators. At a board meeting in September 2017, one NTSB staff member explained: 'The data we obtained was sufficient to let us know the [detection of the truck] did not occur, but it was not sufficient to let us know why' (NTSB, 2017c).

The NHTSA saw the incident as an opportunity for a crash course in self-driving car innovation. Its Office of Defects Investigation wrote to Tesla demanding data on all of their cars, instances of Autopilot use and abuse, customer complaints, legal claims, a log of all technology testing and modification in the development of Autopilot and a full engineering specification of how and why Autopilot does what it does (NHTSA, 2016a).

While working with the ongoing NTSB and NHTSA investigations, Tesla engaged in a very public form of informal technology assessment with numerous others.

‘Do the math’

The day after Tesla revealed the crash, car safety veteran Clarence Ditlow was quoted as saying that, ‘The Tesla vehicles with autopilots are vehicles waiting for a crash to happen – and it did in Florida’ (quoted in Puzzanghera, 2016). Ditlow said that the Autopilot feature should be disabled until regulators were able to advise on its limits. His criticism related not just to whether Autopilot worked as a driving aid, but also whether it was luring drivers into complacency by implicitly over-claiming.

The German transport minister asked Tesla to rename the system, to which the company responded, ‘Autopilot is a suite of technologies that operate in conjunction with the human driver to make driving safer and less stressful’ and published the results of a survey that suggested 98% of German drivers were aware of the limits of the system (Bild, 2016). For some regulators, including in California, (State of California Department of Motor Vehicles, 2016), the name, and the claims surrounding it, implied that the system was not an automatic pilot, for driver-assistance, but an *autonomous* pilot, to replace a driver. Tesla’s competitors shared the concern. Machine learning pioneer Andrew Ng, just after leaving Google to join Baidu, tweeted, in reference to an earlier, non-fatal crash:

> It's irresponsible to ship driving system that works 1,000 times and lulls false sense of safety, then... BAM! (Ng, 2016).

Trent Victor, an engineer at Volvo, who call their system ‘Pilot Assist’, told a technology reporter that ‘[Autopilot] gives you the impression that it’s doing more than it is … [it] is more of an unsupervised wannabe’ (quoted in Golson, 2016). David Caldwell from Cadillac had justified the delay of a new feature called ‘Super Cruise’ in terms of corporate responsibility: ‘We won’t release it just to hit a date, nor will we

"beta test" with customers' (quoted in Davies, 2016). And one driving journalist pointed out that, while Mercedes' 'Drive Pilot' was technically equivalent to Tesla's Autopilot, the German company had deliberately limited the system in order to avoid users thinking they were in a self-driving car (Jaynes, 2016). Tesla, by making promises about imminent autonomy, seemed to be caught, along with their users, in a web of hyperbole.

Other critics suggested that Tesla's real irresponsibility lay in their failure to reveal the crash at the time of a public stock offering that raised $1.46 billion. Alan Murray, the editor of *Fortune* magazine, claimed that the crash was a material fact for the share price, to which Musk responded,

> Indeed, if anyone bothered to do the math (obviously, you did not) they would realize that of the over 1M auto deaths per year worldwide, approximately half a million people would have been saved if the Tesla autopilot was universally available. Please, take 5 mins and do the bloody math before you write an article that misleads the public. (quoted in Loomis, 2016).

In this exchange, Musk's framing is one of near-perfect consequentialism. He imagines the issue as one of relative risk: the acceptability of self-driving cars should be determined merely in comparison to the safety of conventional, human-driven cars. He is attempting to restrict the debate to technological outcomes rather than processes or purposes. In legitimizing only the probabilistic risk quantity, he is ignoring concerns about risk quality, such as those revealed through social science relating to control, trust, fairness, catastrophic potential, novelty and uncertainty (Irwin et al., 1999; Lupton, 1999; Renn, 1998). Societies may know that air travel is far safer than car travel and still justifiably take issue with aeroplane crashes, for example. A system in which conventional cars are replaced by self-driving cars may see the decline of conventional accidents, but the arrival of new categories of catastrophe relating to 'mode confusion' (NHTSA, 2014, RAND, 2016) or software vulnerabilities. If

regulators merely ‘do the math’, they should be relatively relaxed about occasional technological failure and exceptional circumstances, as long as the aggregate performance improves upon alternatives. Musk’s insistence upon ‘doing the math’ led him to a further rebuttal of journalists’ criticisms:

> If, in writing some article that’s negative, you ... dissuade people from using an autonomous vehicle, you're killing people (quoted in McGoogan, 2016).

Musk’s competitors do not share his certainty. Most would admit that they are a long way away from safe self-driving cars. Gill Pratt from Toyota has claimed that reasonable reassurance about safety would only come after a trillion miles of testing (Guizzo and Ackerman, 2015; also, Kalra and Paddock, 2016). Given that this scale of advance testing is impossible, as it was when Alvin Weinberg (1972) made a similar point about nuclear power stations, the uncertainties, and therefore the politics of experimentation, will never be settled.

Constructing human deficiency

In January 2017, the NHTSA issued its report on the May 2016 Tesla crash. The agency’s initial aim was to ‘examine the design and performance of any automated driving systems in use at the time of the crash’. The conventional outcome of such investigations is a decision on the necessity of a product recall. In this case, however, the NHTSA noted how the system under evaluation had already changed markedly since the crash through wireless software updates.

The report contained two largely separate assessments. The first took an engineering approach, sidestepping the connections with grander visions of full autonomy. The report emphasized that the Tesla Autopilot was a long way from full autonomy. NTHSA broke down Autopilot into its constituent systems: lane centering

control, automatic emergency braking (AEB) and traffic-aware cruise control. This choice to evaluate the technology as a set of merely incremental innovations highlighted some important specifics while overlooking larger questions. The report notes, for example, that AEB is designed for rear-end collisions:

> Braking for crossing path collisions, such as that present in the Florida fatal crash, are outside the expected performance capabilities of the system (NHTSA, 2017).

The second strand of analysis focussed on what the NHTSA called 'human factors'. Their approach was one of 'naïve sociology' (Wynne, 1989), in which technologies, when assessed, are assumed to operate within a world far tidier and more predicable than reality. The criticisms of Tesla amounted to a suggestion that the company's warning information was 'perhaps not as specific as it could be'. The agency notes that advanced driver assistance systems are correlated with increased instances of distractions greater than seven seconds. This is a well-known feature of automated systems, bolstered by research in self-driving car simulators (Körber et al., 2015), in which the 'vigilance decrement' of humans becomes more problematic as automation improves. Nevertheless, the NHTSA were satisfied by Tesla's own research into 'mode confusion' and chose to direct their major recommendation at users: 'Drivers should read all instructions and warnings provided in owner's manuals for ADAS [advanced driver-assistance systems] technologies and be aware of system limitations' (NHTSA, 2017). The NHTSA missed an opportunity for social learning and wider applicability with its assessment. Had, for example, a pedestrian or the driver of another car been killed in the crash, the agency's framing and subsequent assessment may have been very different (Faife, 2017).

The identification and blaming of human deficits has been a common feature of self-driving car innovation. Much of the early justification for autonomous cars

made reference to the more than 90% of car accidents caused by human error (NHTSA, 2017). As people started to encounter Google's test vehicles, reports of low-speed crashes caused by bemused or star-struck humans were common. Google's self-driving lead Chris Urmson (2015) justified their activities as part of a longer trend in car design: 'for the last 130 years we've been working around the least reliable part of the car – the driver'. (When a Google car eventually made a mistake, deciding to pull out in front of a bus and crashing, the tone of news reporting was understandably gleeful).

For Tesla, human unreliability has become clear as they have unfolded their Autopilot project. Following the revelation of the fatal crash, the company's then Director of Autopilot Programs tweeted:

> Human complacency is a serious but separate issue best addressed with education, monitoring & enforcement, not dumbed down safety systems. (Anderson, 2016)

Soon afterwards, the company announced a set of changes to Autopilot that, if not 'dumbing down', imposed substantial extra conditions on the feature. Despite claiming, in common with many innovators after accidents, that the crash was unforeseeable and exceptional, the company sought to make a repeat less likely with new Autopilot software uploaded to every car. Musk claimed that these updates would 'minimize the possibility of people doing crazy things with it' (quoted in Charton, 2015). An initial claim that 'Autosteer now navigates highway interchanges' was removed by the time of the update's final release (Westbrook, 2016). Drivers were warned that they would have to touch the steering wheel more often and that, if they didn't demonstrate that they were paying attention, the system would shut them out, enforcing conventional driving for the rest of the trip. The way that the cars' sensors are used was adjusted to increase the dependence of radar, which is better

than conventional cameras at spotting a thing like a white truck against a white sky. The NHTSA (2017) noted in its report that 'Tesla has changed its driver monitoring strategy to promote driver attention to the driving environment'. By January 2017, the company was referring to Autopilot as a 'hands-on experience'.

For cars built with the new hardware, Tesla activated Autopilot in January 2017, with Musk saying 'please be cautious' (Musk, 2017b). The algorithms had to relearn how to employ the car's new sensors, which meant a step backwards in performance. A class action law suit from a group of frustrated Autopilot users claimed that Tesla had sold them a product that failed to live up to its self-driving promises and was unsafe when used (Sheikh et al. v Tesla, 2017). The maximum speed of the new Autopilot was initially limited to 45 mph – 'for heavy traffic, where it is needed most', Musk tweeted (Musk, 2017c) – while they tested the system.

These changes did not satisfy the NTSB. Their final word on the probable cause of the Tesla crash added a concern with Autopilot's 'operational design, which permitted [the driver's] prolonged disengagement from the driving task and his use of the automation in ways inconsistent with guidance and warnings from the manufacturer' (NTSB, 2017c). The board considered that merely asking drivers to touch the steering wheel more was an inadequate response. Tesla, in the words of one NTSB staffer, 'did little to constrain the use of autopilot to roadways for which it was designed' (ibid.) A statement from Brown's family, rather than blaming Tesla, emphasised social learning: 'When rail systems, metro systems, and personal vehicles (etc.) were constructed, fatalities occurred and we learned from them ... Part of Joshua's legacy is that his accident drove additional improvements making the new technology even safer.' (Landskroner Grieco Merriman, 2017).

Tesla's public modulation of Autopilot's capabilities represents a new mode of engagement with car customers. Drivers are expected to be clear about the limits of the technology, but these limits are continually being redefined by the company and tested by users, a subset of whom are engaged in a form of alternative online pedagogy. Some claim to have hacked the Tesla's software or invented ways to quash the car's warnings to hold the steering wheel (e.g., MEtv Product Reviews, 2016). YouTube is replete with hands-free Tesla driving displays and other haphazard experimentation (Brown and Laurier, 2017). In one much-shared case, the driver appears to be asleep while his Tesla moves along in traffic. While official bodies such as the NTSB may insist that the Tesla is not a self-driving car, a significant number of drivers are behaving as if they disagree. These users would seem to be at the bottom of the 'certainty trough', committed to the technology but distant enough to be unaware of its contingencies (MacKenzie, 1990). Musk seems to recognize the dangers of this false certainty, admitting that 'Autopilot accidents are far more likely for expert users' (quoted in Ramsey, 2017).

The connection back to the company is unclear. Tesla claims that it listens to drivers, such as with an update removing the speed limit at which Autopilot could be set. The company is keen to demonstrate that its innovation lies in its software, and that this allows for a new form of responsiveness. However, the model of learning is highly privatized. When evidence of misuse arises, the response is typically to blame users' ignorance and backtrack on expectations. Tesla's emailed response to *Consumer Reports'* call for a moratorium on Autopilot was that, '[w]hile we appreciate well-meaning advice from any individual or group, we make our decisions on the basis of real-world data, not speculation by media' (quoted in *Consumer*

Reports, 2016). This example of a company hiding behind opaque, proprietary data is a governance concern to which I will return in the conclusion.

Technological alternatives

Although they avoided publicly connecting the crash with their cars' flaws, Tesla made a set of technological tweaks in the weeks following the announcement of the accident. Rather than using the 'EyeQ' system from Israeli company Mobileye that also powers other companies' more humble safety systems, they announced a move to develop their own software for use with Nvidia hardware. As Tesla parted company with Mobileye, differences in their versions of events exposed some of the contingencies of machine learning in cars. Tesla didn't mention the crash, but Mobileye's chairman expressed his frustration: '[W]e need to be there on all aspects of how the technology is being used, and not simply providing technology' (quoted in Ramsey, 2016). The company's head of communications added: 'This incident involved a laterally crossing vehicle, which current-generation AEB [automatic emergency braking] systems are not designed to actuate upon' (quoted in Fierman, 2016), a detail that would later be highlighted by the NHTSA.

Nvidia have fewer qualms. Following revelations about the potential of their GPU chips for use in deep neural networks in the late 2000s, the company has grown its machine learning business. Their emphasis is on 'end-to-end' machine learning, in which the system works out the rules that need to be prioritized (detecting the edges of roads, for example) for the solution of its task (Bojarski et al., 2016a). This unsupervised, self-optimizing approach concentrates less on the rule-setting process of formal algorithm design and more on training data, such as from the steering wheel of a human driver, allowing the network to learn rules by itself (Bojarski et al.,

2016b). This creates a hunger for new forms of input. In January 2017, Nvidia announced that they would be adding cameras facing inwards and well as outwards, to learn about drivers' behaviour from their faces as well as their interface with the car's controls.

Competitors have taken issue with this approach, arguing that the seemingly improved average performance of such systems can obscure the tiny but vital fraction of situations in which they fail. Mobileye's founder described the trouble with end-to-end learning in dealing with 'corner cases', the rare events that happen outside normal parameters but which humans may nevertheless find relatively manageable. His argument was that a single digital neural network that has taught itself a set of rules will be less able than multiple specialized subsystems with responsibilities such as detecting pedestrians, detecting lanes or detecting road signs, each trained with formal logic, 'domain expertise' and machine learning, to deal with the unexpected. He pointed out that 'We're talking here about a lot of work' to get from a relatively good algorithm that works almost all the time to one that is trustworthy 99.9% of the time (Shashua, 2016). Mobileye claims that its 'semantic abstraction' approach requires more up-front effort in teaching the vocabulary of driving but fewer training examples (Shaley-Shwartz and Shashua, 2016). Rather than treating self-driving as a single problem, making it look deceptively 'solvable', the challenge is broken up, making complexities more visible.

Other self-driving start-ups emphasize the need for formal logic as a route to verification, interpretability and accountability, making it possible to know the whys and wherefores of algorithmic decision-making (Ackerman, 2016). All approaches adopted by companies make use of multiple forms of data and learning, but the balance varies, and depends on matters of political economy as well as engineering.

For the car industry, autodidactic deep learning is far more disruptive, because it downplays more than a century's worth of accumulated expertise and cumulative learning relating to the sociotechnical system comprising cars and their material and social infrastructure. Deep learning instead presumes that driving is merely another task that can be learnt from human practitioners, mastered and improved upon. Rule-making trumps rule-following.

Tesla claim that its Nvidia-powered 'Tesla Vision' deep neural network 'deconstructs the car's environment at greater levels of reliability than those achievable with classical vision processing techniques' (Tesla, 2017a). Their hope is that this increase in brainpower will compensate for a lack of formal education.

As of October 2017, Tesla sees no problem with its sensors. It claims that its combination of GPS, radar, ultrasound and eight cameras is now capable of full autonomous driving. It is just waiting for its software and the regulators to catch up. Other driverless innovators are unconvinced that such hardware is up to the job. While Tesla relies on ultrasonic sensors for short distances and cameras and radar for detecting objects further away, other companies have invested heavily in LiDAR – a laser-based step-up from radar. LiDAR has a longer range than ultrasound and is better than radar at precisely spotting small objects made from a wide range of materials. But the technology is, as of 2016, prohibitively expensive and bulky for a private car. Competitor companies see affordable LiDAR as critical and expect its price to follow a downward trajectory similar to that of radar.

The opening up of such technological disagreements suggests that, despite the existence of various online explainers for 'how self-driving cars work', and the imagined gaps between engineers' and public understandings of this, there are some subtle but profound differences of approach currently in play. Given these

contingencies, policymakers have a more active part to play in the development of approaches, the setting of standards and the integration of technology into the built environment.

Democratizing learning

As I have described, autonomous vehicles are not as heroically independent as their enthusiasts would have us believe. Nor are they as autodidactic. The story of autonomy is a way of downplaying a car's connections with other vehicles, the built environment and the infrastructure of regulation. It is a story that deserves to be challenged. The emergence of self-driving cars will be a process of social learning that can and should be democratized.

Much of the noisiest excitement surrounding self-driving cars has come from a culture of innovation that has little experience of the material, non-digital world and is unused to intense regulation. Large (though still young) Silicon Valley companies such as Tesla and Google have been joined by start-ups like Comma.ai, a company that promises to 'solve self-driving' by offering a build-it-yourself self-driving car kit. Car manufacturers with long histories that have come to accommodate (and in some cases define) substantial government oversight have, through acquisitions and partnerships, sought to take advantage of these new possibilities.

This clash of hardware and software cultures raises immediate questions of governance. For example, these two worlds understand product liability very differently. Cars are conventionally designed, tested, and released as finished products with an ever-present threat of product recalls, fines or civil law suits if they are deemed defective. Software, however, is governed in most jurisdictions as a service rather than a product, and granted substantial leeway (Chander, 2014). As

Nissenbaum (1994) observed more than twenty years ago, we must 'accept that the producers of computer systems are not, in general, fully answerable for the impacts of their products. If not addressed, this erosion of accountability will mean that computers are 'out of control' in an important and disturbing way.' The norm is to evaluate software liability once defects are detected and consider whether innovators could reasonably have foreseen them. The rapid uptake of machine learning looks set to exacerbate the irresponsibility that Nissenbaum feared. As the stakes of software deployment rise – in online security, social media and robotics – we may well see self-driving cars as a test case for the hardening of software regulation.

JafariNaimi (2017) argues that the self-driving car presents an opportunity for reframing transport governance. With the automobile in the 20th century, a strong idea about what a car was – an everyday object like a bicycle rather than a sociotechnical system like a train and its tracks – led to regulatory regimes that concentrated power with cars, their drivers and manufacturers. As with other emerging technologies (Rayner, 2004), makers of self-driving cars see their unarguable potential being held back by lags and deficits – in infrastructure, law and public understanding. However, the technology's promise is open-ended and its effects are impossible to reliably calculate. Rather than taking the technology as fixed and looking to plug the deficits of law or public understanding that are imagined around it, policymakers should instead see self-driving cars as an opportunity for more active engagement in the shaping of technological systems, prioritizing social learning and knitting self-driving cars back into their social worlds.

The emergence of self-driving cars, with all the missteps and misadventures that will occur as they mix with other modes of transport, will represent an expansion of what is already a form of disorganized social experimentation. Good governance

will mean resisting the privatization of learning that is happening. It will mean engaging not just with technological outcomes, which, given the complexity of transport systems, will be radically indeterminate, but also with the processes and purposes – inscribed and implicit – of innovation.

As I describe above, there are clear tensions between social learning and the pure form of machine learning manifest in deep neural networks. The opacity of machine learning systems, both deliberate and accidental, offers an excuse for innovators and a barrier for governance. However, a closer look at self-driving car innovation reveals some constructive alternatives. Engineers have already had to engage in a form of socialized machine learning, building on research in social robotics (e.g., de Greeff and Belpaeme, 2015). As algorithms meet the material world, social machine learning becomes unavoidable and engineers' responsibilities come to the fore (Nourbakhsh, 2013). If engineers are able to respond, then the narrow sense of the 'social' already programmed into some cars' neural networks – in which, for example, it is assumed that users are error-prone and pedestrians are just another part of the passive environment – need not be imposed on the world. There is some evidence that companies are already differentiating their approaches to self-driving technology (or, as some companies more modestly put it 'driver assistance technology').

Algorithmic efficacy has attracted substantial attention, but it will be only part of the innovation required to make self-driving cars work. Engineering efforts to improve the interpretability of deep learning systems challenge the narrative of inevitable opacity that has until recently provided an easy excuse for irresponsibility. Alongside the debate about interpretability, mitigating 'mode confusion' has become an important target for design. While some have argued that AVs should be allowed

to travel unlabelled, so that other road users do not take advantage of their presumed generosity, there is growing recognition among engineers of the need for vehicles to actively communicate their presence, their intentions and their capabilities to other road users (Surden and Williams, 2016). One AI researcher has suggested the need for a 'Turing red flag law' (Walsh, 2016), mandating the clear labelling of all autonomous systems.

The politics of novelty surrounding self-driving cars is unpredictable. However, if the whole system is to be as transformative as is claimed, its novelty should not be defined merely by technical advances in algorithms. Socializing machine learning demands the closer integration of insights from human-computer interaction and collaborative design into engineering rather than a presumption that a self-driving car merely means replacing a person with a computer.

Most corporate and regulatory statements on self-driving have overlooked the contingencies of machine learning to focus on human deficiencies. As well as blaming human error for crashes (NHTSA, 2017) policy analyses have focussed on the need for public education (GHSA, 2017; Policy Network 2016; Waymo 2017). This hubris is likely to lead to a model of accidental governance in which car manufacturers set the terms of experimentation, and events such as crashes come to define, in the minds of publics and regulators, the trajectory of technology. Countering this requires the deployment of what Jasanoff (2003) calls 'technologies of humility', devices for engaging with the profound uncertainties of innovation. This means reimagining public participation not as education, but as democracy.

An important entry point for governments into the process of learning and experimentation with self-driving cars is through the sharing of data. A self-driving car can already generate a gigabyte of data each second. The investigations of the

Tesla crash provide a window into the politics of data sharing. The NTSB were able to rapidly learn, from data volunteered by Tesla, what happened inside the car in the minutes before the crash, but Autopilot's decision-making remained opaque. Airline regulators mandate the inclusion of flight data recorders (nicknamed 'black boxes') that capture conversations between pilots as well as flight data which are then shared. One NTSB mantra is that 'anybody's accident is everybody's accident'. Some car companies are starting to emphasize accountability. They conclude that, in the event of a crash, it is important not just to know what happened but why it happened: one self-driving research project funded by Toyota is called **'**The car can explain**'**. Tesla is one of the few carmakers not to follow NHTSA guidance on event data recorders, which means that regulators must rely on the company's generosity when crashes happen. The need to improve social learning has led some to call for the inclusion of 'ethical black boxes' in robotic systems (Winfield and Jirotka, 2017). The reasons for doing so go beyond the investigation of accidents.

Data is the fuel for machine learning and it is a source of competitive advantage for car companies. It is impossible to predict precisely how data will be monetized, because of the wide range of possibilities of future transport systems, but we can anticipate that aggregated or personalized data relating to geography, driving behaviours, traffic, people flow and more will become an important currency for future innovation. The economies of scale will be substantial, tending towards new concentrations of economic power.

Once we reject the narrative of autonomy and recognize the thicket of connections between cars and the outside world, we can imagine new possibilities for machine learning in the service of social learning. If the development of self-driving algorithms is to realize some of the public value that its developers suggest, then there

is a strong case for collaboration rather than competition. If cars learn more effectively as fleets, then it is reasonable to expect responsible car companies to share their learning with others. However, if algorithms and the data that feed them are imagined to be, as seems likely, a source of competitive advantage, then the public value of self-driving technology will be diminished.

Some tentative governance proposals in the US have urged greater data-sharing. Guidance from the NHTSA, launched by President Obama in September 2016, uses the language of 'learning' and 'group learning' to justify its call for data-sharing, particularly where urban trials are licensed by government authorities. The NHTSA also suggest that companies should collect and analyse data on 'near misses and edge cases', join an 'early warning reporting program' and find ways for their cars to communicate with one another (NHTSA, 2016b). The NTSB concluded its investigation with a similar recommendation: 'We don't think each manufacturer of these vehicles need to learn the same lessons independently. We think by sharing that data, better learning and less errors along the way will happen' (NTSB, 2017c).

The initial policy focus, as with local governance measures, is on safety. The Californian Department of Motor Vehicles initiated a draft policy in December 2016 demanding that companies provide data not just on accidents involving AVs, but also on 'disengagements', the moments when self-driving technology fails and demands human input (State of California Department of Motor Vehicles, 2016). The disengagement reports submitted by companies reveal the gap between informal experimentation and formal compliance. Tesla claims that, by the end of 2016, its customers had already covered more than a billion miles in Autopilot mode. Meanwhile, the company's report submitted to the Californian DMV presents data on

its four test vehicles, which covered only 550 miles in 2016 and disengaged on average every three miles (Tesla, 2017b).

What counts as a disengagement is largely left to the companies to decide. But, if nothing else, such reports begin to organize social learning from self-driving car experimentation. As it stands, much of the NHTSA guidance is voluntary, albeit with a thinly veiled threat of pre-market approval and proactive regulation if car companies misbehave. Some companies have already sought to demonstrate their responsibility in data sharing in order to head off top down controls. Uber, for example, has volunteered aggregate data on ride sharing for the benefit of transport planning. Industry representatives have responded that self-driving car data will be commercially valuable and must therefore be proprietary, except in situations where safety is a priority (Hawkins, 2016).

Even if sufficient data is forthcoming in the event of crashes, this will be only a small part of a much larger process of social learning. US leadership in self-driving car innovation has meant the inheritance of a mode of governance in which, among other characteristics, cars are seen as self-evidently beneficial, risks are governed retrospectively (often through the courts), concerns about liberty are relatively elevated over those of public safety and public transport receives little support. The default has been to govern self-driving cars according to this framework, defining risks narrowly while emphasizing the need to attract investment and create new markets. Other countries have adopted a modulated form of this technology-first approach. However, there are notable examples of policies that, by starting with a focus on transport rather than technology, reframe the challenge. On the narrow issue of road safety, for example, Vision Zero, a strategy for reducing road deaths that began in Sweden, offers an alternative model of social learning that puts car

innovation alongside infrastructure, law and social norms in redistributing responsibility for safety (Eriksson, 2017; JafariNaimi, 2017). A social learning approach to governing self-driving cars would be similarly well rounded, putting the promise of machine learning in its place.

Acknowledgments

This paper was written during a sabbatical year at the University of Colorado, Boulder. Thanks are due to Roger Pielke Jr, Max Boykoff and colleagues for hosting me there. Versions of this paper were presented at the University of Wyoming, Arizona State University, University of California Berkeley, Colorado School of Mines, the University of Colorado Boulder and the 2017 meeting of the Science and Democracy Network at Harvard. Thank you to these audiences for helping to sharpen the ideas. The paper received extensive and thoughtful critique from three anonymous reviewers and Sergio Sismondo. It has also benefitted from conversations with Jon Agar, Melissa Cefkin, Phyllis Illari, Andrew Maynard, Brent Mittelstadt, Clark Miller, Harry Surden, Michael Veale and Jamey Wetmore. Thanks also to Celeste Maldonado for help with the references.

Notes

[1] A good discussion of the feasibility and desirability of such a legal right can be found in Goodman and Flaxman (2016), Wachter et al. (2017) and Edwards and Veale (2017)

[2] 'Demo 97', for example, with a consortium of Government funders and car manufacturers, embedded magnets in a stretch of Interstate north of San Diego (in 1997) so that cars could follow the road's twists and turns. One company claimed this was 'an integrated vehicle control system that is helping move automated highways from science fiction to reality' (quoted in Wetmore, 2003). In Europe, the PROMETHEUS Project (PROgraMme for a European Traffic of Highest Efficiency and Unprecedented Safety) brought together a large number of universities and car companies to develop trial systems in the late 1980s and early '90s. Such grand projects never got beyond the trial phase, however.

[3] Thanks to Michael Veale for this point.

[4] Researchers from Carnegie Mellon claimed to have made a similar journey in 1995, dubbed 'No Hands Across America', with their car driving itself 98% of the time (Jochem and Pomerleau, 1996).

References

Ackerman E (2016) After mastering Singapore's streets, NuTonomy's robo-taxies are poised to take on new cities. *IEEE Spectrum*. 29 December. Available at: http://spectrum.ieee.org/transportation/self-driving/after-mastering-singapores-streets-nutonomys-robotaxis-are-poised-to-take-on-new-cities (accessed 12 March 2017).

Ananny M (2016) Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology & Human Values* 41(1): 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*. DOI: 10.1177/1461444816676645.

Anderson A (2016) Human complacency is a serious but separate issue best addressed with education, monitoring & enforcement, not dumbed down safety systems. [Twitter]. 6 July. Available at: https://twitter.com/sterling_a/status/750588620378087424 (accessed 14 March 2017).

Baker D (2017) Driverless milestone: No hands across America. *San Francisco Chronicle*, 14 July.

Bandura A (1988) Organisational applications of social cognitive theory. *Australian Journal of Management* 13(2): 275–302.

Berkes F and Turner, NJ (2006) Knowledge, learning and the evolution of conservation practice for social-ecological system resilience. *Human Ecology* 34(4): 479–494.

Bijker WE, Hughes TP, and Pinch TJ (eds.) (1987) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology.* Cambridge: MIT Press.

Bild (2016) AutoTesla weist behördenkritik am «Autopilot»-namen zurück. 17 October. Available at: http://www.bild.de/geld/aktuelles/wirtschaft/kba-fordert-tesla-auf-nicht-mehr-mit-autopilot-48311846.bild.html (accessed 12 March 2017).

Bojarski M et al. (2016a) End to end learning for self-driving cars. *arXiv*, 25 April. Available at: https://arxiv.org/pdf/1604.07316v1.pdf (accessed 12 March 2017).

Bojarski M et al. (2016b) End-to-end deep learning for self-driving cars. *Nvidia*. 17 August. Available at: https://devblogs.nvidia.com/parallelforall/deep-learning-self-driving-cars/ (accessed 12 March 2017).

Bonnefon JF, Shariff A, and Rahwan I (2016) The social dilemma of autonomous vehicles. *Science* 352(6293): 1573–1576.

Borup M, Brown N, Konrad K and Van Lente H (2006) The sociology of expectations in science and technology. *Technology Analysis & Strategic Management* 18(3-4): 285–298.

Both G (2014) What drives research in self-driving cars? Available at: http://blog.castac.org/2014/04/what-drives-research-in-self-driving-cars-part-2-surprisingly-not-machine-learning/ (accessed 11 March 2017).

Bradley R (2016) Tesla Autopilot. *MIT Technology Review*. Available at: https://www.technologyreview.com/s/600772/10-breakthrough-technologies-2016-tesla-autopilot/ (accessed 12 March 2017).

Bradshaw JM, Hoffman RR, Woods DD and Johnson, M (2013) The seven deadly myths of "Autonomous Systems". *IEEE Intelligent Systems* 28(3): 54–61.

Brooks RA and Mataric MJ. (1993) Real robots, real learning problems. In: Connell JH, and Mahadevan S (eds) *Robot Learning.* Springer US, 193–213.

Brown B and Laurier E (2017) The trouble with autopilots: Assisted and autonomous driving on the social road. CHI 2017, May 06-11, 2017, Denver, CO, USA

Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data & Society* 3(1) DOI: 10.1177/2053951715622512.

Chander A (2014) How law made Silicon Valley. *Emory Law Journal* 63(3): 639–694.

Charton A (2015) Tesla owners need to stop doing crazy things with Autopilot, says Elon Musk. *International Business Times*. 4 November. Available at: http://www.ibtimes.co.uk/tesla-owners-need-stop-doing-crazy-things-autopilot-says-elon-musk-1527121 (accessed 12 March 2017).

Cohen T, Jones P, and Cavoli C (2017) Social and behavioural questions associated with automated vehicles. *Scoping study by UCL Transport Institute. Final report*. London: Department for Transport.

Collingridge D (1980) *The Social Control of Technology*. London: Open University Press

Collins HM (1988) Public experiments and displays of virtuosity: The core-set revisited. *Social Studies of Science* 18(4): 725–748.

Consumer Reports (2016). Tesla's Autopilot: Too much autonomy too soon. 14 July. Available at: http://www.consumerreports.org/tesla/tesla-autopilot-too-much-autonomy-too-soon/ (accessed 12 March 2017).

Crawford K and Calo R (2016) There is a blind spot in AI research. *Nature* 538: 311–313.

Cummings ML and Thornburg KM (2011) Paying attention to the man behind the curtain. *IEEE Pervasive Computing* 10(1): 58–62.

Davies A (2016) Cadillac's delaying its first whack at a self-driving car. *Wired*. 14 January. Available at: https://www.wired.com/2016/01/cadillacs-delaying-its-first-whack-at-a-self-driving-car/ (accessed 12 March 2017)

de Greeff J and Belpaeme T (2015) Why robots should be social: Enhancing machine learning through social human-robot interaction. *PLoS One* 10(9): e0138061.

De Vos G (2016) Testimony of Glen W. De Vos – Senate Commerce, Science and Transportation Committee Hearing. Delphi Automotive. Available at: https://www.commerce.senate.gov/public/_cache/files/86053bb6-58d8-4072-a033-03f36766d0c3/0EEB8D39E9E224E1A35D274D5325C181.2016-march-15---delphi-de-vos-testimony-final.pdf (accessed 12 March 2017).

Dewey J (1916) *Essays in Experimental Logic*. Chicago: University of Chicago Press.

Dolgov D (2017). Accelerating the pace of learning. Available at: https://medium.com/waymo/accelerating-the-pace-of-learning-36f6bc2ee1d5#.d0dtdd8g1 (accessed 12 March 2017).

Edwards L and Veale M (2017) Slave to the algorithm? Why a 'right to explanation' is probably not the remedy you are looking for. Available at: https://ssrn.com/abstract=2972855

Elish MC (2016) Moral crumple zones: Cautionary tales in human-robot interaction. Available at: https://papers.ssrn.com/sol3/papers.cfm (accessed 31 March 2017).

Eriksson M (2017) The normativity of automated driving: A case study of embedding norms in technology. *Information & Communications Technology Law* 26(1): 46–58.

Etherington D (2016) Musk targeting coast-to-coast test drive of fully self-driving Tesla by late 2017. *TechCrunch*. 19 October. Available at: https://techcrunch.com/2016/10/19/musk-targeting-coast-to-coast-test-drive-of-fully-self-driving-tesla-by-late-2017/ (accessed 11 March 2017).

European Environment Agency. (2001) *Late Lessons from Early Warnings: The Precautionary Principle 1896-2000*. P. Harremoës (Ed.). Luxembourg: Office for Official Publications of the European Communities.

Faife C (2017) Drivers use Tesla Autopilot at their own risk, investigators conclude. *Vice Motherboard*. 19 January. Available at: http://motherboard.vice.com/en_ca/read/drivers-use-tesla-autopilot-at-their-own-risk-investigators-conclude (accessed 12 March 2017).

Fehrenbacher K (2015). How Tesla is ushering in the age of the learning car. *Fortune*. 16 October. Available at: http://fortune.com/2015/10/16/how-tesla-autopilot-learns/ (accessed 12 March 2017).

Fierman W (2016) BMW just announced that self-driving cars are coming in less than 5 years. *Business Insider.* 1 July. Available at: http://www.businessinsider.com/bmw-2021-autonomous-cars-2016-7?r=UK&IR=T (accessed 12 March 2017).

Filipowicz A, Liu J and Kornhauser A (2017) Learning to recognize distance to stop signs using the virtual world of Grand Theft Auto 5. Paper for the Transportation Research Board.

Finn E (2017) *What Algorithms Want: Imagination in the Age of Computing*. Cambridge: MIT Press.

Fischer, F. (2000) *Citizens, Experts and the Environment: The Politics of Local Knowledge*. Durham: Duke University Press.

Frankel T (2016) Elon Musk says Tesla's Autopilot is already 'probably' better than human drivers. *The Washington Post*, 1 November.

Friedmann J and Abonyi G (1976) Social learning: A model for policy research. *Environment and Planning A* 8(8): 927–940.

Golson J (2016) Volvo autonomous car engineer calls Tesla's Autopilot a 'wannabe'. *The Verge*. 27 April. Available at: http://www.theverge.com/2016/4/27/11518826/volvo-tesla-autopilot-autonomous-self-driving-car (accessed 12 March 2017).

Gomes L (2016) Google self-driving car will be ready soon for some, in decades for others. *IEEE Spectrum*. 18 March. Available at: http://spectrum.ieee.org/cars-that-think/transportation/self-driving/google-selfdriving-car-will-be-ready-soon-for-some-in-decades-for-others (accessed 12 March 2017).

Goodman B and Flaxman S (2016) European Union regulations on algorithmic decision-making and a' right to explanation'. *arXiv*: 1606.08813.

GHSA (Governors Highway Safety Association) (2016) Autonomous vehicles meet human drivers: Traffic safety issues for states. Available at: http://www.ghsa.org/resources/spotlight-av17 (accessed 17 March 2017).

Gross M (2010) *Ignorance and Surprise: Science, Society, and Ecological Design*. Cambridge: MIT Press.

Guizzo E and Ackerman E (2015) Gill Pratt discusses Toyota's AI plans and the future of robots and cars. *IEEE Spectrum*, 11 Sept.

Guston DH (2014) Understanding 'anticipatory governance'. *Social Studies of Science* 44(2): 218–242.

Hajer M (2003) Policy without polity? Policy analysis and the institutional void. *Policy Sciences* 36(2): 175–195.

Hall PA (1993) Policy paradigms, social learning, and the state: The case of economic policymaking in Britain. *Comparative Politics*: 275–296.

Hawkins A (2016) Self-driving car makers don't sound super excited to share data with the federal government. *The Verge*. 20 September. Available at: http://www.theverge.com/2016/9/20/12991302/google-uber-ford-lyft-volvo-self-driving-car-reaction-dot-nhtsa (accessed 12 March 2017).

Hoppe R (2011) *The Governance of Problems: Puzzling, Powering and Participation*. Policy Press.

Irwin A, Simmons P and Walker G (1999) Faulty environments and risk reasoning: The local understanding of industrial hazards. *Environment and Planning A* 31(7): 1311–1326.

JafariNaimi N (2017) Our bodies in the trolley's path, or why self-driving cars must *not* be programmed to kill. *Science, Technology, & Human Values*. DOI: 10.1177/0162243917718942

Jasanoff S (1995) Product, process, or programme: Three cultures and the regulation of biotechnology. In Bauer M (ed) *Resistance to New Technology: Nuclear Power, Information Technology and Biotechnology*. London: Cambridge University Press, 311–331.

Jasanoff S (2003) Technologies of humility: Citizen participation in governing science. *Minerva* 41(3): 223–244.

Jasanoff S (2016) *The Ethics of Invention: Technology and the Human Future*. WW Norton & Company.

Jaynes N (2016) Mercedes Drive Pilot isn't as precise as Tesla Autopilot, and that's OK. *Mashable*, 16 June. Available at: http://mashable.com/2016/06/16/2017-mercedes-benz-e-class-first-drive/#HpXhH3dwMqqd (accessed 22 August 2017).

Jochem T and Pomerleau D (1996) Life in the fast lane: The evolution of an adaptive vehicle control system. *AI magazine* 17(2): 11.

Kalra N and Paddock S (2016) Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Available at: https://www.rand.org/content/dam/rand/pubs/research_reports/RR1400/RR1478/RAND_RR1478.pdf (accessed 1 August 2017).

Kearnes M, Grove-White R, Macnaghten, et al. (2006) From bio to nano: Learning lessons from the UK agricultural biotechnology controversy. *Science as Culture* 15(4): 291–307.

Kelly K (2010) *What Technology Wants*. New York: Penguin.

Körber M, Cingel A, Zimmermann M and Bengler K (2015) Vigilance decrement and passive fatigue caused by monotony in automated driving. *Procedia Manufacturing* 3: 2403–2409.

Kosner AW (2013) Why is machine learning (CS 229) the most popular course at Stanford? *Forbes*. 29 December.

Krizhevsky A, Sutskever I and Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems 25*: 1097–1105.

Kuzma J (2016) Reboot the debate on genetic engineering. *Nature* 531(7593): 165–167.

Lambert F (2016) Transcript: Elon Musk's press conference about Tesla Autopilot under v8.0 update [Part 2]. *Electrek*. 11 September. Available at: https://electrek.co/2016/09/11/transcript-elon-musks-press-conference-tesla-autopilot-under-v8-0-update-part-2/ (accessed 12 March 2017).

Landskroner Grieco Merriman (2017) Statement from the Family of Joshua Brown, 11 Sept 2017, Available at: https://docs.wixstatic.com/ugd/63eaea_19dfd6ee29ae4a60a4eaeeb54786f279.pdf (accessed 19 October 2017)

Lane A (2016) 'Lo and Behold' and 'Mia Madre' reviews. *The New Yorker.* 29 August.

Latour B and Venn C (2002) Morality and technology: The end of the means. *Theory, Culture & Society* 19(5-6): 247–260.

LeCun Y, Bengio Y and Hinton G (2015) Deep learning. *Nature*, 521(7553): 436–444.

Leonardi PM (2010) From road to lab to math: The co-evolution of technological, regulatory, and organizational innovations for automotive crash testing. *Social Studies of Science* 40(2): 243–274.

Lipson H and Kurman, M (2016) *Driverless: Intelligent Cars and the Road Ahead*. Cambridge: MIT Press.

Loomis C (2016) Elon Musk says Autopilot death 'not material' to Tesla shareholders. *Fortune*. 5 July. Available at: http://fortune.com/2016/07/05/elon-musk-tesla-autopilot-stock-sale/ (accessed 12 March 2017).

Lupton D (1999) *Risk and Sociocultural Theory: New Directions and Perspectives*. London: Cambridge University Press.

Mackenzie A (2015) The production of prediction: What does machine learning want?. *European Journal of Cultural Studies* 18(4-5): 429–445.

MacKenzie DA (1990) *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*. Cambridge: MIT press.

McGoogan C (2016) 'You're killing people': Elon Musk attacks critics of self-driving cars. *The Telegraph*. 20 October. Available at: http://www.telegraph.co.uk/technology/2016/10/20/youre-killing-people-elon-musk-attacks-critics-of-self-driving-c/(accessed 12 March 2017).

MEtv Product Reviews (2016) Silence and defeat the Tesla Autopilot nanny feature in v8.0. *YouTube*. 16 September. Available at: https://www.youtube.com/watch?v=wHHRnHnm1xk (viewed 12 March 2017).

Millstone E, Brunner E and Mayer S (1999) Beyond 'substantial equivalence'. *Nature* 401(6753): 525–526.

Mittelstadt BD, Allo P, Taddeo M, et al. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2): 2053951716679679.

Mnih V, Kavukcuoglu K, Silver, D, et al. (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540): 529–533.

Moore MM and Lu B (2011) Autonomous vehicles for personal transport: A technology assessment. Available at: https://papers.ssrn.com/sol3/papers2.cfm?abstract_id=1865047

Morgan P, Alford C and Parkhurst G (2016) Handover issues in autonomous driving: A literature review. *Project Report*. University of the West of England, Bristol, UK.

Morozov E (2013) *To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems that Don't Exist*. London: Penguin UK.

Muoio D (2016) I was behind the wheel when a self-driving Uber failed – here's what happens. *Business Insider*. 24 December. Available at: http://www.businessinsider.com/uber-self-driving-car-fails-2016-12 (accessed 12 March 2017).

Musk E (2016a). Great rebuttal by a Tesla owner to those calling for Autopilot to be disabled. Was written with zero input from us. [Twitter]. 17 July. Available at: https://twitter.com/elonmusk/status/754786895343779840 (accessed 12 March 2017).

Musk E (2016b). Master plan, part deux. *Tesla*. 20 July. Available at: https://www.tesla.com/blog/master-plan-part-deux (accessed 12 March 2017).

Musk E (2017a). @TOCHOTE Our target is a 90% reduction with HW2 as the software matures. [Twitter]. 19 January. Available at: https://twitter.com/elonmusk/status/822131228585271296 (accessed 12 March 2017).

Musk E (2017b). Autopilot for HW2 rolling out to all HW2 cars today. Please be cautious. Some cars will require adjustment of camera pitch angle by service. [Twitter]. 21 January. Available at:

https://twitter.com/elonmusk/status/822922507535560705?ref_src=twsrc%5Etfw (accessed 12 March 2017).

Musk E (2017c). Auto steer limited to 45 mph on highways for now, i.e. heavy traffic, where it is needed most. Limit will raise as we get more data. [Twitter]. 22 January. Available at: https://twitter.com/elonmusk/status/823234503355158528 (accessed 12 March 2017).

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

Ng A (2016) It's irresponsible to ship driving system that works 1,000 times and lulls false sense of safety, then… BAM! [Twitter]. 27 May. Available at: https://twitter.com/AndrewYNg/status/736265938782167040 (accessed 12 March 2017).

NHTSA (National Highway Traffic Safety Administration) (2008) National motor vehicle crash causation survey. Report to Congress. Available at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059 (accessed 12 March 2017).

NHTSA (2014) Human factors evaluation of level 2 and level 3 automated driving concepts. Available at: https://www.nhtsa.gov/DOT/NHTSA/NVS/Crash%20Avoidance/Technical%20Publications/2014/812044_HF-Evaluation-Levels-2-3-Automated-Driving-Concepts-f-Operation.pdf

NHTSA (2016) Letter from Jeffrey Quandt to Mathew Schwall. 8 July. Available at: https://www-odi.nhtsa.dot.gov/acms/cs/jaxrs/download/doc/UCM533397/INIM-PE16007-64338.pdf (accessed 12 March 2017).

NHTSA (2016b) Federal automated vehicles policy. Available at: https://one.nhtsa.gov/nhtsa/av/pdf/Federal_Automated_Vehicles_Policy.pdf (accessed 31 March 2017).

NHTSA (2017) Office of defects investigation, investigation PE 16-007. Available at: https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF (accessed 31 March 2017).

Nissenbaum H (1994) Computing and accountability. *Communications of the ACM* 37(1): 72–81.

Nourbakhsh IR (2013) *Robot Futures*. Cambridge: MIT Press.
NTSB (National Transportation Safety Board) (2016) HWY16FH018 Preliminary Report Highway. Available at: https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY16FH018-Preliminary-Report.pdf (accessed 2 August 2017).

NTSB (2017a) HWY16FH018 Driver assistance system. Specialist's Factual Report. Available at:

https://dms.ntsb.gov/pubdms/search/document.cfm?docID=453441&docketID=59989&mkey=93548 (accessed 2 August 2017).

NTSB (2017b) HWY16FH018 Human performance factors group chairman's factual report. Attachment 4: Witness Interview. Available at: file:///Users/jackstilgoe/Downloads/604743.pdf (accessed 2 August 2017).

NTSB (2017c) Transcript of board meeting, 12 September 2017. Available at: https://recapd.com/w-91b39c/ (accessed 14 September 2017).

Nvidia (2017) Reading an AI car's mind: How NVIDIA's neural net makes decisions. Available at: https://blogs.nvidia.com/blog/2017/04/27/how-nvidias-neural-net-makes-decisions/ (accessed 2 August 2017).

Obama B (2016). Barack Obama: Self-driving, yes, but also safe. *Pittsburgh Post-Gazette*, 19 September. Available at: http://www.post-gazette.com/opinion/Op-Ed/2016/09/19/Barack-Obama-Self-driving-yes-but-also-safe/stories/201609200027 (accessed 12 March 12, 2017).

Owen, R. (2014). Solar radiation management and the governance of hubris. *Geoengineering of the Climate System* 38: 212.

Parson EA and Clark WC (1995) Sustainable development as social learning: Theoretical perspectives and practical challenges for the design of a research program. In Holling ES, Gunderson LH, and Light S (eds) *Barriers and Bridges to the Renewal of Ecosystems*. New York: Columbia University Press, 428–460.

Paquet, G. (2005). The new geo-governance: A baroque approach. University of Ottawa Press.

Pasquale F (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge: Harvard University Press.

Perrow C (1984) *Normal Accidents: Living with High Risk Systems*. New York, Basic Books.

Pinker S (1995) *The Language Instinct: The New Science of Language and Mind*. London: Penguin UK.

Puzzanghera J (2016) Fatal Tesla crash exposes lack of regulation over autopilot technology. *Los Angeles Times*. 1 July.

Ramsey M (2016). Mobileye ends partnership with Tesla. *The Wall Street Journal*. 26 July.

Ramsey J (2017) The way we talk about autonomy is a lie, and that's dangerous. *The Drive,* 8 March. Available at: http://www.thedrive.com/tech/7324/the-way-we-talk-about-autonomy-is-a-lie-and-thats-dangerous (accessed 17 March 2017).

RAND (2016) Autonomous vehicle technology: A guide for policymakers. Available at: http://www.rand.org/pubs/research_reports/RR443-2.html

Rayner S (2004) The novelty trap: Why does institutional learning about new technologies seem so difficult? *Industry and Higher Education* 18(6): 349–355.

Rayner S and Cantor R (1987) How fair is safe enough? The cultural approach to societal technology choice. *Risk Analysis* 7(1): 3–9.

Recode (2017) CES 2017: We liveblogged all the news and updates from Las Vegas. *Recode.* 6 January. Available at: http://www.recode.net/2017/1/3/14154122/ces-2017-liveblog-news-updates-las-vegas (accessed on 12 March 2017).

Redebo (2016) Mothra strikes back! Knocks me out of autopilot on a lonely stretch of road… *Reddit*. Available at: https://www.reddit.com/r/teslamotors/comments/4irtae/mothra_strikes_back_knocks_me_out_of_autopilot_on/ (accessed 12 March 2017).

Reed M, Evely AC, Cundill G, et al. (2010) What is social learning?. *Ecology and Society* 15(4): r1.

Renn O (1998) The role of risk perception for risk management. *Reliability Engineering & System Safety* 59(1): 49–62.

Reynolds K (2016) Testing (semi) autonomous cars with Tesla, Cadillac, Hyundai, and Mercedes. *Motor Trend*. 5 July. Available at: http://www.motortrend.com/news/testing-semi-autonomous-cars-tesla-cadillac-hyundai-mercedes/ (accessed 12 March 2017).

Rheinberger HJ (1997) *Toward a History of Epistemic Things. Synthesizing Proteins in the Test Tube.* Stanford: Stanford University Press.

Rip A (1986) Controversies as informal technology assessment. *Knowledge* 8(2): 349–371.

Scoltock J (2015) Driving the future. *Automotive Engineer.* Available at: http://ae-plus.com/features/driving-the-future/page:2 (accessed on 12 March 2017).

Shalev-Shwartz S and Shashua A (2016) On the sample complexity of end-to-end training vs. semantic abstraction training. *arXiv*:1604.06915.

Shashua A (2016) Autonomous car – what goes into sensing for autonomous driving? Mobileye. *YouTube*. 17 April. Available at: https://www.youtube.com/watch?v=GCMXXXmxG-I&feature=youtu.be&t=169 (viewed 12 March 2017).

Sheikh et al. v Tesla Motors (2017) Class action complaint submitted to United States district court northern district of California San Jose division. Available at: https://www.hbsslaw.com/uploads/case_downloads/tesla_ap2/teslaclassactioncomplaint.pdf (accessed 6 August 2017).

Shladover SE. (2009) Cooperative (rather than autonomous) vehicle-highway automation systems. *IEEE Intelligent Transportation Systems Magazine* 1(1): 10–19.

Shorrock S (2013) 'Human error' – The handicap of human factors, safety and justice. Available at: http://www.skybrary.aero/bookshelf/books/2566.pdf

State of California Department of Motor Vehicles (2016) Express terms, title 13, division 1, chapter 1, article 3.7 – autonomous vehicles. Available at: https://www.dmv.ca.gov/portal/wcm/connect/211897ae-c58a-4f28-a2b7-03cbe213e51d/avexpressterms_93016.pdf?MOD=AJPERES (accessed 12 March 2017).

Stayton E (2015) Driverless Dreams; Technological Narratives and the Shape of the Automated Car. MSc Thesis, Cambridge: MIT. Available at: https://dspace.mit.edu/handle/1721.1/97997

Stilgoe J, Owen R, and Macnaghten P (2013) Developing a framework for responsible innovation. *Research Policy* 42(9): 1568–1580.

Stilgoe J (2016) Geoengineering as collective experimentation. *Science and Engineering Ethics* 22(3): 851–869.

Surden H and Williams MA(2016) Technological opacity, predictability, and self-driving cars. *Cardozo Law Review* 38: 121.

Szerszynski B, Kearnes M, Macnaghten P, et al. (2013) Why solar radiation management geoengineering and democracy won't mix. *Environment and Planning A* 45(12): 2809–2816.

Tesla (2013) Tesla Model S achieves best safety rating of any car ever tested. 19 August. Available at: https://www.tesla.com/blog/tesla-model-s-achieves-best-safety-rating-any-car-ever-tested?redirect=no (accessed 12 March 2017).

Tesla (2015) Your Autopilot has arrived. 14 October. Available at: https://www.tesla.com/blog/your-autopilot-has-arrived (accessed 11 March 2017).

Tesla (2016a) All Tesla cars being produced now have full self-driving hardware. 19 October. Available at: https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware (accessed 11 March 2017).

Tesla (2016b) A tragic loss. 30 June. Available at: https://www.tesla.com/blog/tragic-loss (accessed 11 March 2017).

Tesla (2016c) Upgrading Autopilot: seeing the world in radar. 11 September. Available at: https://www.tesla.com/blog/upgrading-autopilot-seeing-world-radar (accessed 12 March 2017).

Tesla (2017a) Self-driving hardware on all cars. Available at: https://www.tesla.com/autopilot (accessed 12 March 2017).

Tesla (2017b) Letter to California Department of Motor Vehicles. Available at: https://www.dmv.ca.gov/portal/wcm/connect/f1873c87-4f21-4beb-b665-050cada6db7a/Tesla_disengage_report_2016.pdf?MOD=AJPERES (accessed 1 August 2017).

Toyota USA (2017) Toyota CES 2017 Live Stream | Toyota. *YouTube*. 4 January. Available at: https://www.youtube.com/watch?v=CFTa2IjMNwM&feature=youtu.be&t=20m34s (viewed 12 March 2017).

Traffic Crash Records (2016) Florida traffic crash report. Highway Safety & Motor Vehicles. Available at: http://documents.latimes.com/tesla-accident-report/ (accessed 12 March 2017).

Urmson C (2015) How a driverless car sees the road. *YouTube*. 26 June. Available at: https://www.youtube.com/watch?v=tiwVMrTLUWg (viewed 12 March 2017).

Vanderbilt T (2012) Let the robot drive: the autonomous car of the future is here. *Wired*. Available at: https://www.wired.com/2012/01/ff_autonomouscars/2/ (accessed 11 March 2017).

Vellido A, Martín-Guerrero JD and Lisboa PJ (2012) Making machine learning models interpretable. In *ESANN* (Vol. 12): 163–172.

Vinsel L (forthcoming) *Taming the American Idol: Cars, Risks, and Regulations*.

Wachter S, Mittelstadt B and Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law* 7(2): 76–99.

Walsh T (2016) Turing's red flag. *Communications of the ACM* 59(7): 34–37.

Waymo (2017) On the road to fully self-driving. Waymo Safety Report. Available at https://storage.googleapis.com/sdc-prod/v1/safety-report/waymo-safety-report-2017-10.pdf (Accessed 20 October 2017)

Webler T, Kastenholz H and Renn O (1995) Public participation in impact assessment: A social learning perspective. *Environmental Impact Assessment Review* 15(5): 443–463.

Weinberg AM (1972) Science and trans-science. *Minerva* 10(2): 209–222.

Westbrook J (2016) PSA: don't try to take highway off-ramps with Tesla's Autopilot just yet. Jalopnik. 30 September. Available at: http://jalopnik.com/psa-dont-try-to-take-highway-off-ramps-with-teslas-aut-1787206951 (accessed 12 March 2017).

Wetmore, J (2003) Driving the dream, The history and motivations behind 60 years of automated highway systems in America. *Automotive History Review* 7: 4–19.

Wiener EL (1977) Controlled flight into terrain accidents: System-induced errors. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 19(2): 171–181.

Winfield AF and Jirotka M (2017) The case for an ethical black box. In: *18th Conference Towards Autonomous Robotic Systems*: 262–273.

Winner L (1977) *Autonomous Technology: Technology-out-of-Control as a Theme in Political Thought*. Cambridge: MIT Press.

Winner L (1980) Do artifacts have politics? *Daedalus*: 121–136.

Wired (2016) Barack Obama, neural nets, self-driving cars, and the future of the world. *Wired*. 24 August. Available at: https://www.wired.com/2016/10/president-obama-mit-joi-ito-interview/ (accessed 12 March 2017).

Wynne B (1988) Unruly technology: Practical rules, impractical discourses and public understanding. *Social Studies of Science* 18(1): 147–167.

Wynne B (1989) Frameworks of rationality in risk management: towards the testing of naive sociology. *Environmental Threats: Social Sciences Approaches to Public Risk Perceptions*: 33–45.

Wynne B (1992) Risk and social learning: Reification to engagement. In: Krimsky S and Golding D (eds) *Social Theories of Risk*. Westport: Praeger, 275–300.

Wynne B (1993) Public uptake of science: A case for institutional reflexivity. *Public Understanding of Science* 2(4): 321–337.

Xkwizit (2016) Autopilot – note to drivers and Consumer Reports. *Tesla Motors Club*. 16 July. Available at: https://teslamotorsclub.com/tmc/threads/autopilot-note-to-drivers-and-consumer-reports.73715/ (accessed 12 March 2017).

Xrayvsn (2016) How Autopilot has added years to my life. *Tesla Motors Club*. 13 February. Available at: https://teslamotorsclub.com/tmc/threads/how-autopilot-has-added-years-to-my-life.62607/ (accessed 12 March 2017).

Yoshida J (2017) Uber crash exposes V2X politics. *EE Times*, 28 March Available at: http://www.eetimes.com/author.asp?section_id=36&doc_id=1331532 (accessed 12 August 2017).

You Y, Pan X, Wang Z, & Lu C (2017) Virtual to real reinforcement learning for autonomous driving. *arXiv*:1704.03952.

Author biography

Jack Stilgoe is senior lecturer in Science and Technology Studies at University College London. He is interested in the governance of emerging technologies such as

machine learning, geoengineering and nanotechnology. He teaches undergraduate and graduate courses in science and technology policy. He previously worked at the Royal Society and the think tank Demos, and continues to advise policymakers in the UK and Brussels. In 2016-17 he was a visiting researcher at the University of Colorado, Boulder. He co-edits the Guardian's Political Science blog.