POLICY INSIGHTS

Thinking ahead for Europe

No 2018/02, January 2018

# Ethics, algorithms and self-driving cars – a CSI of the 'trolley problem'

Andrea Renda

## Summary

Many experts argue that focusing too much on how automated cars will solve the famous 'trolley problem' isn't going to get us very far in the debate on the ethics of artificial intelligence (AI). But it's hard to resist if you are a philosopher, an ethicist, a futurist, or simply a geek – and it's fun. Still, this ethical dilemma can reveal a number of outstanding policy issues that are often neglected in the public debate.

This paper performs a 'crime scene investigation' to find some of the missing parts in the ethics/AI quandary. These include the need to preserve human control over machines; the need to take data governance and ownership seriously; algorithmic accountability and transparency; various forms of user empowerment and their tension in relation to overall system control; the need for modernised tort rules; and more generally, a discussion about whether algorithms should reflect, exacerbate or mitigate the biases existing in our society.

The investigation concludes that current legal systems are insufficiently equipped to cope with most of these issues, and that a mapping of outstanding ethical and policy dilemmas is a useful starting point for a thorough overhaul of public policies in this ever-expanding domain.

Andrea Renda is Senior Research Fellow at CEPS and Chair in Digital Innovation at the College of Europe.



978-94-6138-664-9

Available for free downloading from the CEPS website (www.ceps.eu)

# Contents

# Ethics, algorithms and self-driving Cars – A CSI of the 'trolley problem'

Andrea Renda

CEPS Policy Insight No. 2018-02 / January

Of the dozens of ethical dilemmas posed by the emergence of artificial intelligence and advanced machine learning, the 'trolley problem' (already a 'must' in ethics and moral philosophy) is undoubtedly the most popular.[1] Scholarly papers, op-eds, blog posts and even dedicated websites discuss how automated cars should behave whenever some form of collision is inevitable; and whether they should prioritise the life of their occupants, that of pedestrians, or neither.[2] The MIT's 'Moral Machine' website guides users through various scenarios.[3] Some academics discuss the need for "utilitarian" cars, others would prefer "prioritarian cars".[4] Other academics use game theory to conclude that pedestrians would impose their presence on risk-averse self-driving cars, especially if they behave with impunity.[5] Scholars imagine cars equipped with an "ethical knob", which could set key patterns of behaviour such as "full altruist", "fully egoist", or "impartial".[6] Others opt for simple "value of life" models as opposed to neural networks, since the latter are "black boxes", and as such would be largely unacceptable to end users.[7]

Market players and regulators have not been silent either. Last year the German car manufacturer Mercedes took an explicit stance on this ongoing *querelle* by announcing that indeed, their cars were being designed to prioritise occupants over pedestrians; but the German Ministry of Transportation immediately challenged this plan by anticipating the content of a new regulatory intervention, which would make the Mercedes priority scale

[1] See https://en.wikipedia.org/wiki/Trolley_problem for a definition of the trolley problem. Also, see M. J. Costa (1986), "The Trolley Problem Revisited", *Southern Journal of Philosophy* 24 (4):437-449.

[2] See i.a. S. Nihols and J. Smids (2016), "The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?", *Ethical Theory and Moral Practice*, November 2016, Volume 19, Issue 5, pp 1275–1289.

[3] See http://moralmachine.mit.edu/

[4] Bonnefon et al. (2017) recently argued that "regulating for utilitarian algorithms may paradoxically increase casualties by postponing the adoption of a safer technology". See J.F. Bonnefon, A. Shariff, I. Rahwan (2016), "The social dilemma of autonomous vehicles", *Science* 352, 1573–1576.

[5] See A. Millard-Ball (2016), "Pedestrians, Autonomous Vehicles, and Cities," *Journal of Planning Education and Research*,https://people.ucsc.edu/~adammb/publications/Millard-Ball_2017_Autonomous_vehicles_pedestrians_cities_preprint.pdf.

[6] See G. Contissa, F. Lagioia and G. Sartor (2017), "The Ethical Knob: ethically-customisable automated vehicles and the law", *Artificial Intelligence and Law*, September 2017, Volume 25, Issue 3, pp 365–378.

[7] See L. R. Sütfeld, R. Gast, P. König and G. Pipa (2017), "Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios: Applicability of Value-of-Life-Based Models and Influences of Time Pressure", *Frontiers of Behavioural Neuroscience* 2017 11:122.

illegal.[8] A recent guidance paper by the US National Highway Traffic Safety Administration (NHTSA) went in the same direction, by observing that in some, relatively rare situations, "the safety of one person may be protected only at the cost of the safety of another person", and that "algorithms for resolving these conflict situations should be developed transparently using input from Federal and State regulators, drivers, passengers and vulnerable road users, and taking into account the consequences of [automated vehicles] on others".[9] Meanwhile, the three Waymo-enabled FCA cars introduced on the streets of Phoenix, Arizona in October 2017 are apparently not endowed with any specific moral value and are expected to simply go for the "smaller object":[10] but even having no specific rule, as the NHTSA itself acknowledges, is a rule.

So far, so good. But how useful is the trolley problem in prompting a useful policy conversation? Or are we just being mesmerised by another chapter of this well-known ethical saga, repackaged in a high-tech *salsa*? Probably the latter is true: experts typically argue that so many things must have gone wrong for such a problem to eventually occur in reality, that the dilemma is practically irrelevant for policy purposes. And some have also observed that there are structural differences between the "human" and the "robotic" version of the trolley problem.[11] One expert recently put it quite explicitly in a tweet: "Why do people think the trolley problem is critical for self-driving cars? The trolley problem wasn't critical even for trolleys".[12] Amitai and Oren Etzioni recently took a similar stance by warning about relying too much on extreme scenarios such as the trolley problem to conceptualise the incorporation of ethics into AI.[13]

Hence, as Zen monks would put it, the answer to the question asked in the trolley problem is 'mu', which simply means: don't ask (or 'un-ask') the question. But a deeper and more critical reflection suggests that the problem can unveil a number of unresolved issues that remain below the surface in the public and scholarly debate on the future of AI. As such, the trolley problem is not a useless dilemma, but focusing on it too much might overlook more important overarching issues, just as looking at the finger is meaningless when someone points to the moon. The moon, not the finger, is the subject matter of this short reflection.

[8] The principle mentioned is "A car never distinguishes between humans based on categories such as age or race". See https://www.newscientist.com/article/mg23130923-200-germany-to-create-worlds-first-highway-code-for-driverless-cars/.

[9] See NHTSA, Federal Automated Vehicles Policy, Accelerating the Next Revolution In Roadway Safety, at https://www.transportation.gov/sites/dot.gov/files/docs/AV%20policy%20guidance%20PDF.pdf.

[10] See http://uk.businessinsider.com/self-driving-cars-already-deciding-who-to-kill-2016-12?r=US&IR=T

[11] See https://link.springer.com/article/10.1007/s10677-016-9745-2#Fn1

[12] See https://twitter.com/andrewyng/status/791648421291528197

[13] See A. Etzioni and O. Etzioni (2017), "Incorporating Ethics into Artificial Intelligence", *J. Ethics*, December 2017, Volume 21, Issue 4, pp 403–418.

## A crime scene investigation

So, imagine that you are detectives just arrived on site to open a crime scene investigation, and need to ask yourselves a few questions about the circumstances that led to the incident. For example, what made collision possible? What did the car know about the humans involved? What do we know about how the car reached a decision? Who was in charge of deciding over the best course of action? Was the self-driving car supposed to behave like a reasonable human being? And ultimately, who is liable for the damage incurred? Below, we address these questions one by one. And as anticipated, we do not try to solve the trolley problem.

### *Problem 1: How did the car end up there?*

In its version featuring automated cars, the trolley problem assumes that there will be interactions between cars and pedestrians. We simply imagine cars as they are today, on streets that look as they do today, but without a human being behind the steering wheel. We assume that pedestrians will walk on pavements at the side of the street as they do today. We implicitly assume that they will adopt the same standard of care and the same patterns of behaviour when crossing the street. Are we sure this is going to be the case?

The answer is uncertain, for many reasons. First, automated cars and trucks may initially be deployed in highways, possibly on dedicated lanes, in order to avoid difficult interactions with human-driven vehicles.[14] In densely populated areas, they would have to interact with both human-driven vehicles and unpredictable pedestrians, which would make their job a lot more complicated: as a matter of fact, even bad weather conditions can jeopardise the correct scanning of the surrounding environment by an autonomous vehicle. This is the reason why Waymo is trialling its three FCA cars in the suburbs of Phoenix (AZ), where the street map is very simple, the speed is low and the weather is almost always good. In other conditions, problems would become much harder to solve for the vehicle: even if it is true that the recent performance of algorithms in competitions such as the 'Stanford dog' project and *Imagenet* suggest that one day those problems will become more tractable for machines,[15] moving images are still an enigma for computers: the most successful algorithms still have remarkably low success rates when identifying cyclists and their direction, even when the weather is good.[16]

Moreover, even if automated cars are allowed in densely populated, metropolitan areas, this does not automatically imply that pedestrians will use the road in the same way: dedicated

[14] See i.a. D. Sabin (2017), "New Algorithm Lets Self-Driving Cars Merge With Traffic Like a School of Fish", at https://www.inverse.com/article/27119-algorithm-merge-autonomous-highway. And Meeder et al. (2017), "Autonomous vehicles: Pedestrian heaven or pedestrian hell?", ETH Zurich, at http://www.strc.ch/2017/Meeder_EtAl.pdf

[15] See A. Krizhevsky, I. Sutskever, G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks". In *Advances in Neural Information Processing Systems*, https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[16] See https://www.theguardian.com/cities/2017/jun/14/street-wars-2035-cyclists-driverless-cars-autonomous-vehicles.

lanes, pedestrian bridges or underway passes may be built to avoid interference between autonomous cars and humans (outside of them).[17] At a minimum, new technologies based on stigmergy such as Starling Crossing could be deployed to allow pedestrians to cross the road only when it is safe to do so.[18] Duke University's Human and Autonomy Lab is also studying various ways in which self-driving cars will communicate to pedestrians in case of interaction, something that should make the trolley problem even less likely.[19]

Our first lesson is thus that we can still decide whether we want the trolley problem to occur in practice. There is no need to design rules in a way that puts algorithms and AI in charge of deciding over human lives. And no need to imagine the future as a slight variant on what exists today. On the contrary, we can make design decisions in ways that preserve the outstanding potential of automated transportation, at the same time avoiding the delegation of tragic choices to machines. This will also be useful for self-driving car manufacturers: as a matter of fact, no matter how many lives will be saved by automated cars, the ones that are ended will count more in the eyes of the citizens. The level of acceptability of a killer machine veering onto a wall to kill the occupant of a car or changing direction to crush a small group of bystanders is much lower than that of a drunk-driver running over an innocent pedestrian. And acceptability is essential for new technologies to succeed.

Our first lesson also leads to a first rule: whenever possible, policy decisions should give priority to alternatives that do not place robots or self-learning algorithms in a position to decide over human lives. This rule potentially applies even if other alternatives appear more efficient from a cost-benefit perspective. This rule partly overlaps with the need for "human control" specified by rule 16 of the Asilomar principles:[20] but it goes beyond that, embracing a specific approach to the precautionary principle.

### *Problem 2: What did the car know?*

In our crime scene investigation, do we know what information was available to the car? The answer to this question is crucial, and not only for liability purposes: some individuals might be willing to delegate the decision to a very well-informed car, but not to a car that only relies on its own, highly imperfect sensors. In other words, in order to appraise the risk associated with our decision it is essential to factor into the analysis the degree of confidence we have in the adequacy of the inputs received by the machine; in the soundness of the process it follows to elaborate the input; and in the quality and accuracy of the output (the decision itself). As

[17] Incidentally, the fact that we picture automated cars in metropolitan areas is a significant victory for the automotive industry: until a few years ago scenarios for urban mobility did not necessarily contemplate the use of automated cars: we were discussing *whether* there would be cars in future smart cities; today, we discuss *how* cars will interact with the environment in smart cities. See also http://www.strc.ch/2017/Meeder_EtAl.pdf and the Guardian Article cited above.

[18] See https://www.fastcompany.com/40481550/the-crosswalk-of-the-future-moves-and-changes-to-prioritize-pedestrians.

[19] See https://www.inverse.com/article/35686-driverless-cars-pedestrian-safety-future.

[20] "Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives". See futureoflife.org/ai-principles.

Mittelstadt et al. (2016) put it, one of the key ethical issues associated with algorithms is the so-called 'garbage in, garbage out' problem, which makes the quality of the output (and most likely, the desirability of the outcome) dependent on the quality of the input received.[21] The twin problems of inconclusive or misguided evidence are never factored into the analysis of the trolley problem, but are indeed essential to guide our initial decision on whether to trust the system or not.

What is happening, in this respect, in the world of self-driving cars? There is a raging battle between some industry players (e.g. large tech giants), who would like self-driving cars to base their decision only on what their sensors signal (as in the case of Waymo's purely offline cars, but also in line with rumours about Apple's recent patent application for Autonomous Navigation Systems);[22] players that advocate the use of information received from other cars (Vehicle-to-Vehicle, or V2V communication); and yet another group of players (in particular, European telecommunications companies and the whole of the EU) that propose that self-driving cars use information received from the infrastructure (Vehicle to Infrastructure, V2I) or from the whole environment (Vehicle-to-Environment, V2E).[23] The latter implies reliance on 5G networks, LiDAR sensors, and other fixed and wireless infrastructure sources. This debate is still ongoing: in the US, significant investment was made so far on V2V systems, but in Europe the story might unfold differently. A likely evolution is that self-driving cars will end up using only sensors and wireless technology when in sparsely populated areas, but may also use 5G and fixed infrastructure (e.g. embedded in street lighting) when the density of the population allows for the deployment of more expensive technologies.

But the amount of data that might be processed by a self-driving car at any given point in time could also be much greater, if technological development allows. For example, distributed ledger architectures could be used to build a real-time shared description of the state of the road infrastructure (including the presence of humans). This might reduce the need for heavy investment in multi-technology systems to be installed on the vehicle and within the surrounding infrastructure. Cars would then coordinate on the basis of a single, verifiable description of the current state of the road. Not surprisingly, car manufacturers such as Toyota, Porsche and Daimler have invested in blockchain applications for self-driving cars: however, such applications will have to overcome considerable technological hurdles, not least latency, security and privacy concerns, before they can become viable. Apple is reportedly going in the opposite direction by minimising cars' reliance on maps.[24] Moreover, the use of quantum computing might lead to better processing of information and more advanced traffic

[21] See B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter and L. Floridi (2016), "The ethics of algorithms: Mapping the debate", *Big Data & Society*, July–December 2016: 1–21.

[22] A car typically does not always distinguish a dog from a frog. See, on Apple's recent patent application, https://www.cnbc.com/2017/12/21/apple-patents-navigation-system-for-self-driving-cars.html

[23] See i.a. S. Karnouskos and F. Kerschbaum (2018), "Privacy and Integrity Considerations in Hyperconnected Autonomous Vehicles", *Proceedings of the IEEE 2018*, http://www.fkerschbaum.org/pieee18.pdf

[24] See http://www.businessinsider.com/new-apple-patent-self-driving-cars-minimize-map-usage-2017-12?IR=T

optimisation flows, and technologies in this respect are being developed by large industrial conglomerates, including a new joint venture between Volkswagen and Google.[25]

In other words, self-driving cars will in the future be able to use all the information we want them to use to solve the trolley problem, if it ever presents itself. For example, if street lights use cameras, the presence of pedestrians could also be detected by these fixed installations; should they have facial recognition capabilities, the amount of information available to the self-driving car would also include the identity of the individuals; or, if the pedestrians that could possibly be hit by the car hold a smartphone or use wearables, information on their position could also be used from those sources, and could include many more personally identifiable details (I will get back to this issue later). And the use of wearables could also enable some form of alert warning sent by the car to the inadvertent pedestrian.[26]

Finally, in circumstances in which an accident might be caused by more than one vehicle it would be interesting to find out what amount of information was available to each car on the expected behaviour of other cars. Here, it is important to distinguish between at least three different situations: i) a self-driving car is dealing with a human-driven car; ii) a self-driving car is dealing with another self-driving car, and the behavioural criteria of both cars are regulated and/or standardised; and iii) a scenario in which there are two or more self-driving cars, but their ethical or behavioural settings have been reprogrammed by their users in a way that other cars cannot incorporate. The second scenario is the easiest to manage for a self-driving car, whereas the other two might create significant problems due to the need to avoid the interaction of two or more algorithms, which creates significant problems in terms of outcomes and related liability (see below).

In summary, knowing what the car knew is important for our crime scene investigation. And it gives us an important warning about the variety of business and regulatory models that still have to be discussed in depth and selected as the technology becomes commercially viable. This variety raises one issue that will apply to the remainder of this paper: where do we draw the line between accuracy of information and protection of personal data? Or, put differently, where do we draw the line between the efficiency of the algorithm and the fairness of its outcomes?

### *Problem 3: What do we know about how the car decided?*

Even if we know what information went into the algorithm that was driving the car, we might not know how the algorithm decided to process the information. Algorithms can use a multitude of criteria to reach decisions, and these criteria might not always be transparently shared with users and authorities. Algorithms might be patented and proprietary, and as such difficult to audit or monitor. Subjecting them to close inspection or requiring that their internal

[25] See https://www.cnet.com/roadshow/news/vw-google-announce-quantum-computing-partnership/ VW started its first quantum-computing project in March in China to optimise traffic for 10,000 taxis in Beijing, using another technology supplier.

[26] See http://cityobservatory.org/self-driving-cars-versus-pedestrians/

functioning be open to the public might be criticised as stifling the incentive to innovate by weakening property rights. Still, not being able to observe what algorithms do, and how they reach their decisions, might create insurmountable problems for law enforcement. Hence, a heated debate on algorithmic accountability has emerged over the past few years, especially in Europe, starting from competition law cases but soon expanding into many other aspects of our daily lives, from recruitment to street police, advertising, financial services and e-commerce. Some authors demonise the use of algorithms based on big data analytics, defining them as black boxes (Pasquale 2016); "Weapons of Math Destruction" (O'Neill 2016); or focusing on how they can lead to dilution of liability and at the same time enable new, more sophisticated forms of cartels (Stucke and Ezrachi 2016).[27]

The issue becomes even more problematic if one considers that the type of algorithm that will be used in self-driving cars is still an enigma. Neural network algorithms are seen by some as the most efficient and accurate, and by others as too much of a black box.[28] Adaptive, self-learning algorithms might be seen as the most technically advanced, but also as less predictable than others, including by their own developers.[29] Clustering and pattern recognition algorithms need enormous amounts of data to become acceptably accurate, and as already noted, have significant problems in recognising moving images. Inevitably, the choice of the data sources used will have an impact on the type of algorithm used, and this creates a double layer of uncertainty in the expected evolution of algorithms in self-driving cars.[30]

The debate among experts has led to some first results in terms of defining principles for algorithmic transparency and accountability. In particular, the Association for Computing Machinery issued in 2017 a "Statement on Algorithmic Transparency and Accountability";[31] and the Institute of Electrical and Electronics Engineers released at the end of 2016 a draft for public discussion titled "Ethically aligned design", which underlined the need for accountability that can help in "proving why a system operates in certain ways to address legal issues of culpability, and to avoid confusion or fear within the general public".[32] The Fairness, Accountability and Transparency in Machine Learning (FATML) interdisciplinary academic community worked on defining key principles of accountability, which include responsibility, explainability, accuracy, auditability and fairness (on which, see below under problem 4).[33] Kroll et al. (2017) examine a number of avenues for improving the accountability of algorithms, including privacy-compliant

[27] See F. Pasquale (2015), *The Black Box Society. The Secret Algorithms That Control Money and Information*. Harvard University Press, Cambridge, MA. C. O'Neill (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishers, New York. M. Stucke and A. Ezrachi (2016), *Virtual Competition*, Harvard University Press, Cambridge, MA. And see also Mittelstadt et al. (2016), *supra* note 21.

[28] See Sütfeld et al. (2017) *supra* note 7.

[29] Ibid.

[30] See i.a., https://www.kdnuggets.com/2017/06/machine-learning-algorithms-used-self-driving-cars.html

[31] Association for Computing Machinery US Public Policy Council (USACM), available at http://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

[32] See Ethically Aligned Design. IEE. Available at http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf.

[33] See https://www.fatml.org/resources/principles-for-accountable-algorithms.

mechanisms such as zero-knowledge proofs and fair random choices.[34] Kim (2017) argued in response that additional empirical verification of outcomes through constant, ongoing monitoring would be needed to ensure that algorithms are accountable for the way they reach decisions.[35]

That said, the issue of whether algorithms should be fully observable from third parties or policymakers is far from settled in the policy arena. The debate has been further promoted by the adoption of the EU General Data Protection Regulation, which arguably contains a right for end users to receive an explanation of how the algorithm works, although the scope of this right is far from clear in the text (and more specifically, in the recitals) of the Regulation, and disappeared by the final text of the European Parliament resolution on civil law rules for robotics, which initially advocated compliance of smart autonomous robots with the GDPR (Wachter et al. 2017).[36] All in all, the legal system appears underdeveloped and dangerously fragmented, with many robotic systems likely to fall outside the scope of any attempt towards accountability.

### *Problem 4: Better than us, or like us?*

Even if the problems of algorithmic transparency and accountability were solved, the criteria self-driving cars should rely upon to solve potentially fatal dilemmas would remain nebulous at best. So far, car manufacturers have admittedly taught their vehicles to look for the smaller object: but whatever lies behind this smaller object won't be known, and the equation 'smaller object equals smaller damage', intuitively, can prove very misleading. Now assume, in this respect, that an algorithm is taught to seek the 'smallest damage', and as such pursues the goal of minimising social cost. This would entail that the algorithm behaves rationally, i.e. performs a benefit-cost analysis of alternative courses of action. This, too, would lead to a number of undesirable consequences.

First, the algorithm would need to assess a variety of consequences, which (by definition) would include the loss of human lives, but possibly also damage to property. For example, the car might be confronted with an option A (e.g. hit five pedestrians) that generates no loss in property, and an option B (e.g. hit a wall and hurt, possibly fatally, four passengers), which would create damage to a highly valued monument (e.g. if the wall is an ancient Roman wall). If the algorithm expressed damage to assets (including cultural and environmental assets) in monetary terms and summed them up with estimates of the monetary equivalent of the human lives lost (e.g. using the Value of a Statistical Life formula),[37] then it might end up choosing option B despite the fact that it kills more people. Alternatively, the algorithm could use a lexicographic ordering *à la* John Rawls, by giving absolute priority to minimising the

[34] See http://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=9570&context=penn_law_review

[35] Pauline T. Kim (2017), "Auditing Algorithms for Discrimination", *166 U. Pa. L. Rev. Online* 189.

[36] See S. Wachter, B. Mittelstadt and L. Floridi (2017), "Transparent, Explainable and Accountable AI for Robotics", *Science Robotics*, 2(6).

[37] See e.g. W. K. Viscusi and J. E. Aldy. (2003), "The Value of a Statistical Life: A Critical Review of Market Estimates throughout the World". *Journal of Risk and Uncertainty* 27(1): 5-76.

impact on humans: in that case, cars might end up sacrificing invaluable cultural and environmental assets in the attempt to save one individual; or, in case of environmental damage, they might have to appraise the possible, more scattered damage to human health that might derive from a single course of action (e.g. polluting a river; an explosion in a factory).[38] Finally, the best decision-making criterion for self-driving cars might be found through a 'Rawlsian' method, by asking individuals what behaviour they would prefer, independently of whether they would be involved in the trolley problem scenario, and in what role.[39] But again, the veil of ignorance is not an easy *escamotage*: for example, Bonnefon et al (2016) find that people would choose a utilitarian car: but of course, provided that they are not the ones sitting inside.

Second, the problem of using some form of rational decision-making criterion does not wipe away the issue of algorithmic discrimination. In this respect, it is essential to note that our society is already substantially biased when it comes to evaluating human lives, or calculating damage compensation for the case of accidents: the VSL formula contains both income-related and age-related elements, which might be specifically factored into the analysis when deciding which course of action to take; and tort law explicitly takes foregone earnings (*lucrum caessans*) as a basis for compensating damages from personal injury. So, moving from a situation in which the car looks for the smaller object to one where it looks for the lesser damage inevitably introduces new elements of discrimination in the behaviour of self-driving cars.

The culmination of all these concerns would coincide with a scenario in which a non-transparent algorithm is taught by its manufacturer to look for the least damage for the company, in terms of minimisation of liability exposure. In that case, cars would really need to incorporate an assessment of the prospective damage compensations that the company (or its insurance provider) would face. Public authorities, however complacent, would not be able to fully observe that the car, despite officially trying to minimise social cost, is indeed trying to minimise private cost: discriminatory consequences would be as significant as they are hard to police.

More generally, the jury is still out on whether algorithm developers and vendors should be considered liable for discrimination whenever their algorithms' biases reflect existing biases in our society. This debate is now finally moving from the initially, largely unsatisfactory premise that algorithms should be 'neutral', and is now moving towards more meaningful dilemmas.[40] For example, Google was criticised because its search engine is reportedly far more likely to advertise a highly paid executive job to what it perceives to be a white male compared to an

[38] Isaac Asimov's three laws of robotics apparently aim at this type of lexicographic ordering. They go as follows: i) A robot may not injure a human being or, through inaction, allow a human being to come to harm, ii) A robot must obey orders given it by human beings except where such orders would conflict with the First Law; iii) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

[39] See the so-called "veil of ignorance" in J. Rawls (1971), *A Theory of Justice*, Harvard University Press, Cambridge MA.

[40] See A. Renda (2015), "Searching for Harm, or Harming Search? A look at the European Commission's antitrust investigation against Google", *CEPS Special Report No. 118 / September 2015*.

African-American woman (Sweeney 2013).[41] The PredPol software that implemented "minority report" by indicating to police stations where to send their troops as a crime is probably "about to happen" always send troops to the same urban ghettos. Algorithms suggest the inspection of people that look dark-skinned, or appear likely to be Muslim. In all these circumstances, discrimination exists independently of algorithms, but may be exacerbated by their systematic use. Then the question emerges: should algorithms behave better than we do, or exactly as we do?

Take this problem to the domain of self-driving cars. And now, assume that the car could take into account personally identifiable data when deciding how to solve the dilemma, and who to save. If the car can only detect, however roughly, the age of the pedestrians involved, then it might decide to opt for the younger ones (but not too young, as compensating damage for the death of children is, unfortunately, likely to be cheaper than compensating the life of a wealthy young adult). But if the car knew more, for example thanks to facial recognition technologies, then it could choose based on the perceived contribution to social welfare of the individuals involved, and even information on their "social credit score".[42] So, imagine again that the car must choose between option A (killing five people) and option B (killing four): then, it might use a variety of information sources (e.g. recognition of data in the individual's smartphones and wearables; facial recognition coupled with social credit score information; probabilistic determinations based on the identification of individuals in transit in that area over the past two minutes; images recorded by preceding cars and transmitted through V2V; police records, etc.). As a result, there can clearly be cases in which the four people in the car will be considered more 'valuable' than the five people in the street, and the final decision will, again, be an optimisation one.

Removing algorithmic biases and reaching a satisfactory balance between algorithmic efficiency and adequate protection of users' privacy are two almost intractable problems today; in addition to the inevitable trade-offs that one incurs when trying to provide answers, it is worth recalling that the 'right' thing to do very much depends on the values expressed by our communities, which can differ significantly across the globe. For example, a life/death decision taken by a machine on the basis of a social credit score might be acceptable in South Korea, but not in Europe; and a more informed decision based on personally identifiable data might be more acceptable in the US (where credit history already leads to significant discriminations in access to consumer goods) than in Europe. These boundaries might also change over time, of course, but as of today the existence of a single lens through which to judge what behaviour is right or wrong across countries and populations is pure utopia. The differences in responses given to the trolley problem are, if anything, a good instance of how wide-ranging individual perceptions are of right and wrong: scholars have also investigated this issue, prompting

[41] See L. Sweeney (2013), Discrimination in Online Ad Delivery. Available at https://ssrn.com/abstract=2208240.

[42] See https://www.wired.com/story/age-of-social-credit/.

Buddhist monks with a variant of the trolley problem. From the Kantian categorical imperative to utilitarianism and Buddhist ethics, all angles have already been explored.[43]

The difficulty of solving these problems has led to an unexpected development in scholarly literature. Scholars are not trying to model human behaviour to ensure that cars, rather than behaving better, behave exactly like us, and thus impulsively, rather than rationally. Sütfeld et al. (2017) observe that "simple models based on one-dimensional value-of-life scales are suited to describe human ethical behaviour" in these circumstances, and as such would be preferable to more arcane decision-making criteria, which might ultimately appear too complex, insufficiently transparent, and difficult to predict. This solution appears counterintuitive but oriented towards two essential aspects of technology and society: the acceptability of outcomes, and the predictability thereof, which is fostered by a greater alignment of robots with human behaviour. Not surprisingly, rule 11 of the already mentioned Asilomar principles prescribes that "AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity". But then again, why design autonomous vehicles if then we ask them to replicate our well-known biases and imperfections?

### *Problem 5: Who is going to be liable?*

Our crime scene investigation is mostly aimed at determining what happened, and who is liable for the damage done. But the two results don't come at once: with algorithms, even if the whole accident and causation chain is identified, attributing liability can be a daunting task. There are two different issues that are worth highlighting in this respect. The first relates to the possible distancing between the tortfeasor and the damage incurred that can result from the use of an algorithm. In some circumstances, a car manufacturer could object that the damage occurred due to an unforeseen event, which even a high-capacity computer could not process; or that external circumstances intervened (e.g. a cyber attack), which did not depend on the negligence of the manufacturer, the vendor, or the transportation company (e.g. if a company like Uber or Lyft had provided the self-driving car). This led a few commentators to argue for a "strict liability" (or "no-fault liability") principle that would extend to the sphere of control of the tortfeasor whatever course of action was decided by the AI system in the self-driving car (Floridi 2016).[44] This could occur along the lines of the products liability directive, which however contemplates the possibility that tortfeasors escape liability whenever the likelihood that the product was defective, and thus likely to cause damage to consumers, could not be known based on the state of the art of scientific research at the time the product was placed on the market.

[43] See https://www.theatlantic.com/technology/archive/2017/06/how-do-buddhist-monks-think-about-the-trolley-problem/532092/.

[44] L. Floridi (2016), "Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions", *Philosophical Transactions of the Royal Society*, Volume 374, issue 2083. See also D.C. Vladeck (2017), "Machines Without Principals: Liability Rules and Artificial Intelligence", *Washington Law Review*, Vol. 89:117.

Second, an even more complex problem arises whenever it is the interaction between algorithms that causes damage. Some readers will recall the famous 'flash crash of 2.45', the trillion-dollar stock market crash of May 2010 that was caused by a clash of algorithms, and left millions of dollars burnt and no one responsible for compensating them; or the mind-boggling price surge of the book "The Making of a Fly", which due to an unforeseen interaction between algorithms ended up costing more than 23 million dollars on Amazon.[45] We do not have a legal rule that is tailored to these accidents, even if rules on joint and several liability, or rules that apportion liability based on the degree of contribution to the damage, might be adapted for use in the algorithm age.

More generally, there is reason to believe that the more AI systems become smart and able to learn, the more their decisions will be taken with a degree of autonomy, the greater the potential distance between their 'creators' and the damage they will generate. The European Parliament, in its report on civil law rules for robotics, even argued that smart autonomous robots should be given "rights and duties".[46] But the level of autonomy of these robots might not become, at least in the foreseeable future, advanced enough to really earn these machines the right to be treated as legal entities. On the contrary, even partly autonomous machines might still be linked, from a legal perspective, to their developers: this would mean that robots are treated as 'slaves' to their 'masters': and indeed, this is the etymological meaning of the word 'robot'.[47] A different approach would be to treat robots as animals under civil law rules, which entails that any damage they cause would be attributable to their owners whenever negligence in custody can be proved (so-called *culpa in vigilando*).[48]

## Intermezzo: Ubiquitous robots, the fat guy on the bridge, and the need for system-level analysis

There is no limit to imagination when it comes to the possible benefits and possible risks that could occur from the delegation of decision-making powers to robots. Most importantly, the solution of the trolley problem would need to take into account the whole transportation system, rather than focusing on an individual vehicle.[49] Compared to the traditional trolley problem, in future robots might intervene to solve the problem in many other ways. For example, in the traditional example of the trolley problem, there could be AI systems both in the vehicle and in the switch management system; and other robots could be available nearby,

[45] The book was being sold by two Amazon.com marketplace sellers, each of which was basing its price on the other.

[46] European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

[47] See https://www.fatml.org/resources/principles-for-accountable-algorithms.

[48] See R. Kelley, E. Schaerer, M. Gomez and M. Nicolescu (2010), "Liability in Robotics: An International Perspective on Robots as Animals", *24 Advanced Robotics 13* (2010).

[49] See J. Borenstein, J. R. Herckert and K. W. Miller (2017), "Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis", *Science and Engineering Ethics*, November 2017, pp. 1-16.

and might possibly intervene to move the pedestrians from their current position. Take this popular variant of the trolley problem:

> You are on a footbridge, and next to you there is a very fat man. You realise that by pushing the man down the bridge you would manage to slow down the train and save the five lives of the pedestrians stuck on the tracks.

In ethics, most people would decide not to push down the fat man. But what if instead of you, the bystander on a footbridge was a robot? Should that robot, not directly involved in the dynamics of the trolley problem, be instructed to act to minimise the damage by killing the fat man? In principle, this would be possible, and consistent with a supreme imperative given to robots, to maximise social welfare. Connected robots would be able to communicate seamlessly and in real time, to find the most effective remedy to a compromised situation: the fact that the remedy entails the sacrifice of an innocent bystander would not represent an obstacle for a highly robot-intensive society, since the key option of preserving human lives is already impossible to achieve. But then, would we be comfortable living in a society in which robots obey a supreme imperative, which might require the sacrifice of human lives? Again, this would not be the first time we take such a decision: we have gone to war knowing that we would lose lives; we have built infrastructure such as bridges, tunnels and railroad tracks knowing that someone would lose his or her life; and we decide on pollution standards and spending resources for prevention and post-disaster relief trying to optimise a cost-benefit equation rather by minimising the number of lives lost.

Under these assumptions, there is a need to decide whether the degree of autonomy of a connected robot should only apply to the immediate proximity of its behaviour, and the sphere of control of its master. Alternatively, robots could also be endowed with the possibility of taking action to avoid major damage to the benefit of society as a whole, but only in specific circumstances, which would need to be defined. The etymological root of the word 'robot' originates in the Czech word for 'slave' or 'forced labour': this would suggest a close relationship between the behaviour of the robot and the sphere of control of its master.

## Conclusion: A quick mapping of the key ethical issues of the AI age

How useful is the trolley problem for self-driving cars? And how dangerous is it? In this paper, I have argued that rather than being useful *per se*, the problem can become a very good starting point for a pervasive mapping of the unresolved issues related to the use of robots in various aspects of our daily life. As some experts start imagining a complete transition towards self-driving cars within the next two decades, and a boom in the number of smart and connected objects over the same timeframe, the impact of our choices today in terms of interoperability, connectivity, human control, data availability, legal liability, privacy protection and user empowerment can make a real difference for the sustainability of our social model in years to come. This paper has tried to 'un-ask' the trolley question by showing that the surrounding policy issues are so much more important and complex, that preventing the dilemma from ever happening becomes a compelling choice. The trick is to take a step backwards and consider

that human control should always be prioritised and even enhanced by regulatory decisions. Treating robots as independent, autonomous legal persons would not go in that direction. Ironically, even if self-driving cars must be recognised as having outstanding potential in terms of saving human lives, there seems to be good reasons to avoid creating the conditions for them to interact with pedestrians and, if needed, to decide on human lives in a split second.

The key findings of our crime scene investigation are summarised below.

First, whenever possible, the delegation of choices related to human life to robots or algorithms should be avoided. Any infrastructural choice that can prevent robots from interfering with human behaviour should be pursued, unless absolutely impractical. This should be taken as part of an overall attempt to preserve human control in the long run.

Second, modelling the functioning and effectiveness of algorithms is impossible if one does not discuss the underlying technology and data. Delegating specific functions to a blind car is different from delegating them to a fully informed vehicle that receives information in a secure way from a mix of sources. In this respect, the more data is available to robots (subject to confidentiality constraints), the better the final outcome: however, the weight of this technology on the operation of the vehicle would also need to be considered. The choice of which technologies to use is also largely an industrial policy dilemma, especially if one considers that 5G telephony might be preferred by EU and Chinese players, whereas V2V technology has already been publicly and privately funded in the United States.

Third, algorithm accountability and decision-making criteria must be subject to further elaboration. To establish trust in the algorithmic process, some degree of observability is needed, be it in the form of auditing, transparency obligations, black box recording (*ex post* observation), etc. This, once settled, still does not solve the issue of which criterion should be used for decision-making, and whether users would be able to modify this criterion.

Fourth, it is important that algorithms follow a logic of user empowerment rather than one of protecting passive citizens. This could even go as far as enabling citizens to modify their algorithms, to the extent that this is known to other machines. Without this possibility, the criteria for algorithmic decision-making would need to be standardised and remain unmodifiable by end users.

Fifth, we need liability rules that are not only adapted by a robot-intensive world, but also to a world in which algorithms interact with each other, possibly creating unpredictable outcomes, for which attributing liability would not be easy unless the whole process is fully traceable, and apportionment criteria are available.

Sixth, there is no final word on whether algorithms should be allowed to reflect, exacerbate or mitigate the biases existing in our society. The original answer is that they should improve outcomes compared to what biased legal rules and patterns of human behaviour normally produce: but the meaning of ‘improve’ appears to be utterly obscure. Should organ donation algorithms award priority to younger people, as this would maximise the impact on life-years added? Or to older people, as they won’t be able to profit from subsequent technological

developments? Or to those who are mostly likely to evolve into highly paid human beings? Should banking algorithms deny loans to what they perceive as unlikely to repay them, based on their personal characteristics and big data analytics? And if they do, will malicious algorithms attempt to take advantage of decisions based on past statistics? Will there be a need to 'rebalance' the role of the state?

In summary, ethical issues range from process-related (e.g. the transparency of the algorithms) to outcome-related (e.g. discrimination). Our legal system is currently insufficiently equipped to cope with all these issues, and the emergence of largely self-regulated governance schemes can only exacerbate the problem. In other words, the trolley problem is welcome inasmuch as it prompts us to take action to solve all these issues, rather than to choose one course of action over another.

| Problem | Policy challenge/response |
|---|---|
| 1. *How did the car end up there?* | - Avoid delegating life-threatening decisions to machines<br>- Preserve human control as a key item in policy shaping |
| 2. *What did the car know?* | - Adopt a clear and predictable data policy for self-driving cars, balancing privacy and efficiency<br>- Test the use of privacy-compliant distributed ledgers for automated vehicles<br>- Experiment with forms of differential privacy in algorithms to strike the balance between efficiency and privacy |
| 3. *What do we know about how the car decided?* | - Clarify the legal framework for algorithmic accountability and transparency<br>- Clarify the applicability and scope of the right to explanation under the GDPR<br>- Establish an obligation for *ex post* inspection of automated vehicle 'black boxes' (or action logs) |
| 4. *Better than us, or like us?* | - Define a set of principles for algorithmic decision-making, including clear criteria for separating lawful from unlawful discrimination<br>- Work on anti-polarisation strategies to avoid the AI-powered exacerbation of existing biases |
| 5. *Who is liable?* | - Define strict liability principles for algorithm-powered decision-making<br>- Define legal rules for damages caused by the interaction between algorithms |