Ethics and Information Technology (2021) 23:635–647
https://doi.org/10.1007/s10676-021-09602-1

ORIGINAL PAPER

How can we know a self-driving car is safe?

Jack Stilgoe[1]

Accepted: 18 June 2021 / Published online: 30 June 2021

Abstract
Self-driving cars promise solutions to some of the hazards of human driving but there are important questions about the safety of these new technologies. This paper takes a qualitative social science approach to the question 'how safe is safe enough?' Drawing on 50 interviews with people developing and researching self-driving cars, I describe two dominant narratives of safety. The first, safety-in-numbers, sees safety as a self-evident property of the technology and offers metrics in an attempt to reassure the public. The second approach, safety-by-design, starts with the challenge of safety assurance and sees the technology as intrinsically problematic. The first approach is concerned only with performance—what a self-driving system does. The second is also concerned with why systems do what they do and how they should be tested. Using insights from workshops with members of the public, I introduce a further concern that will define trustworthy self-driving cars: the intended and perceived purposes of a system. Engineers' safety assurances will have their credibility tested in public. 'How safe is safe enough?' prompts further questions: 'safe enough for what?' and 'safe enough for whom?'

Keywords Autonomous vehicles · Self-driving cars · Risk assessment · Governance · Public dialogue

Introduction

Foremost among the justifications offered for self-driving cars is that they will offer dramatic improvements in road safety. The promise is based on an assumption that the automation of driving, an activity prone to numerous human failings, will be possible in the short term thanks to rapid developments in artificial intelligence. If computers can take over the tasks of sensing and interpreting the world, predicting the behaviours of objects within it, planning a safe path and controlling a car's speed and direction, the idea is that human performance can be rapidly matched and then exceeded. A well-known public health catastrophe—more than a million global road deaths each year, a hundred per day in the US alone—provides a strong motivation for radical improvement, with technology offering powerful options. However, new technologies raise questions about safety as well as offering answers. If self-driving cars are to earn public trust, we should ask, at an early stage, how safety can be assured, demonstrated and improved over time. This challenge stretches beyond engineering (Koopman & Wagner, 2017). Questions of regulation and safety assurance have been given insufficient attention as self-driving car developers focus on demonstrating the technology's potential. There has been little research to find out what the public thinks about self-driving car safety. Postponing debates about safety presents hazards for the public and reputational risks for developers who may be undone by their own or others' recklessness.

In 2016, a Tesla that was in Autopilot mode crashed in Florida, killing its sole occupant. This offered a stark reminder that technologies that attempt to automate at least part of the job of driving were less safe than their proponents claimed. In their crash investigation report, the National Transportation Safety Board were eager to point out that this vehicle was not a self-driving car, even though the data extracted from the vehicle suggests that its owner was behaving as though it was one (see Stilgoe, 2018 for a discussion of this case and its implications). The NTSB went on to investigate other Tesla Autopilot crashes as well as a crash in Tempe, Arizona in March 2018 in which a self-driving car operated by Uber hit and killed a woman who was walking her bicycle across the road. These collisions, and the investigations that have followed, have revealed not just a carelessness among some developers, but also a lack of

✉ Jack Stilgoe
j.stilgoe@ucl.ac.uk

[1] University College London, Gower Street, London WC1E 6BT, UK

consensus about how to assess risk, and an absence of clear regulation or standards to govern the testing or approval of new self-driving technologies. These incidents remind us not just of a technology's limits, but also of the flaws of a mode of governance that leaves technology developers to their own devices.

The 2020 independent expert report for the European Commission on the ethics of connected and automated vehicles (CAVs) (European Commission, 2020) recommended that the technology should reduce overall risk, be designed to prevent unsafe use and have clear standards for testing on public roads. These principles offer a strong regulatory ideal, but any approach to governance must engage with a political and economic reality. Self-driving car companies talk about a race to develop the technology. Governments' enthusiasm for innovation has seen them buy into this story, which has meant lax governance regimes.

The question of how we can know a self-driving car is safe is complicated. It depends on assumptions about how safe is safe enough, who needs to be persuaded and what constitutes a self-driving car. This paper explores these qualitative aspects of risk using qualitative data from interviews and workshops with members of the public. My team and I conducted 50 interviews with self-driving car developers, researchers and policymakers in the UK, US and Europe as part of the "Driverless Futures?" project. The interviews lasted between 30 and 90 minutes and took place between 2019 and 2021. The aim of the interviews was to go beneath public accounts of the benefits and risks of self-driving cars, the hypothesis being that the people closest to research and development would have a clearer sense of the uncertainties, complexities and contingencies of the technology (following MacKenzie, 1998) and would be able to articulate these during long interviews. Interview quotes are anonymised here using numbers. In addition, I draw on transcribed conversations from a large public dialogue exercise that took place in 2018 and 2019, commissioned by the UK Department for Transport and Sciencewise. The CAV public acceptability dialogue was the world's first substantial attempt at deliberation designed to inform policy for CAVs. It involved 150 public participants in five locations, over three weekends, informed by expert visitors. Participants were recruited to reflect the diversity of the UK population. I was part of the team designing, facilitating and reporting on the process.[1] This research reveals the diversity of understandings from people inside and outside the community of innovators. While the discussion is currently dominated by engineers, there is a clear need to include perspectives from other stakeholders and members of the public.

How safe is safe enough?

In 1969, Chauncey Starr asked how, in weighing the benefits and costs of new technologies, we might consider variations in the acceptance of different types of risk. His question, "How safe is safe enough?", prompted consideration of the dimensions of safety that couldn't be captured by a calculus of probabilities and outcomes (Starr, 1969). The 1970s and 80s saw growing interest in research on risk perceptions that examined the importance of psychological biases and heuristics in explaining individuals' attitudes to risk (e.g. Slovic, 1987). Risks that were seen as new, uncontrolled, catastrophic and artificial were found to be consistently exaggerated.

For transport, we see markedly different societal assessments of acceptable risk. The risks of many transport systems are by now well known, but the evolution of law, regulation, technology and culture suggests that people are much less willing to accept risks from modes of transport they regard as highly centralised and out of their control. UK train operators are willing to spend, according to one estimate, tens of millions of pounds per life saved on the railways (Wolff, 2006); meanwhile, there is chronic underinvestment in affordable and available technologies for car safety (Vinsel, 2019), even though cars are far more dangerous.

Revealed risk preferences are observable in hindsight. However, formal risk assessment demands the calculation of outcomes and probabilities, both of which will be uncertain for new technologies. Alvin Weinberg, a Cold War physicist and nuclear energy enthusiast, described the difficulty of risk assessment for rare, catastrophic events in complex technologies, such as a nuclear accidents:

> Because the probability is so small, there is no practical possibility of determining this failure rate directly - i.e., by building, let us say, 1000 reactors, operating them for 10,000 years and tabulating their operating histories. (Weinberg, 1972, p.211).

Social scientists (Stirling, 2007) (Funtowicz & Ravetz, 1993) have analysed the limits of conventional modes of science and policymaking in conditions of uncertainty or ignorance and the tendency to treat incalculable uncertainties as controllable risks. To use Hansson's (2009) analogy, the tools of risk assessment trick us into believing we are in an environment like a casino, where risks are known, when we are actually surrounded by a thick jungle of unknown and possibly unknowable hazards. The development of new technologies, from this view, is a form of experiment whose variables and metrics cannot be well-defined in advance (Krohn & Weyer, 1994) (van de Poel, 2016).

[1] The full Sciencewise report is available here https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/951094/cav-public-acceptability-dialogue-engagement.pdf, accessed 25 March 2021.

Starr's challenge and subsequent research into risk perception are limited by a one-dimensional view of risk. An alternative research agenda that symmetrically problematises new technologies as well as public responses to them has seen risk as multidimensional. The politics of new technologies mean that questions of risk may be unavoidably bound up in questions of equity or freedom. The assessment of risk, even though it is often discussed as a value-neutral activity, is political and ethical as well as scientific (Rayner & Cantor, 1987) (Irwin, 1985).

The language of technology governance tends to presume a separation and a sequencing between scientific risk assessment and risk management. Risk assessment is assumed to be scientific and risk management is where questions of trust, acceptability, uncertainty and politics come in. The public, it is assumed, only have an interest in the management of risks. The rise of research on risk perception has hardened rather than blurred this boundary. Engineers' appreciation for public views on risk cemented the view that theirs was the correct assessment, and that it was the divergences from this that needed social science explanation. If we pay attention to the framing assumptions of risk assessment, we can however see the limits of such a model both empirically and politically (Stirling, 2007). The assumptions behind risk assessment are revealed to be brittle when tested in terms of public credibility.

In the face of social and technological uncertainties, different groups will seek to draw parallels and precedents that either problematise or downplay novelty. Proponents of genetically modified crops, for example, sought to argue that the technology was 'substantially equivalent' to its conventionally bred counterparts, while advocates for precautionary regulation highlighted novelties and uncertainties (Millstone et al., 1999). A successful innovator must finely balance claims of novel benefits with reassurance that their technologies do not require radical regulatory attention (Rayner, 2004).

Performance, assurance and reassurance

As sociotechnical systems have become more complex, more dependent on digital and automated technologies and more privatised (Leveson, 2004), regulators have sought to make their rules less prescriptive and more 'performance-based' (Gann et al., 1998). The idea is to give innovators an end goal rather than tell them how to get there. The hope is that this approach encourages innovation and allows for more focussed regulation. The model presumes, first, that we are clear on how to judge performance and, second, that the public has no interest in what is going on behind the scenes. It represents a way of knowing as well as a way of governing: a presumption that innovators know best and will be able to account for the public interest. According to one study of the recent crashes of Boeing 737 Max aeroplanes, if a performance-based approach is going to encourage safety, it should resist simple metrics and have independent verification (Sgobba, 2019). Without external scrutiny, performance-based approaches look less like genuine safety assurance and more like naïve attempts at public reassurance.

Unlike some technologies, developed in a laboratory before being released into the world, self-driving cars are being developed in public. Their developers have therefore been compelled to build public stories of their safety that they hope will be sufficient to secure a social license to operate on public roads. When the technology was brand new, these stories reflected a 'technological sublime' (Nye, 1996). As the technology has become entangled in real-world complexity, the stories have been modulated in response to others' questions and concerns (Tennant and Stilgoe, in press). The stories provide first drafts for risk assessment that could become hugely consequential. From my interviews and publicly available sources, I have extracted two competing narratives. The first, safety-in-numbers, starts by presuming self-driving cars are a solution to a perennial safety problem and looks for metrics to show progress. The second, safety-by-design, starts with the question of safety assurance and problematises self-driving innovation.

Safety in numbers

In April 2020, during a conference presentation on 'AI for full self-driving' Tesla's senior director of artificial intelligence announced that the company's cars had driven 3 billion miles on Autopilot. Autopilot is a limited automated system, but the number is meant to reinforce the impression that a self-driving Tesla is just around the corner. The claim is twofold: first, that Tesla are harvesting more data than their competitors and, second, that their system has a track record of safety.

The safety-in-numbers narrative starts with a simple calculation: we know the risks of human driving; self-driving aims to eradicate that risk; therefore, as long as the technology works and there are numbers to show it works, there will be safety improvements. The stated justification for developing self-driving systems is to solve a problem of safety; the system's adequate performance is therefore a demonstration of safety. This approach ignores the question of who needs to be convinced; the developers' own assessment of performance is the relevant criterion. It's an approach that has defined self-driving car development since Google first funded its self-driving car project in 2009. Google's engineers were given a target: if they were able to clock

up 101,000 self-driving miles hiding in plain sight on California's roads, they would receive large financial bonuses. By the time these secret tests were publicised, they were almost complete. The New Yorker later revealed that the company's self-driving cars had been involved in multiple incidents during this time, but there were no laws compelling the company to report them. The response from the company, which had by then been spun out from Google and renamed Waymo, to the New Yorker revelations is telling:

> The Google self-driving car project was founded with a mission to improve road safety, and that's the standard we hold ourselves to in everything we do. Over the past near-decade, we've carefully developed a comprehensive testing program that includes more than 10 million miles on public roads.[2]

Waymo's claim rests on a statistic of number of miles driven without a death or serious injury. This superficial metrics demands further analysis. A series of reports from RAND (Kalra and Paddock, 2016) (Kalra & Groves, 2017) (Fraade-Blanar et al., 2018) (Blumenthal et al., 2020) have taken on the question of measuring safety. The first concludes that, if self-driving cars are to demonstrate improved average safety over human driving, they would have to rack up 275 million miles without a mistake. Their conclusion is that "developers of this technology and third-party testers cannot drive their way to safety" (Kalra and Paddock, 2016, p. 3). However, the assumption behind the RAND reports is that average improvements in safety still justify the rapid deployment of self-driving cars (Kalra & Groves, 2017). In the search for measures that might be both useful and publicly persuasive, RAND popularised a distinction between leading and lagging metrics in a report that was commissioned by Uber (Fraade-Blanar et al., 2018). The argument is that lagging metrics of outcomes might be easy to measure, but leading metrics, including the number and type of minor transgressions a self-driving car makes, might allow for the prediction of performance. Judging self-driving performance by number of fatalities per million miles driven might be possible after substantial experience of the technology, but this information is useless in regulatory terms and a poor indicator of performance. Humans move on from the embarrassing everyday near misses that characterise our imperfect driving. Self-driving cars and their regulators can and should learn from the crashes that don't quite happen as well as those that do.[3]

From the safety-in-numbers viewpoint, the technology's safety benefits are jeopardised by irrational public risk perceptions that mean we underestimate the safety of modes of transport that we presume to control, such as driving, while overestimating the risks of systems that are out of our control and seem uncanny, such as flying. Some early research (Liu et al., 2019) suggests that a sizeable proportion of the public wants self-driving cars to be at least a hundred times safer than conventional cars.

Interviews with self-driving technology developers and researchers provide an opportunity to get beneath the superficial story of safety and probe some of the claims being made. The public narrative of how self-driving cars 'work' hides a broad range of views even among those trying to get the technology to 'work'. Companies are adopting diverse strategies, with some emphasising safety and responsibility while others, particularly smaller start-ups, find it hard to divert core engineering resources to address safety assurance. The most optimistic enthusiasts for self-driving see self-evident safety benefits, meaning that public persuasion becomes just an extension of engineering. One of my interviewees, a leading artificial intelligence researcher, argued that the statistics would inevitably force the hands of regulators:

> At some point in the near future, it's hard to predict when… you will have [self-driving] cars that are maybe, on average, ten times safer than humans. It will be three times, then five times, then ten times safer. It's a matter of statistics… I'm not sure whether the factor of ten is sufficient, maybe you need a factor of 100, but at some point they're going to be mandatory. (Interview 1)

Some interviewees entertained a consequentialist argument that there might be short term hazards from a technology under development, but the long-term safety benefits would justify the means. Other researchers engaged with the reality of public risk perceptions. One concluded that, even if average safety improvements were unarguable, "That's a hard one to deal with when it's your child that got run over by the vehicle" (2). The recognition here is that the technology would change the qualitative as well as the quantitative aspects of safety, making questions of responsibility inescapable. Other interviewees referred to "algorithm aversion" (3), a hypothesis that members of the public might exaggerate the hazards of automated systems.[4]

[2] 'A Google self-driving car reportedly caused a crash in 2011 after a former engineer changed its code to drive where it wasn't supposed to', Sean Wolfe, Business Insider Oct 20, 2018 https://www.businessinsider.com/anthony-levandowski-google-self-driving-car-crash-2018-10?r=US&IR=T.

[3] Others have pointed out the limits of KSI (killed or seriously injured) statistics in determining safety (Ryerson et al., 2021).

[4] One psychology paper (Shariff, Bonnefon and Rahwan, 2017) recommends that innovators and regulators should manage "public overreaction with 'fear placebos' and information about actual risk levels." This paper shares the view of some of our interviewees that public perceptions represent "psychological roadblocks" to inevitable adoption.

While constructing numbers that they hope will offer reassurance, self-driving companies have sought to normalise the technology with public displays of flawless driving. YouTube is replete with companies' demonstrations of the technology working, but these videos only report on success, offering little assurance on the technology's limits. Even though self-driving car developers, as one interviewee put it, "need to be out on the road… racking up the miles" (4), they know that this costs money and that not all miles are equal. One interviewee from a large car company criticised self-driving car start-ups who

> drive for millions of miles to prove a point… it's endless… We have to let go of that. Three or four years ago that used to be the criterion: How many miles have you driven? That's not the issue any more… if I'm driving a hundred miles between one intersection and another one – a straight, simple road – the fact that I drove a hundred miles is insignificant. (5).

Another researcher argued "You could drive up and down the Nevada desert. A hundred million miles. It doesn't help me if I'm going to use it in London" (6). When Google's engineers were given their target, the company recognised the variability of driving: 100,000 miles could be ticked off on North California's easy roads, but there were also 10 predefined routes on more challenging terrain, including Lombard Street, known as "the crookedest street in the world".

While engineers recognise the qualitative variation, the pure numbers remain seductive. Waymo continues to announce milestones of incident-free distance. Another company's CEO has argued that progress could be measured out by increases in the number of "miles per disengagement" (a disengagement is a moment of system failure).[5] One interviewee, an investor in self-driving companies, ran with this idea of a "Moore's law for self-driving vehicles", starting with the rough calculation that human driving in the US produces a fatal crash every hundred million miles:

> How long will it take to get to one disengagement every hundred million miles?... 15, 16 years, something like that… We're not going to tolerate machines killing people at the rate of 40,000 a year in the United States. So they've got to be maybe an order of magnitude more safe. Add another order of magnitude to that? That's sort of the timeline. (7)

One British self-driving company, seeking to emphasise its responsible approach to safety assurance, has taken issue with the "disengagement myth":

> It's now clear to everyone that simply measuring progress as improvements in miles between disengagements hides many failures that might not bubble up to the level of disengagement, whilst at the same time enforcing an extremely slow development cycle. That's not to mention the need to physically drive hundreds of millions of miles to be statistically confident.[6]

The continued presence of simple statistics in the public debate even while engineers agree that they are flawed is an echo of self-driving's origin myth. The technology's feasibility, according to this story, is enabled by recent and rapid advances in artificial intelligence. The technology therefore takes its inspiration not from other mobility technologies, but from technologies like machine translation, which requires little linguistic expertise, relying instead on what the researchers behind Google Translate called "the Unreasonable Effectiveness of Data" (Halevy et al., 2009). Machine learning, at its root, is statistical. The hope is that, with enough data, performance can become superhuman even if the way machine learning works is utterly unlike human learning and is usually opaque (Burrell, 2016).

The approach to self-driving that prioritises data for machine learning is sometimes called "brute force", but the question of safety has proven hard to force. Some interviewees talked about getting the technology to work safely in terms of percentages:

> I think we're doing a pretty good job with technology. It is really close to actually working. it's always tough to get the last one or 2% out of these things, it's easy to do 80%... 20 is hard. The last 2 is really hard. The last 0.2 is really, really hard. (8)

An interviewee who once ran a self-driving start-up concluded that "scaling safety is going to be so hard and take so long" (9). As discussed below, an alliance between probabilistic machine learning and probabilistic risk assessment will also struggle to achieve public credibility.

Self-driving car developers, some of whom have switched over a decade from regarding the technology as impossible to seeing it as inevitable, now find themselves asymptoting towards an ideal of safety that may always be out of reach. More data and more miles produce better systems, but they also reveal more 'edge cases'—circumstances that the model cannot account for. Engineers recount the unusual

[5] 'The Moore's Law for Self-Driving Vehicles', Edwin Olson, Feb 27, 2019 https://medium.com/may-mobility/the-moores-law-for-self-driving-vehicles-b78b8861e184, accessed 1 March 2021.

[6] 'Laying Out the Challenges in AI Safety', Five AI, Jun 4 2020 https://medium.com/fiveai/laying-out-the-challenges-in-ai-safety-9f51f91107ea.

things—balloons, ducks, wheelchairs, kangaroos—that their sensors have seen but which their software has struggled to make sense of. The sheer complexity of their challenge can lead to frustration. Interviewees often expressed disdain for unruly pedestrians or poorly maintained roads. Predicting the movements of pedestrians is impossible with certainty, but pedestrians are unavoidable. And yet, obviously, they must be avoided. Developers' usual answer to this challenge is that it will be met with more data from which the system can learn.

As they confront the challenge of safety, some admit the impossibility of perfection. (One interviewee said, "I hate perfection, because I know I can't attain it" (8)). But all developers will remain troubled by the normal abnormalities that exist in a world designed by and for humans, whose autonomy and mobility in the environments that self-driving cars seek to occupy is, for now, relatively unconstrained. As one safety engineer explained, complex systems are impossible to describe with one probabilistic risk assessment:

> If you don't understand the design of this system, how do you know that number's right?… We're trying to solve problems that are actually impossible. We throw numbers at them, we almost make them up, but they don't apply to that design. (10)

This engineer points to a gap between the numbers that are available and the numbers that regulators might need to assess a system. The data that are of interest to developers as they seek to get their technology to work may not be relevant for safety assurance. Others may have a very different sense of what it means for the technology to 'work'.

The safety-in-numbers narrative is superficially impressive, especially when weighed against the known risks of driving and as long as self-driving cars aren't implicated in high-profile crashes. It fits a prevailing regulatory assumption that what a system does is more important than why it does it, and that we should trust innovators to show us what they can do. The limitations of the narrative become more apparent when, rather than taking safety as self-evident, we start with the challenge of designing safe systems.

Safety by design

While most self-driving car developers focus on improvements in AI and demonstrating safety through performance, there are engineers emphasising safety-by-design who are more likely to have had experience with hardware, requirements engineering and safety assurance of other complex systems. Looking at a prototype self-driving car, they see a potentially lethal safety–critical system; a heavy robot travelling at speed in an uncertain, uncontrolled environment. From this standpoint, the challenge of safety assurance looks daunting. This group draws attention to issues that they see as important but neglected in the simplistic safety-in-numbers story. These issues, discussed below, include human–machine interaction, system safety, redundancies, interpretability and simulations.

The self-driving ideal removes the human from responsibility, if not always from the driving seat, but engineers with experience of humans in-the-loop know that people can never be completely automated out of sociotechnical systems (Bainbridge, 1983; Mindell, 2015). Safety engineers now have decades of experience with aeroplane autopilots and other systems involving human–machine interaction (Cummings & Thornburg, 2011). They have warned about the risks of 'mode confusion', 'skill detriment' and handovers between human and machine responsibility, issues that self-driving car developers are now coming to terms with, sometimes in reckless ways (Stilgoe, 2018). Some regard such issues as temporary, worthy of attention while prototype systems are being developed, overseen by safety drivers who are expected to take control in the event of a technological failure. But other engineers have called for users to be a permanent part of a new approach to 'informed safety' (Khastgir, 2018). As Lisanne Bainbridge argued almost 40 years ago, "there will always be substantial human interaction and involvement with automated systems" (Bainbridge, 1983). Even if the humans in control of a vehicle are completely reliable, interactions with humans outside the vehicle multiply the complexity, and cannot be easily engineered away. One software engineer interviewee said, "in the context of self-driving cars, something that we don't yet know how to do is handle the humans in-the-loop and interaction with human-driven cars" (11). These interactions, which, for a human, define driving, must be reinterpreted by engineers to become amenable to a technological fix.

The assessment of a vehicle's risk necessarily involves more that the vehicle itself. The vehicle's context, in engineering terms, is sometimes called an 'operational design domain' (ODD). The ODD represents the conditions in which a self-driving car can reliably operate, and may include material features like road types, weather, other road users and infrastructure as well as digital systems like high-definition maps (which need to be constantly updated as the environment changes) and communication between vehicles and the outside world. Many of these bits will be outside the control of a self-driving car company. One interviewee discussed the necessity of

> narrowing the ODD… The idea that you'll be able to flip a switch in a Tesla and it'll drive you anywhere there's a road is in my mind fantasy… If you can restrict an ODD… you can characterize the types of interactions that the vehicle is more likely to encounter. You're narrowing this whole available pool of scenarios to something smaller. (3)

In practical terms, this might mean 'geofencing' a vehicle to prevent it from straying into spaces that are too unpredictable, or it might mean changing the outside world to make a particular domain operational by, for example, restricting the movements of other road users or upgrading infrastructure. One engineer, discussing so-called 'smart infrastructure' that would be able to communicate directly with a vehicle, concluded "there are some things where it's so difficult to be able to assure safety without the infrastructure helping" (12). Another said there would be a "need to instrument the environment, for example, for self-driving cars, so that they can read traffic signs maybe automatically" (11). An interviewee from a company trialling the technology wondered, "are we going to have to tell people not to walk out in front of cars? I think we might" (13). Even setting aside the political ramifications of reshaping the outside world to suit a new technology, one can see the complexity of a safety-first view that sees risk as a product of complex systems rather than individual machines.

For a safety engineer, the challenge might seem intractable. Innovators' emphasis on getting their technology to work obscures consideration of what to do when it doesn't. Some interviewees defended their technologies by pointing to redundancies and fail-safes. One self-driving car developer claimed that their company maps the environment because "we want to know in advance to expect traffic lights to be in a certain position. That's what gives us the redundancy in our solution. That's what makes us confident that we're not missing things." (14) But a safety engineer wondered "when the RSU ['roadside unit', for communication between vehicles and infrastructure] fails, what's my back up?". This interviewee said there was a need to "build in redundancy and diversity" (12).

Some engineers were particularly troubled by a dependence on AI systems that they regard as opaque and brittle. Asked about the challenge of understanding why a machine learning system does what it does, one engineer responded,

> Companies out there who sell GPUs… [graphics processing units—a type of chip used for training neural networks] claim that you can do everything inside the car. They will never build a car. They will never take the responsibility of cars driving automatically outside in the world… end-to-end learning? Having a neural network which takes in camera data, Lidar data, radar data and then operates the brake and the steering wheel? That's a nice showcase… but it will never happen on public roads… a system which we bring to the road always needs to be 100% deterministic... if you say, 'Well, I don't know what happened, there's a deep neural network', that won't work… it needs to be completely deterministic. (15)

The end-to-end learning referred to by this interviewee is an AI approach in which one model—a neural network—learns how to turn inputs such as the images from a camera into outputs such as turning the steering wheel. This interviewee worried that such systems were usually black boxes, and "engineering is pure responsibility", which meant the need for "somebody inside the company who signs off the system and is personally responsible." This person would not be able to "sign off a deep neural network" (15) whose decisions were uninterpretable and non-deterministic. One AI specialist in a self-driving start-up offered a counter-argument:

> Vaccines are not deterministic. You don't understand them. There is no transparency. There is no determinism. You use it based on the statistical guarantee that they give more value than damage, so I don't think it's a good argument against machine learning. (16)

Another machine learning advocate compared the interpretability challenge with a massive software package like Microsoft Windows:

> Millions and millions of lines of code are in that software package. Is that really explainable?... You can't understand really why certain decisions will be made. It's the same with neural networks… We should really look at the tests that are used to assess the software, not making sure it's human readable. (17)

This interviewee did however see the value of explainable AI as a PR tool:

> Every accident that an AV [autonomous vehicle] is going to be involved in, there's going to be press articles about it. Anything we can use to debunk any incorrect information I think would be very beneficial. (17)

The indeterminacy of both the worlds in which self-driving cars operate and the software that drives their decision-making persuades some engineers of the need for a fundamental rethink of the approach to self-driving that seeks a like-for-like replacement between a human driver and a computer:

> We shouldn't be replicating a human, who has such a lousy ability to perceive time and space and speed… We can actually define how the vehicle should respond so it will be safe. I don't want to mimic a human's faulty decision making. I can do it safer than it's done today. By this, we can save lives. (5)

This interviewee's design view encompasses the whole system—roads, other road users and more—rather than just

a vehicle (cf Blyth, 2016). Others agreed that the safety challenge could only be met by rethinking whole systems. Interviewees involved in self-driving start-ups were more likely to argue that if new technologies were unable to meet established standards of safety assurance, those standards would need updating. Some companies have proposed rules that they hope will guarantee safe driving while reassuring a sceptical public.[7] These rules are egocentric; they are about protecting the vehicle and, crucially, protecting its makers from blame, rather than designing for collective safety. However, as with other self-driving innovations, these rules still require real-world testing.

Human drivers must, in most jurisdictions, pass a driving test that certifies a certain level of aptitude, the assumption being that skills will be transferable across different driving conditions. In addition, their vehicles need to be certified as roadworthy. A self-driving test would blend elements of both tests, and would be confronted with some of the challenges discussed above: How could we test for a potentially infinite variety of edge cases? Would a self-driving car's capabilities be as adaptable as a human's? How could we avoid technology developers 'teaching to the test'? Would a licence be localised or portable to other ODDs? Would certification still count after a software upgrade? A self-driving test could, like its commonplace counterpart, combine a set of scenarios.[8] These could be examined at a test track, on public roads or in computer simulation. Some developers would see such tests as straightforward extensions of what they are doing anyway to verify their driving algorithms. Simulation has quickly become a vital tool for risk averse engineers. (One developer argued, rather dramatically, "we should be killing hundreds of millions of kids in simulation" (18)). But a test might only postpone the question of assurance. A safety engineer wondered:

> You've closed the world to build the simulation… When you go from the simulation to the real world, do any of the things that you've left out in building the simulation really matter? (12)

One developer argued that this meant a need to understand the internal processes of a vehicle's decisionmaking rather than just its performance:

> There is no world in which, as we develop this technology, we can actually test all of the possible permutations of things, so we have to understand at some logical level how it's understanding, interpreting the environment and making decisions so that we know that the methods that we are using to test it have appropriate coverage (19)

An AI researcher speculated on what a test for AI driving the real world would look like:

> Maybe you could have a simulated test… a billion miles in simulation? If it has fewer than ten accidents, you say 'you're good to go'. And it could be a completely uninterpretable one, as long as we have confidence in the test, which we don't yet, but maybe we could someday. Yeah, maybe that'd be fine. (20)

But one developer took issue with "people trying to introduce digital driving tests and rules about having to test autonomy in somebody else's simulator" before the technology 'works': "test companies and test facilities, test processes, they're talking about validation, certification, verification, digital driving tests, digital MOTs. We don't have anything for them to test yet" (14).

As self-driving car excitement, investment, testing and development have expanded, the initially straightforward narrative of safety has been troubled by the perspectives of others, some of whom have become enrolled in the technology and some who are watching from the sidelines. It is unclear whether these perspectives can constructively mesh or whether they are incompatible.

Safety first; safety last

In discussions surrounding standards, tests and possible regulations, there is a clash of cultures. For one group, safety is a self-evident property of the technology. The challenge is therefore one of public reassurance. For the other, safety is a vital design criterion, supported by extensive domain expertise. The former would be sceptical of the latter, suspecting they are conservatively defending their own incumbency. The latter group's response would be that the upstarts are creating real and reputational risks through irresponsibility. One safety engineer was critical of "start-ups who don't have to lose anything. They only have to win venture capital" (15) and a former self-driving CEO argued.

> Very few Silicon Valley companies have ever had to ship a safety critical thing… Google seem to think that if they code a program well enough, no one will ever die. That's just not how the world works. That's not how you build a safety protocol system… The lack of maturity about that has really hurt the industry overall. (9)

One interviewee criticised "a lax attitude… within the industry to safety" (11). Others were more diplomatic:

[7] Examples include Intel/Mobileye's Responsibility Sensitive Safety, Aptiv's Rulebooks and Nvidia's Safety Force Field.

[8] The German government's PEGASUS project is, at the time of writing, working to develop a set of scenarios for verification and validation https://www.pegasusprojekt.de/files/tmpl/Pegasus-Abschlussveranstaltung/PEGASUS-Gesamtmethode.pdf (Weber et al., 2019).

"I'm not sure I'd be rude to the AI people but often all of them working in this area don't understand a lot of standard safety engineering" (12). On the question of whether AI can explain its actions, one AI researcher was critical of colleagues for encouraging "a culture of people building uninterpretable models and getting paid a lot of money for that" (20). Viewed optimistically, this antagonism within and between disciplines might be constructive, destabilising assumptions and leading to more robust systems. One software researcher who works on self-driving systems admitted that, after taking his code into the world,

> I now understand the safety question in more detail… There are a lot of easy things that people say in machine learning that aren't really true…. Speak to any machine learning researcher today and they say, well, you just have to get more and more data and everything is done… There is an implicit assumption in everything that actually getting more data is easy… And then people say, 'Well, we can do it in simulation'. But this is a chicken and egg problem, because then how do you make the simulation good enough for it to be useful? (21)

Notwithstanding this interviewee's politeness, the two cultures and their approaches to safety need work to improve their compatibility, and there is a danger that the momentum of the safety-in-numbers approach railroads the other. Some interviewees' reflections on culture were prompted by other companies' early missteps. The NTSB's report on the Uber crash points not just to technological flaws, but also to a woeful safety culture (Stilgoe, 2019). In this case, engineers' desire to demonstrate their success led to what Diane Vaughan (1996) calls a 'normalization of deviance'. The risks were a product as much of economics as of technology. The Uber crash also revealed that the governance of trials was, in some places, threadbare. As technologies are being tested, gaps and disagreements between the safety-in-numbers and a safety-by-design approaches are papered over by safety cases aimed at reassuring local authorities that uncertainties are under control.

Safety case are documents put together by an organisation that attempt to persuade a regulatory authority that their system is acceptably safe.[9] The argument for safety cases is that they should inculcate safety by making a company redesign its operations from the bottom up. Most safety case approaches are performance-based rather than looking to open technological black boxes (Leveson, 2011). The presumption is that the organisations developing the technology are best placed to identify issues and that they are willing and able to regulate themselves. Early attempts at safety cases for self-driving are criticised by some safety engineers because, as one interviewee put it, "they're very focused on the goal. They're showing all of the arguments why they believe their goal is met. That is a positive-oriented argument." (10). An alternative would be to "start by assuming this thing is going to lead to a loss of life… regardless of what you think, let's start by assuming there's something wrong with it. We're going to find everything that's wrong with this" (10). Such a philosophy would represent a dramatic shift in the burden of proof.

In the short term, many safety cases rest on a safety driver, a human on-the-loop who is behind the wheel and, in principle, able to compensate for the technology's shortcomings. The safety driver acts as scaffolding while the software is under construction. The danger is that, without external safety assurance, the scaffolding could be removed by a self-confident technology developer and the safety cases that govern testing could become de facto rules of the road.

Although technologies are still experimental and legal frameworks unclear, a consensus is crystallising around safety cases and other performance-based approaches. However, as I will discuss in the final section, a premature lock-in to this mode of regulation would foreclose more prescriptive approaches that may be more publicly credible. The claims of engineers need exposure to public attitudes in order to appreciate the multidimensionality of self-driving safety.

Safety and the public

As engineers from different standpoints attempt to persuade themselves and each other that their approaches to safety are good enough, they are imagining who the public are, what the public think and what the public want. Engineers' private effort to know about safety is extrapolated into a public project of reassurance. The role imagined for the public is, at present, an exceptionally narrow one. Some engineers recognise that people may disagree on levels of acceptable risk, but the presumption is that the public should be kept out of risk assessment:

> The complexity that you need to go into is not going to be of any benefit to the public… people aren't really interested in that. I think with the public, you can only really prove safety through experience. (17)

The widespread assumption is that non-experts should not know or care how a system works or why it does what it does. Interviewees described non-experts as prone to biases, including "algorithm aversion", that would skew their risk perceptions. But the more pragmatic safety engineers saw such perceptions as unignorable. Ultimately, regulation

[9] Most self-driving car companies have published safety cases. 'Safety first for automated driving' is a notable collaboration between multiple partners, aiming to lay the groundwork for future standards https://www.daimler.com/documents/innovation/other/safety-first-for-automated-driving.pdf, accessed 7 June 2021.

would be a public matter. One software developer said "what I as an engineer give society is the knob [to balance safety against efficiency]… and I also clearly explain the trade-off" (16). The question is whether society will be content with just dialling up or down risk, or whether there will also be a legitimate public interest in the process through which self-driving cars are developed and assessed.

Our workshops with members of the public revealed a set of complexities that suggest a simple narrative of the technology as a safety solution will lack credibility. People are sceptical that a technology as novel as a self-driving car on public roads can be guaranteed safe. During the public dialogue exercise, participants were quick to highlight complexities and identify what engineers would call 'edge cases'. (One participant mentioned a recent encounter his wife's guide dog had had with a Starship delivery robot in Milton Keynes that had confused both dog and robot). The discussions reflected an ambivalence typical of public attitudes to technology (Kearnes & Wynne, 2007): excitement about the benefits coupled to a concern about the technology's limits and its governance.

The workshop participants were not naïve about the hazards of human-driven cars, and many were optimistic about possible benefits of self-driving, particularly for disabled people, but they were not convinced that weighing these qualities, many of which were highly uncertain, was straightforward. For some, the question of safety was far broader than just road safety. One woman said, "I'm concerned about travelling as a woman in shared [driverless] rides. How can I guarantee that I'll be safe late at night?" Another participant concluded "There will always be vulnerability in technological systems… I wouldn't get in a plane without a pilot." They were used to their computer software crashing, but expected far more from other technologies. They were aware that cars were highly regulated and tested before they left the factory. After one participant expressed concern that self-driving could be regulated more lightly, like computer software, another participant sought to reassure her:

> Cars are different. No one is going to allow a car on the road till it's [proven] that nothing will happen…. whoever is designing these cars and is moving this technology forward has thought about all these things. They haven't put a car on the road and just hoped for the best.

This participant's faith in automotive regulation suggests that self-driving car developers will need to work hard to earn similar levels of public trust. Many were concerned about what Wynne (1983) calls 'social risks', to jobs, businesses and local communities from a rapid introduction of new technologies. Some were worried that the technology could displace public transport. Others worried that their freedom to drive would be at risk in the long term. In the short term, recognising that things would go wrong, they saw risks in the context of responsibilities, as these comments from participants indicate:

> There will be risks, we will learn from accidents, but I don't want my family to be those on the back of which the learning happens.
> If the system fails overall, then someone needs to be accountable for a backup system.
> The algorithms will be written by humans, so humans have some responsibility.

Most of the participants agreed that it would be unwise to leave safety to the market. They saw a need for oversight. In the groups' final sessions, they discussed their messages for Government. Across the five locations, their support for the technology was broadly conditional on the following factors:

- If the technology is proven to be safe and secure
- If the benefits of the technology are widely available
- If the technology is good for society and jobs
- If we're in control of our transport
- If there is clear guidance on accountability
- If new regulatory bodies are created[10]

These early insights from public workshops give a sense of the public credibility challenge. People see safety as highly contextualised, entangling questions of science and technology with those of values. The social assessment of safety is not just a question of how safe is safe enough. It is also linked to the question of what the technology is for and who is seen as benefitting from its development. People engage with safety issues from multiple perspectives: as potential users of the technology, drivers, pedestrians, public transport users, parents and citizens.

The safety-in-numbers story treats safety as an ex-post destination, while the safety-by-design approach sees safety as an ex-ante starting point. In discussing the technology's uncertainties and contingencies with experts and members of the public, it becomes clear that safety will actually be a journey, a collective experiment whose questions and metrics are not set in advance. One interviewee, discussing the various approaches to self-driving safety, said "the answer is it's going to be all of that. There won't be any single thing that says this vehicle is safe", before arguing that "we will learn as we go along" (3).

When considering questions of trust and public credibility, we should ask who the "we" is in this response. Can we

[10] The full Sciencewise report is available here https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/951094/cav-public-acceptability-dialogue-engagement.pdf, accessed 25 March 2021.

trust that the technology developers' questions and metrics are sufficient? Or should the process of learning be opened up? There is a clear case, given that much of the experimentation is happening in plain sight, using publicly-owned roads as a laboratory, for the democratisation of learning. This would mean, at a minimum, sharing safety–critical data for overall safety improvement, enabled by what some authors have called an 'ethical black box' (Winfield and Jirotka, 2018). It could also mean governing the technology in explicitly experimental terms, as clinical trials do for new medicines. This approach, proposed by London and Danks (2018) would see staged approval for tests that would be overseen by government before being scaled up. Once we recognise the limits of narrow approaches to the assessment and management of risk, we can ask what trustworthy governance might look like.

Trustworthy self-driving cars?—Performance, process and purpose

A multidimensional view of technological risk should change how we think about trust in autonomous vehicles. First, if we take public views on risk seriously, we should recognise that trust is out of the control of innovators and regulators. They can design for trustworthiness, but trust is a gift of the public, hard-won and easily lost. Second, rather than talking about trust in self-driving cars as technological artifacts, we should consider trust in the systems that govern the technology. Third, we should think about trust beyond just performance. The Cold War adage 'trust but verify' presupposes that we know what to measure. As we have seen in the debate on self-driving safety, the relevant numbers are at the centre of a controversy that is not just about what self-driving performance, but also about the processes by which they function and the purposes of innovation.

Lee and See's (2004) framework has trust in automation resting on three pillars—competence, integrity and benevolence (see also Mayer et al., 1995). Their focus is on people using automated systems, such as pilots and train operators, but if we see trust as a concern for governance as well as human–machine interaction, we might see trust depending on issues of technological processes and purposes as well as those of performance. Members of the public are likely to pay attention not just to what self-driving cars do, but also to how they work, how they are developed and what the technology is used for.

As it stands, the regulatory debate on self-driving safety emphasises performance-based metrics that rest on companies' own safety cases. If governance becomes locked in to this mode of safety assurance it could be socially brittle. Just as engineers would be concerned with a single point of failure in a technological system, so regulators should worry about balancing public trust on a single pillar. There is a strong case for socially-robust governance that is more pluralist, building on cultural theories of risk that look for the multiple ways in which people prioritise and make sense of hazards (Thompson & Rayner, 1998).

The interviews analysed for this paper reveal innovators' emphasis on self-driving performance and neglect of questions of process and purpose. Public discussions suggest that people will want to scrutinise why a self-driving car does what it does, and who self-driving cars are seen as benefitting. Only one interviewee drew an explicit connection between risk assessment and the question of unequal benefits:

> Risk acceptance will also be different whether you talk about trucks driving on highways from A to B,… autonomous shuttles driving at 30 [kilometres per hour] in a city [or] luxury vehicles driving on motorways up to 130, where only rich people benefit (15).

Given the potential politics of self-driving car innovation, what is the potential for governance to connect technological means and ends?

For genetically-modified crops, competing political constitutions of the technology led to competing governance frameworks in Europe and the US. While American regulators focussed on crops as products and assessed their performance, European regulators saw genetic modification as a novel process, surrounded by uncertainties and creating social as well as physical risks (Jasanoff, 1995). In both cases, regulation was framed by assumptions of what the technology was for: who benefitted and how. For self-driving cars, as standard-setting and other governance processes gear up, regulators should challenge emerging models of de facto governance and seek more deliberate, and more deliberative, alternatives. One interviewee, a researcher at a self-driving company, described a set of dilemmas that faced the company as it scaled up its operations. This researcher, unusual in their reflexivity, saw potential gaps between colleagues' motivations and the public interest:

> The verification/validation side?… From a safety standpoint… I have no doubt about the commitments and intentions of people who are doing that... But, you know, where and how? What does it mean to be validated and verified? What voices are in the room, on what grounds? What are the metrics that get considered and don't get considered?... Who's making those decisions? (22)

Those questions will not melt away; they will intensify as and when the technology scales up. One lesson emerging from our workshops with members of the public is that there will be a public interest in how self-driving cars work and how they are developed and governed. People are unlikely to be satisfied just by public displays of self-driving in action

and statistics showing average improvements in safety. Early proposals such as ethical black boxes and publicising leading metrics such as disengagements, while they need work, at least offer ways to open up the governance of processes of self-driving and its testing. In addition, there will be a public interest in the perceived purposes of self-driving technologies and who they are likely benefit. Members of the public will have different expectations from engineers. The questions 'safe enough for whom?' and 'safe enough for what?' will be unavoidable. Seen in this light, the question of how we know whether a self-driving car is social as well as scientific. Some of the uncertainties will be hard if not impossible to resolve, but the project must bring together a wide range of disciplines and draw on public as well as expert insights.

Funding This study was supported by the Economic and Social Research Council and the Alan Turing Institute (ES/S001832/1).

References

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Blumenthal, M. S., Fraade-Blanar, L., Best, R., & Irwin, J. L. (2020) 'Safe Enough: Approaches to Assessing Acceptable Safety for Automated Vehicles'. Rand Corporation.

Blyth, P. L., Mladenovic, M. N., Nardi, B. A., Ekbia, H. R., & Su, N. M. (2016). Expanding the design horizon for self-driving vehicles: distributing benefits and burdens. *IEEE Technology and Society Magazine, 35*(3), 44–49. https://doi.org/10.1109/MTS.2016.2593199

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Cummings, M. L., & Thornburg, K. M. (2011). Paying attention to the man behind the curtain. *IEEE Pervasive Computing, 10*(1), 58–62. https://doi.org/10.1109/MPRV.2011.7

European Commission. (2020). *Ethics of connected and automated vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility.* LU: Publications Office. Available at: https://data.europa.eu/doi/10.2777/035239. Accessed 6 December 2020.

Fraade-Blanar, L., Blumenthal, M. S., Anderson, J. M., & Kalra, N. (2018) *Measuring Automated Vehicle Safety: Forging a Framework*. Rand Corporation.

Funtowicz, S. O., & Ravetz, J. R. (1993). Science for the post-normal age. *Futures, 25*(7), 739–755. https://doi.org/10.1016/0016-3287(93)90022-L

Gann, D. M., Wang, Y., & Hawkins, R. (1998). Do regulations encourage innovation?—the case of energy efficiency in housing. *Building Research & Information, 26*(5), 280–296. https://doi.org/10.1080/096132198369760

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Hansson, S. O. (2009). From the casino to the jungle. *Synthese, 168*(3), 423–432. https://doi.org/10.1007/s11229-008-9444-1

Irwin, A. (1985). *Risk and the control of technology: Public policies for road traffic safety in Britain and the United States*. Manchester: Manchester University Press.

Jasanoff, S. (1995). Product, process, or programme: three cultures and the regulation of biotechnology. In M. Bauer (Ed.), *Resistance to new technology* (pp. 311–332). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511563706.016

Kalra, N., & Groves, D. G. (2017). *The enemy of good: estimating the cost of waiting for nearly perfect automated vehicles*. Rand Corporation.

Kalra, N. & Paddock, S. M. (2016). 'Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?' Rand Corporation.

Kearnes, M. B., & Wynne, B. (2007). On nanotechnology and ambivalence: The politics of enthusiasm. *Nanoethics, 1*(2), 131–142. https://doi.org/10.1007/s11569-007-0014-7

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies, 96*, 290–303. https://doi.org/10.1016/j.trc.2018.07.001

Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine, 9*(1), 90–96. https://doi.org/10.1109/MITS.2016.2583491

Krohn, W., & Weyer, J. (1994). Society as a laboratory: The social risks of experimental research. *Science and Public Policy, 21*(3), 173–183. https://doi.org/10.1093/spp/21.3.173

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Leveson, N. (2004). A new accident model for engineering safer systems. *Safety Science, 42*(4), 237–270. https://doi.org/10.1016/S0925-7535(03)00047-X

Leveson, N. (2011). *White Paper on the Use of Safety Cases in Certification and Regulation*.

Liu, P., Yang, R., & Xu, Z. (2019). How safe is safe enough for self-driving vehicles? *Risk Analysis, 39*(2), 315–325. https://doi.org/10.1111/risa.13116

London, A. J. and Danks, D. (2018). 'Regulating autonomous vehicles: A policy proposal', In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA. pp. 216–221. https://doi.org/10.1145/3278721.3278763.

MacKenzie, D. (1998). The certainty trough. In R. Williams, W. Faulkner, & J. Fleck (Eds.), *Exploring expertise: issues and perspectives* (pp. 325–329). London: Palgrave Macmillan UK.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review, 20*(3), 709–734. https://doi.org/10.2307/258792

Millstone, E., Brunner, E., & Mayer, S. (1999). Beyond "substantial equivalence." *Nature, 401*(6753), 525–526. https://doi.org/10.1038/44006

Mindell, D. A. (2015). *Our robots, ourselves: Robotics and the myths of autonomy*. Penguin.
Nye, D. E. (1996). *American technological sublime*. MIT Press.
Rayner, S. (2004). The novelty trap: Why does institutional learning about new technologies seem so difficult? *Industry and Higher Education, 18*(6), 349–355. https://doi.org/10.5367/0000000042683601
Rayner, S., & Cantor, R. (1987). How fair is safe enough? The cultural approach to societal technology choice1. *Risk Analysis, 7*(1), 3–9. https://doi.org/10.1111/j.1539-6924.1987.tb00963.x
Ryerson, M. S. et al. (2021). 'Safer City Streets for Pedestrians and Bicyclists', *Issues in Science and Technology*, 24 February.
Sgobba, T. (2019). B-737 MAX and the crash of the regulatory system. *Journal of Space Safety Engineering, 6*(4), 299–303. https://doi.org/10.1016/j.jsse.2019.09.006
Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour, 1*(10), 694–696. https://doi.org/10.1038/s41562-017-0202-6
Slovic, P. (1987). Perception of risk. *Science, 236*(4799), 280–285. https://doi.org/10.1126/science.3563507
Starr, C. (1969). Social benefit versus technological risk. *Science, 165*(3899), 1232–1238.
Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science, 48*(1), 25–56. https://doi.org/10.1177/0306312717741687
Stilgoe, J. (2019). *Who's driving innovation?: New technologies and the collaborative state*. Springer.
Stirling, A. (2007). Risk, precaution and science: Towards a more constructive policy debate. *EMBO Reports, 8*(4), 309–315. https://doi.org/10.1038/sj.embor.7400953
Tennant, C. & Stilgoe, J. (in press). The attachments of 'autonomous' vehicles, Social Studies of Science.
Thompson, M., & Rayner, S. (1998). Risk and governance part I: The discourses of climate change. *Government and Opposition, 33*(2), 139–166.
van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics, 22*(3), 667–686. https://doi.org/10.1007/s11948-015-9724-3
Vaughan, D. (1996). *The challenger launch decision: risky technology, culture, and deviance at NASA*. University of Chicago Press.
Vinsel, L. (2019). *Moving violations: automobiles, experts, and regulations in the United States*. Baltimore: Johns Hopkins University Press. Hagley library studies in business, technology, and politics.
Weber, H., et al. (2019). A framework for definition of logical scenarios for safety assurance of automated driving. *Traffic Injury Prevention, 20*(sup1), S65–S70. https://doi.org/10.1080/15389588.2019.1630827
Weinberg, A. M. (1972). Science and Trans-Science. *Minerva, 10*(2), 209–222.
Winfield, A. F. T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180085. https://doi.org/10.1098/rsta.2018.0085
Wolff, J. (2006). Risk, fear, blame, shame and the regulation of public safety. *Economics & Philosophy, 22*(3), 409–427.
Wynne, B. (1983). Redefining the issues of risk and public acceptance: The social viability of technology. *Futures, 15*(1), 13–32. https://doi.org/10.1016/0016-3287(83)90070-8