Maya Indira Ganesh

ENTANGLEMENT: MACHINE LEARNING AND HUMAN ETHICS IN DRIVER-LESS CAR CRASHES

APRJA Volume 6, Issue 1, 2017
ISSN 2245-7755

Algorithmic regulation of everyday life, institutions and social systems increases with little oversight or transparency, and yet usually with significant social outcomes (Angwin et al.; Pasquale). Therefore, the need for an 'ethics of algorithms' (Ananny; CIHR) and 'accountability' of algorithms (Diakopolous) has been raised. The "constellation of technologies" we have come to refer to as 'artificial intelligence'[1] (Crawford and Whittaker) enable an anxiety that sits alongside the financial speculation, experimentation and entrepreneurial enthusiasm that feeds the Silicon Valley gold rush of 'innovation'. How can machine intelligence optimise its decision-making and avoid errors, mistakes and accidents? Where machines are not directly programmed but learn, then who or what is accountable for errors and accidents, and how can this accountability be determined?

This paper is based on driver-less car[2] technology as currently being developed by Google[3] and Tesla, two companies that amplify their work in the media. More specifically, I focus on the moment of real and imagined crashes involving driver-less cars, and argue that the narrative of 'ethics of driver-less cars' indicates a shift in the construction of ethics, as an outcome of machine learning rather than a framework of values. Through applications of the 'Trolley Problem', among other tests, ethics has been transformed into a valuation based on processing of big data. Thus ethics-as-software enables what I refer to as big data-driven accountability. In this formulation, 'accountability' is distinguished from 'responsibility'; responsibility implies intentionality and can only be assigned to humans, whereas accountability includes a wide net of actors and interactions (in Simon). 'Transparency' is one of the more established, widely acknowledged mechanisms for accountability; based on the belief that *seeing into* a system delivers the truth of that system and thereby a means to govern it. There are however limitations to this mechanism in the context of algorithmic transparency (Ananny and Crawford). This work does not begin with a definition of accountability, but is part of a larger body of ongoing work that asks how accountability may be defined anew in a context where human and non human agents are in interaction.

This paper starts by looking at a recent crash involving a Tesla semi-autonomous car, and then examines literature around aviation crashes as a body of work that narrates how accountability in complex vehicles human-machine systems has been approached. This literature shows that establishing accountability is difficult because of the dense entanglements between human action and machine agency; that identifying the actors and events involved in a crash include complex chains of human and non-human agents. However, in the development of the driverless car, machine learning is being used to practically pre-empt crashes. I show that ethics becomes a framework to guide the development of machine learning, and thus in doing so sets up linear paths of accountability: if the machine can minimise human error by learning how to respond to a vast number of crash scenarios, then accountability becomes something much easier to un-entangle. However, this is the ambition for a fully autonomous consumer vehicle, which does not yet exist, and is unlikely to for at least the next ten years. Based on their documentation of driverless car testing and crashes, Brandon Schoettle and Michael Sivak conclude that the most risky period is that of transition from conventional driving to driverless cars. Moreover, industry predictions suggest that the insurance industry could be transformed by autonomous driving, moving to a model of offering coverage of technical errors rather than personal liabilty, much like cruise ships and airlines (Bertoncello and Wee). Thus I conclude

that there is a pressing need to confront machine learning as it is being applied to the new framing of ethics-as-accountability; and consequently develop new considerations of accountability in terms of, and building on, the reality of entanglements between human and machine agents.

Of Tesla and other crashes

In May 2016, an ex-US Navy veteran was test-driving a Model S Tesla semi-autonomous vehicle. The test driver, who was watching a Harry Potter movie at the time with the car in 'auto-pilot' mode, drove into a large trailer truck whose white surface was mistaken by the computer vision software for the sky. Thus it did not stop, and went straight into the truck. The fault, it seemed, was the driver's for trusting the auto-pilot mode, as the company's condolence statement suggests:

> *It is important to note that Tesla disables Autopilot by default and requires explicit acknowledgement that the system is new technology and still in a public beta phase before it can be enabled. When drivers activate Autopilot, the acknowledgement box explains, among other things, that Autopilot "is an assist feature that requires you to keep your hands on the steering wheel at all times," and that "you need to maintain control and responsibility for your vehicle" while using it. Additionally, every time that Autopilot is engaged, the car reminds the driver to "Always keep your hands on the wheel. Be prepared to take over at any time." The system also makes frequent checks to ensure that the driver's hands remain on the wheel and provides visual and audible alerts if hands-on is not detected. It then gradually slows down the car until hands-on is detected again. (Tesla)*

Tesla goes into detail to clarify that the human is assumed to be in control even though 'auto-pilot', familiar to anyone who has been up in an airplane, implies that the machine is in control. This confusion over the meaning of auto-pilot becomes a critical moment to begin to think about the relationship between the human operator and a complex machine and how challenging it becomes to identify responsibility for errors. The literature from aviation crash histories offers some valuable insights in this direction, and suggests that responsibility for a crash has never been easy to ascertain.

The history of aviation crashes shows that human error tends to be cited as the most common reason for accidents; moreover, there is a tendency to "praise the machine and punish the human" for accidents and crashes (Elish and Hwang). Looking specifically at the history of the role of autopilot, scholars find that even though there has been increasing automation in the cockpit, the responsibility for accidents remains with human pilots (10).

Peter Galison finds that identifying the cause of an aviation accident can be a Byzantine exercise. Examining narratives of aviation crashes, he finds that there is a deep entanglement in accounts of accidents between human actions and the perceived agency of technologies, a "recurrent strain to between a drive to ascribe final causation to human factors and an equally powerful, countervailing drive to assign agency to technological factors" (4). Galison finds that in accidents, human action and material agency are entwined to the point that causal chains both seem to terminate at particular, critical actions as well as radiate out towards

human interactions and organisational cultures (4). Yet, what is embedded in unstable accident reporting is the desire for a "single point of culpability" (Brown 378), which never seems to come.

Galison finds in these multi-causal accounts that there has been a gradual move away from individual action towards examining "mesoscopic world[s] in which patterns of behavior and small-group sociology could play a role" (37). A good example of the role of small-group sociology comes from Diane Vaughan's landmark ethnography of the *Challenger* Space Shuttle crash. She finds that the crash was caused by the 'normalisation of deviance', a slow and gradual loosening of standards for the evaluation and acceptance of risk in an engineering context. This loosening happens because of organisational-cultural issues, and not because of blatant corruption or malafide intent. *Challenger* exploded 73 seconds into its ascent because the 'O rings' on the rocket's boosters broke on that unusually cold January morning; and yet, it was known for over a year that the rings would fail in cold weather. Vaughan found that how engineers, scientists, bureaucrats, and managers communicated and managed risky or faulty engineering was determined by the bureaucratic language or processes of NASA. It got hidden, reframed, minimised, second-guessed, and eventually buried. In unearthing it, Vaughan found a complicated chain of accountable actors.

Madeleine Elish and Tim Hwang acknowledge multiple sites of potential responsibility for crashes and ask, "how do we locate the network of human actors responsible for the actions of computational agents?" (22). Is it the car manufacturer that is responsible, or the software development team that programmed the car's software? In the Tesla case, who is responsible? Is it the driver who lost his life because he misinterpreted what auto-pilot mode means, the computer vision software that wrongly categorised the side of a long truck trailer for the sky, or the manufacturer, Tesla, that did not pre-empt these possibilities? If all of these actors, and others not identified here, are somehow part of the story of how and why the crash happened, then how are they all to be held accountable and to what extent?

What is at stake in how accountability is assigned for crashes involving driver-less cars? In order to answer this question, this paper began by showing that assigning accountability in aviation crashes reveals a complex entanglement between the human operator and machine agent; and that, despite increasing automation, humans are still held responsible for crashes. Next, taking this forward into the driverless car context, I make a detour into machine learning in driver-less car technology; from there I will discuss how machine learning is related to the application of the *Trolley Problem* and the Pascalian *Wager*, which are both used to construct an 'ethics' of autonomous driving. This will then allow me to show how software and big data are implicated in the consequent framing of ethics.

Computer vision and machine learning for accuracy in driver-less cars

The precision and accuracy of driver-less cars comes from software that 'learns' appropriate driving behaviour – merging, driving around construction zones, etc. – through exposure to large datasets that its algorithms are trained on. The combination of computer vision and machine learning is used so the car can detect objects, identify and categorise them, and rely on data it has

been 'exposed' to in order to make a decision about how to respond to objects and avoid accidents.

One of the most significant features of machine learning algorithms is that they determine patterns. Algorithms such as convolutional neural nets, that are used in driver-less car software use their pattern-recognition ability to build internal models for identifying features of a dataset. Eventually, they can learn how those features are related without being explicitly programmed to do so (NVIDIA; Bojarski et al.). Another distinctive feature of machine learning more generally is that it is not always possible to open up the system and identify exactly *how* or *why* a decision was made to categorise and analyse something – machine learning is an inscrutable technology (Knight).

Rather than have to be 'brute-force' programmed, or 'hard-coded', to respond to every single possible situation it might encounter – a near impossible and exhausting software engineering exercise – driver-less car software uses machine learning to establish how to respond to unfamiliar situations through repeated practice (Google, *Self Driving Car Project*). An illustrative parallel to the difference between hard-coded programming and machine learning exists in the history of computers programs that play 'perfect information games', games where all information about the status of the game is available to all players. In the 1980s, *Deep Blue* was an IBM computer program that was brute-force programmed to play Chess; that is, every possible permutation and combination of moves that could be made on a 8×8 board with 32 pieces was programmed. The ancient Chinese game of Go however has a far higher number of possible moves; it is a more complex game than Chess. So in the development of *Alpha Go*, the Google computer program that plays Go, the algorithm looks at millions of games of Go, and discerns patterns in it. It can read which moves, and which combinations of moves, are more or less successful in achieving a winning outcome and then it is able to enact those moves when playing a game (Hassabis, Alpha Go).

Driver-less cars have to learn how to identify objects so they know how to respond to them in a similar way. Cars are fitted with radar, LIDAR ('light detection and ranging') and other sensors with which to perceive the environment around them. Computer vision software identifies an object and breaks up that image into small parts: edges, lines, corners, colour gradients and so on. By looking at billions of images, the neural nets in cars can identify patterns in how combinations of parts come together to constitute different objects. The expectation is that such software can identify a ball, a cat, or a child, and make a decision about how to react based on the data received. Yet, this is a technology still in development and there is the possibility for much confusion. So, things that are yellow, or things that have faces and two ears on top of the head for instance, can be misread until the software sees enough examples that distinguish how things that are yellow, or things with two ears on the top of the head, are different from one another. In the case of the Tesla crash, the software misread the large expanse of the side of the trailer truck for the sky. It is possible that the machine learning was not well-trained enough to make the required distinction.

Depending on what the object is, the driver-less car is expected to respond: stop, go around it, wait for it, and so on. With increasing exposure to good quality data, the software can distinguish between different kinds of objects and eventually make more fine-grained decisions. The more complex something is visually, without solid edges or curves or single colours – or if it is a fast, small, or flexible object on the road – the more difficult it is to understand. So, driver-less

car software is shown to have a so-called 'bicycle problem' because bicycles are difficult to identify, are not a structured shape, and can move at different speeds (Fairley). Being able to identify objects on the road and assess their relative value in relation to each other has become a central aspect of the narrative around ethics in driver-less cars, which the paper now turns to.

Programming ethics in machines: Trolley problems and wagers

Ethics is assumed to be a framework for values governing appropriate actions in society; and often applied in situations that are difficult for the law to regulate, or where laws do not yet exist. 'Machine ethics', 'information ethics', 'computer ethics', and 'robot ethics' are some overlapping fields that deal with ethics in contexts relevant to the present discussion, however it is beyond the scope of this paper to unpack each of these in more detail. Mike Ananny has identified three approaches to ethics in technology across these domains, and these tend to mirror consequentialist, Kantian (or, deontological), and virtue ethics: developing policies and regulations by codifying use of technologies, developing standards, best practices and anticipating future failures; anticipating the ethical outcomes of technologies and how they reconfigure social relationships; and investigating the values of designers and developers of these technologies (95).

In the context of driver-less cars, the accident is framed as a moment when a decision has to be made by software about how to avoid it. This decision-making process is tantamount to 'ethics' and has been framed in terms of Kantian ethics and consequentialist ethics through the Trolley Problem, a popular shorthand for the discussion about ethics in driver-less car contexts (Lin; Google, *Self Driving Car Project*). In the world of the Trolley Problem, an autonomous vehicle is expected to learn to make the optimal choice in the case of the worst scenario imaginable – an autonomous vehicle being involved in the killing of human beings.

The Trolley Problem is a classic thought experiment developed by the Oxford philosopher, Philippa Foot in 1967, originally to discuss the permissibility of abortion. The Trolley problem is presented as a series of hypothetical situations with two or more negative outcomes, in which consequentialist or deontological approaches must be used to find a way to choose the lesser of two negative outcomes. The Trolley Problem is described by Judith Jarvis Thompson in the following way:

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

Thompson goes on to describe versions of the Problem substituting track workmen and the trolley with other characters and circumstances. Each version of The Trolley Problem necessitates a process of reasoning by invoking the tension between Kantian ethics, and consequentialist ethics: does how you *arrive* at the outcome matter more than the *outcome* itself? Is it more important to save more lives (a consequentialist approach), or is it more important to consider *how* people die? (the deontological approach), and in which situations is one approach more valid than the other?

Patrick Lin has developed an application of the Trolley Problem (as described by Bhargava and Kim 2017) as has MIT's Moral Machine Project. In the Lin version, the driverless car is in a situation where it has to decide which of two difficult options to select in order to save itself, such as having to either hit a cyclist wearing a helmet or one that is not; or decide what to do if a child runs out across a road; or how to rationalise potentially harming occupants of a car known to have poor crash test ratings. The Moral Machine Project is an online research exercise based on the Trolley Problem that serves as "a platform for 1) building a crowd-sourced picture of human opinion on how machines should make decisions when faced with moral dilemmas, and 2) crowd-sourcing assembly and discussion of potential scenarios of moral consequence" (Rahwan, Bonnefon and Sharif). In this, the driver-less car has to select which kinds of humans to avoid hitting – children, pregnant women, older people, escaping thieves, athletes, or animals like cats and dogs – in the case of brake failure.

Vikram Bhargava and Tae Wan Kim find however that the Trolley Problem does not address the fact that Kantian and consequentialist cannot be resolved because they are not of the same kind of moral value ("value incommensurability"); that the Problem sets up a situation beset by moral uncertainty; that it does not afford a "view from nowhere", meaning one that is 'objective'. In such an objective view, say the authors, even the driver-less car should be factored in to the question of who or what should be saved in the case of an unavoidable crash; in the Trolley Problem, the driver-less car and its occupants are not assumed to be at risk in a crash, only pedestrians or other vehicles and drivers are. Instead Bhargava and Kim suggest an application of the Pascalian *Wager*, along with a ranking system developed by Andrew Sepielli. In this, calculations to rank different outcomes of crashes are developed to arrive at an 'objective' choice. So, the cases of the car with failed brakes ramming into a child, an animal, or a helmet-wearing cyclist, or destroying itself to save others, are all given numerical rankings. An algebraic calculation processes these rankings to arrive at the most mathematically objective outcome. The authors note that ethics tests properly applied in this way could help to establish accident claims under the law, and allow manufacturers to offer their customers a "moral navigation system", much like a menu of Facebook's privacy settings from a drop-down list; and manufacturers could generate crowdsourcing mechanisms to generate datasets of appropriate, and objective, decisions for machine learning (13-14).

There is a nuanced shift suggested by this scenario. If 'ethics' has become a series of computations that can be augmented by big data, then ethics – and thereby failures of ethics – is seen as a matter of individual morality rather than that of a group of individuals, organisations, laws, or other actants. Through application of the Trolley Problem, it is almost as if the car is imagined to be a sort of neoliberal, individualised, subject. As 'self driving', it is imagined to be an individual moral agent that can act independently and efficiently on the basis of guidelines and feedback (Ganesh).

It is possible that data currently being harvested about driver behaviour from a variety of sources – from highway cameras, police records, social media, insurance records, automotive engineering simulations, and so on – will be used to develop machine learning algorithms that will learn how to make decisions across different situations. Something to this effect has already been in development and testing as Anderson and Anderson discuss in their paper on the possibility of creating an ethical, intelligent machine agent. They cite an approach to applied ethics called *casuistry*: "the branch of applied ethics that, eschewing principle-based approaches to ethics, attempts to determine correct responses to new ethical dilemmas by drawing conclusions based on parallels with previous cases in which there is agreement concerning the correct response" (20). This appears to be very much along the lines of machine-learning for decision-making discussed here. They cite work by Rzepka and Araki that identifies such an approach to machine ethics:

> *it might be safer to have machines "imitating millions, not a few," believing in such "democracy-dependent algorithms" because, they contend, "most people behave ethically without learning ethics." They propose… [to] search the web for opinions, usual behaviors, common consequences, and exceptions, by counting ethically relevant neighboring words and phrases, aligning these along a continuum from positive to negative behaviors, and subjecting this information to statistical analysis. They suggest that this analysis, in turn, would be helpful in the development of a sort of majority-rule ethics useful in guiding the behavior of autonomous systems." (Anderson and Anderson 20)*

However they do not discuss what the practical implications of this sort of application are. For example, that crowdsourced datasets are neither 'raw' nor neutral, and import the errors, biases, and the cultural and local contexts encoded in them. One of the 'promises' of big data is that of insight and prediction, a kind of 'higher' knowledge (boyd and Crawford). Something to that effect is being invoked here in pre-empting crashes. The idea that crash situations can be envisioned is, however, not entirely new to the automotive industry. In the 2000s car manufacturers began to invest large sums in mathematical modelling. Paul Leonardi cites Nigel Gale's work in identifying "road to lab to math" as an industry-wide belief that mathematics-based simulations are more cost-efficient than road and laboratory testing:

> *Math is the next logical step in the process over testing on the road and in the lab. Math is much more cost effective because you don't have to build pre-production vehicles and then waste them. We've got to get out in front of the technology so it doesn't leave us behind. We have to live and breathe math. When we do that, we can pass the savings on to the consumer. (244)*

Thinking outside black box ethics

In the development of driver-less cars we can see an ambition for the development of what James Moor refers to as an *explicit ethical agent* – one that is able to calculate the best action in an ethical dilemma – through big data technologies. In the development of machine intelligence towards this goal, a series

of shifts can be discerned: from accounting for crashes after the fact, to pre-empting them; from ethics that is about values, or reasoning, to ethics as crowdsourced, or based on statistics, and as the outcome of software engineering. Thus ethics-as-accountability moves towards a more opaque, narrow project, and away from the kinds of entanglements that scholars such as Galison and Vaughan identify.

Yet, as the Tesla crash indicates, if there was both an error in the computer vision and machine learning software, as well as a lapse on the part of the test driver who misunderstood what the term autopilot meant, then how are these two conditions to be understood as part of the dynamic that resulted in the crash? What is the relationship between them? How might an ethics be imagined for this sort of crash that comes from an unfortunate entanglement of machine error and human error?

In a 2016 paper, Mike Ananny and Kate Crawford confront the idea of transparency as a mechanism for algorithmic accountability citing ten limitations of the idea of transparency, emphasising that "it is sometimes unnecessary and always insufficient to simply look inside structures"; but that the limitations of the idea of transparency could serve as a starting point for accountability (12-13). In this vein, I conclude with an agenda for future work.

In thinking about a framework for values, and in rethinking accountability, how can the multiple, parallel conditions present in driving be conceptualised? Rather than understanding an 'ethics of driver-less cars' to be a set of programmable rules for appropriate action, could it instead be imagined as a process by which an assemblage of people, social groups, cultural codes, institutions, regulatory standards, infrastructures, technical code, and engineering are framed in terms of their interaction? As Ananny notes:

> *In reality, technology ethics emerges from a mix of institutionalized codes, professional cultures, technological capabilities, social practices, and individual decision making. Indeed, ethical inquiry in any domain is not a test to be passed or a culture to be interrogated but a complex social and* cultural achievement *(Christians et al. 2009). It entails anticipating how the intersecting dynamics of a sociotechnical system—design, interpretation, use, deployment, value—"matter" for the future (Marres 2007)—and figuring out how to hold these intersections accountable in light of an ethical framework. (96; emphasis in original)*

In this conception, ethics is not just a end-point or outcome, but is something that can be imagined as a series of individual and system-level negotiations involving socio-technical, technical, human and post-human relationships and exchanges. The more challenging and intriguing questions of how these actors and their inter-relationships are to be materialised and made visible are still to be answered, but perhaps we may start to discern the shape of the black box.

Notes

[1] In a recent public event, *AI Now*, convened by the White House, 'artificial intelligence' was defined as a constellation of technologies that includes machine learning, natural language processing, and big data. This text ascribes to this definition of AI as a constellation.

[2] 'Autonomous vehicles', 'self driving cars', and 'driver-less cars' are all commonly used terms today referring to the same technology. There are five levels of autonomy in vehicles as defined by the United States' National Traffic and Highway Safety Authority. At present, there is no fully autonomous vehicle in testing or operation, but it is Google's ambition to create one. Tesla is working on a semi-autonomous vehicle. Traditional car manufacturers have been introducing increasing levels of autonomy in existing car models, such as adaptive parking, highway assist, or cruise control. Thus, this paper does not use the word 'autonomous vehicles' but uses the terms 'driver-less cars' or 'self driving cars' to refer to this technology.

[3] Both Google and its self-driving car project have undergone some changes in identity. Google is now known as *Alphabet*, and the self driving car project is called *Waymo*.

Works cited

Ananny, Mike. "Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness." *Science, Technology, & Human Values* 41 (1) (2016): 93-117. Print.

Ananny, Mike and Kate Crawford. "Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability." *New Media & Society* (2016): 1-17. Print.

Anderson, Michael and Susan Leigh Anderson. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28 (4) (2007). Print.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica*, May 23 (2016). Print.

Arnold, Thomas, and Matthias Scheutz. "Against the Moral Turing Test: Accountable Design and the Moral Reasoning of Autonomous Systems." *Ethics of Information Technology* 18: 103–115 (2016). Print.

Bertoncello, Michele and Dominik Wee. "Ten Ways Autonomous Driving Could Redefine the Automotive World." McKinsey & Company, June 2015. www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-could-redefine-the-automotive-world. Accessed March 7, 2017. Web.

Bhargava, Vikram and Kim Tae Wan. "Autonomous Vehicles and Moral Uncertainty" *Roboethics 2.0: From Autonomous Cars to Artificial Intelligence*. Ed. Patrick Lin, Keith Abney, and Ryan Jenkins. London: Oxford UP. 2017 (forthcoming). Print.

Bojarski, Mariusz et al. "End to End Learning for Self-Driving Cars." (2016).

boyd, danah and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5), A Decade in Internet Time: The Dynamics of the Internet and Society (2012). Print.

Brown, Alexander. "Accidents, Engineering and NASA 1967-2003." Ed. Steven J. Dick and Roger D. Launius. *Critical Issues in the History of Spaceflight.* National Aeronautics and Space Administration (2006): 377-402. Print. National Aeronautics and Space Administration.

"Ethics of Algorithms." Centre for Internet and Human Rights, 2015. www.cihr.eu/eoa2015web/. Accessed October 1, 2016. Web.

Crawford, Kate and Meredith Whittaker. "The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term." Symposium report, 2016. www.artificialintelligencenow.com/media/documents/AINowSummaryReport_3.pdf. Accessed October 2, 2016. Web.

Diakopolous, Nicholas. "Algorithmic Accountability: On the Investigation of Black Boxes." Tow Center for Digital Journalism, 2014. www.nickdiakopoulos.com/wp-content/uploads/2011/07/Algorithmic-Accountability-Reporting_final.pdf. Accessed March 12, 2017. Web.

Elish, Madeleine and Tim Hwang. "Praise the Machine! Punish the Human! The Contradictory History of Accountability in Automated Aviation." *Comparative Studies in Intelligent Systems – Working Paper #1. Intelligence and Autonomy Initiative* 1, 24 Feb. 2015. Data & Society. www.datasociety.net/pubs/ia/Elish-Hwang_AccountabilityAutomatedAviation.pdf. Accessed September 23, 2015. Web.

Fairley, Peter. "The Self-Driving Car's Bicycle Problem." *IEEE*, 31 Jan. 2017. www.spectrum.ieee.org/cars-that-think/transportation/self-driving/the-selfdriving-cars-bicycle-problem. Accessed February 6, 2017. Web.

Foot, Philippa. "The Problem of Abortion and the Doctrine of the Double Effect." *Oxford Review* 5 (5-15). (1967). Print.
Galison, Peter. "An Accident of History." Ed. Peter Galison and Alex Roland. Atmospheric Flight in the Twentieth Century. Springer Science and Business Media (2000): 3-43. Print.

Ganesh, Maya Indira. Personal communication with Martha Poon, 27 Jan 2017.

Google. *Self Driving Car Project*. Monthly Report. Aug. 2015. http://static.googleusercontent.com/media/www.google.com/en//selfdrivingcar/files/reports/report-0815.pdf. Accessed December 20, 2016. Web.

Google. *Google's Autonomous Vehicle*. (n.d.). www.googlesautonomousvehicle.weebly.com/ethics.html. Accessed December 4, 2016. Web.

Hassabis, Demis. "Alpha Go: Using Machine Learning to Master the Ancient Game of Go." Google, 2016. www.google-blog.blogspot.de/2016/01/alphago-machine-learning-game-go.html. Accessed April 10, 2016. Web.

Jarvis Thompson, Judith. "The Trolley Problem." *Yale Law Journal* 94 (6), May (1985): 1395-1415. Print.

Lin, Patrick. "The Ethics of Autonomous Cars." *The Atlantic*, 2013. www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/. Accessed December 20, 2016. Web.

Knight, Will. "If a Driver-less Car Goes Bad We May Never Know Why." *MIT Technology Review*. 7 July 2016. www.technologyreview.com/s/601860/if-a-driver-less-car-goes-bad-we-may-never-know-why/?set=601871. Accessed March 7, 2016. Web.

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

"End to End Learning for Self Driving Cars." *NVIDIA*, 2006. https://devblogs.nvidia.com/parallelforall/deep-learning-self-driving-cars/. Accessed December 15, 2016. Web.

Rahwan, Iyad, Jean-Francois Bonnefon and Azim Shariff. *Moral Machine*, 2016. http://moralmachine.mit.edu. Accessed March 31, 2017. Web.

Simon, Judith. "Distributed Epistemic Responsibility in a Hyperconnected Era." Ed. Luciano Floridi. *The Onlife Manifesto: Being Human in a Hyperconnected Era*. Springer International, 2016. Print.

Sivak, Michael and Brandon Schoettle. "Road Safety with Self-Driving Vehicles: General Limitations and Road Sharing with Conventional Vehicles." University of Michigan Transportation Research Institute, 2015. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/111735/103187.pdf?sequence=1&isAllowed=y. Accessed online March 15, 2017. Web.

"A Tragic Loss." Tesla, 2016. https://www.teslamotors.com/blog/tragic-loss. Accessed March 15, 2017. Web.

Vaughan, Diane. *The Challenger Launch Decision: Risky Technology, Culture and Deviance at NASA*. Chicago: University of Chicago, 1997. Print.