The Moral Psychology of AI and the Ethical Opt-Out Problem

Jean-François Bonnefon[1], Azim Shariff[2], and Iyad Rahwan[3,4]

[1]Toulouse School of Economics (TSM-R), CNRS, Université Toulouse-1 Capitole, Toulouse, France
[2]Department of Psychology & Social Behavior, University of California, Irvine, USA
[3]The Media Lab, Massachusetts Institute of Technology, Cambridge MA 02139, USA
[4]Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge MA 02139, USA

Abstract

This chapter discusses the limits of normative ethics in new moral domains linked to the development of Artificial Intelligence. In these new domains, people have the possibility to opt out of using a machine, if they do not approve of the ethics that the machine is programmed to follow. In other words, even if normative ethics could determine the best moral programs, these programs would not be adopted (and thus have no positive impact) if they clashed with users' preferences—a phenomenon that we label the "Ethical Opt-Out." The chapter then explores various ways in which the field of moral psychology can illuminate public perception of moral AI, and inform the regulations of such AI. The chapter's main focus is on autonomous vehicles, but it also explores the role of psychological science for the study of other moral algorithms.

1 Introduction

Most people are happy to use technology driven by Artificial Intelligence (AI), as long as they are not fully aware they are doing so. They enjoy music recommendations, email filters, and GPS advice without thinking too much about the machine learning algorithms that power these products. But

as people let AI-driven technology take an ever greater place in their lives, they also express anxiety and mistrust about things labeled AI. Leaving aside fears of super-intelligent robots lording over humanity (Bostrom, 2014), only 8% people would trust the mortgage advice offered by an AI program—a shade lower than the 9% who would trust their horoscope for investment advice (*Trust in technology*, 2017).

Of course, shopping recommendations and GPS routes arguably do not have a critical impact on people's life outcomes. AI-driven technology, though, is progressively extending into realms in which it will have such an impact, and thus make decisions that fall in the moral domain. Self-driving cars will need to make decisions on how to distribute risk among road users; organ donation algorithms prioritize who will get a transplant; and algorithms already advise judges about who should get probation, parole, or a longer jail sentence.

All these decisions inescapably incorporate ethical principles and complex moral trade-offs. Should self-driving cars always strive to minimize casualties, even if it means sometimes sacrificing their own passengers for the greater good? Should children always have priority for organ transplants, even when an older patient is a better genetic match for an available organ? Should sentencing algorithms always seek to minimize rearrest, even if this minimization results in an unfair rate of false alarms for black and white defendants?

It is not always clear who should be consulted to answer these questions. Should we seek responses from ethicists, or listen to laypersons' opinions? Even though ethicists do not necessarily behave better than laypersons, and even though their initial intuitions may not be better than that of laypersons', their training allows them to think more deeply about these questions and provide solid justifications for their conclusions. Laypersons' intuitions, in contrast, are often untrained and uninformed.

It would be tempting, then, to discard laypersons' intuitions and preferences about the complex ethical issues raised by algorithms and AI-driven technology. But that would be a grave mistake. To understand why, one must realize that if people are not satisfied with the ethical principles that guide moral algorithms, they will simply *opt out* of using these algorithms, thus nullifying all their expected benefits.

Self-driving cars provide the starkest example of the effect of such an opting-out. Imagine (for the sake of the argument) that some ethicists would agree that self-driving cars should always strive to minimize casualties under a veil of ignorance—that is, that self-driving cars should always take the action that minimizes harm, even if this action is dangerous for their own

passengers. This would seemingly guarantee the greatest safety benefits for all road users–measured by the smallest overall number of traffic fatalities. But it would also mean that self-driving cars might autonomously decide to sacrifice (or at least imperil) their own passengers to save other road users—and this possibility is so aversive to consumers that they might opt out of buying self-driving cars, thus forfeiting all their expected safety benefits (Bonnefon, Shariff, & Rahwan, 2016; Shariff, Bonnefon, & Rahwan, 2017).

In other words, even if ethicists were to agree on what they believe to be the best ethical principles to guide a moral algorithm, their work would be made null and void if many laypersons were to strongly disagree with them, to the point of opting out of using the algorithm. This Ethical Opt-Out can take several forms. People opt out of using self-driving cars by not buying them. People opt out of organ donation by either not registering as donors, or registering as non-donors. Finally, people can opt out of judicial algorithms by electing state court judges who vow not to use them (in the US), or by turning to alternative, community-based justice such as Sharia councils (in the UK).

One may still argue that if ethicists were in fact able to come to a consensus about the normative principles guiding moral AI in a given domain, then laypersons should be educated, rather than listened to. In other words, that the best way forward would be to persuade laypersons by rational argument (Walton, 2007) or implicit nudging (Thaler & Sunstein, 2008), rather than to adjust the principles to make them closer to what laypersons spontaneously find acceptable. As a matter of fact, we are agnostic when it comes to this debate. What we note is that *whichever way is actually taken,* public policy will require understanding what people find acceptable—whether with the aim of coming closer to their preferences, or of persuading them that their preferences should be abandoned.

In sum, many benefits of AI technology require people to opt-into an *algorithmic social contract*–an agreement between citizens, mediated by machines (Rahwan, 2018). To facilitate such agreement, we must understand what principles people expect moral AI to follow, lest they opt out from using, enabling, or allowing beneficial AI-driven technology—and we need this understanding regardless of whether we think people should be educated or accommodated. The problem, then, is how we can achieve this understanding. Here we can draw inspiration from the tools and techniques developed in the field of moral psychology, but applying these tools to the field of moral AI raises methodological as well as second-order ethical challenges, which we now address in turn.

2 Methodological challenges

Assessing moral preferences is a complicated matter—one that has drawn in not just the field of moral psychology (Greene, 2014; Haidt, 2007), but also subfields of experimental economics and human behavioral ecology (Gintis, Bowles, Boyd, Fehr, et al., 2005). Moral preferences are fluid, multifaceted and nuanced. To measure them is to accept that much complexity is lost in the measurement, and that some measurement techniques inevitably amount to presenting people with highly simplified, stylized problems—problems that sacrifice realism in order to cut at the joints of moral preferences. Different domains of application call for different degrees of such simplification, as we consider in this section through three examples: autonomous vehicles, kidney paired donation, and algorithmic sentencing.

2.1 Autonomous vehicles

The most famous stylized moral dilemma is known as the *trolley problem* (Foot, 1967). In its most common version, the trolley problem presents people with a scenario in which a trolley car is barreling down on five persons, with no time to stop. If nothing is done, these five persons will die. The only way to save these persons is to pull a lever that would redirect the car on another line. One person, though, is currently on that line and would be killed by the car, with no time to react. The question is whether it would be morally acceptable (or even obligatory) to pull the lever.

This specific scenario is frequently criticized as unrealistic. How many times did such a situation actually occur in the real world? Why can't the car just stop? Why are these people standing there instead of walking a few steps, away from harm? These are all legitimate questions, but experimental psychologists (or experimental philosophers, for that matter) simply ask people to accept the premises of the problem, in order to discover fundamental principles and processes underlying moral judgment. As a result, the trolley problem has led to many important insights about human morality, despite (or thanks to) its unrealistic simplicity.

Consider now the AI version of the trolley problem, in which an autonomous car is barreling down on five persons, and cannot stop in time to save them. The only way to save them is to swerve into another pedestrian, but that pedestrian would then die. Is it morally acceptable (or even obligatory) for the car to swerve? This scenario is clearly as unrealistic as the classic trolley scenario. Why is the car driving at unsafe speed in view of a pedestrian crossing? And why are the only options to stay or swerve—

surely, the sophisticated AI that powers the car should be able to come up with other solutions?

Just like the trolley problem and most experimental stimuli in the behavioral sciences, this autonomous car dilemma is a *model,* not a *reflection* of reality. To borrow a turn of phrase, it is meant to be taken seriously without being taken literally: it captures the gist of many genuine ethical trade-offs that go into the algorithms of autonomous cars, and does so in a way that laypersons can understand.

In the real world, every complex driving maneuver influences relative probabilities of harm to passengers, other drivers, and pedestrians (Goodall, 2014). A car that is programmed to favor a certain set of maneuvers may thus have a higher probability of harming pedestrians, and a lower probability of harming passengers. Though these maneuvers may only minutely shift the risk profile for any individual, the trade-offs that are being made will become apparent when aggregating statistics over thousands of cars driving millions of miles. And these statistics will prompt the same questions as the stylized dilemma does (Bonnefon, Shariff, & Rahwan, in press). For example, imagine that accidents involving one car have a 1-to-2 ratio of passenger to pedestrian fatalities, while another car exhibits a 1-to-7 ratio. Will society accept this discrepancy? Will consumers flock to the second car? Should regulators intervene? Note that we have been there before. For example, 'killer grilles' (also known as 'bull bars') were banned by many regulators because they disproportionately harmed pedestrians and passengers in other vehicles. Regulators identified the ethical trade-off embedded in a physical feature of the car, and acted in the interest of all stakeholders. Should they do the same for the ethical trade-offs embedded in self-driving car software?

By capturing ethical trade-offs embedded in software in a form that all people understand immediately, the stylized dilemma empowers them not to leave ethical choices in the hands of engineers, however well-intentioned these engineers are. To dismiss the stylized dilemma as an abstract philosophical exercise is to hide ethical considerations where lay individuals cannot see them. Most would agree that ethical algorithms should be developed transparently—but transparency is useless if the trade-offs are too obscure for the public to understand. Stylized dilemmas like the trolley problem have a critical role to play to overcome this psychological opacity.

The need for stylized dilemmas should accordingly be assessed as a function of the complexity of the domains to which we apply moral AI. In some domains, it might be possible to measure moral preferences using problems which are actually very close to the real thing. In the next section, we consider one such domain, organ transplants.

		APKD	OPTN
Zero-antigen mismatch	Yes	6	200
High PRA	PRA $\geq$ 80%	10	125
	PRA $\geq$ 50%	6	0
Travel distance	Same region	0	25
	Same center	3	25
Pediatric recipient	Age $\leq$ 5	4	100
	Age $\leq$ 17	2	100
Prior donor	Yes	6	150

Table 1: Examples of criteria used in the kidney allocation algorithms of the Alliance for Paired Kidney Donation (APKD) and the Kidney Paired Donation program of the Organ Procurement and Transplantation Network (OPTN), before their 2016 update. PRA = Panel reactive antibodies.

2.2 Kidney paired donation

Too frequently, candidates for kidney donation have access to a living donor who is unfortunately a poor match for them. To optimize the efficiency of kidney allocation, kidney paired donation (KPD) consists of entering candidates and donors in a database, which is then fed to an algorithm that seeks 2-way, 3-way, or complex chains of donations such that as many candidates as possible find a compatible donor.

The algorithm does not only seek to maximize the number of donations, though. It also uses a scoring rule to determine the priority of each donation (see below), in order to find chains that maximize the number of high-priority donations. While the chain-seeking part of the algorithm might be too complex for laypersons to understand, the same is not true of the scoring rules that determine the priority of each donation. Most criteria in these scoring rules can be readily understood, and the tradeoffs they imply may be explained almost straightforwardly to citizens, and to potential donors in particular.

Consider for example the criteria shown in Table 1, together with the priority points they get under two scoring rules. While the interpretation of the zero-antigen mismatch criterion and the controversies surrounding its use are perhaps best left to specialists (Casey, Wen, Rehman, Santos, & Andreoni, 2015), the other criteria are easy enough for laypersons to under-

stand. Three of the criteria are straightforward (travel distance, recipient's age, recipient's prior donor status). The Panel Reactive Antibodies (PRA) score indicates the proportion of the population against which the candidate is immunized, which accordingly restrict the pool of potential donors for this candidate. A candidate with a PRA score of 80 is thus unable to receive a kidney from 80% of donors.

With this information, laypersons can readily assess some of the tradeoffs implied by the scoring rules, as well as some of their problematic aspects. Consider the problems raised by using cutoffs for continuous criteria such as age and PRA. Why would a 5-year old candidate receive more points than a 6-year old candidate, while the 6-year old candidate does not receive more points than a 7-year old candidate? Is it fair that a candidate with a PRA of 80 gets a massive point gain compared to a candidate with a PRA of 75, while a candidate with a PRA of 98 receives the same number of points than a candidate with a PRA of 80? These are questions which laypersons can easily understand, without the need for researchers to invent stylized dilemmas.

Consider now the relative importance of the criteria, and the fact that they can largely differ between the two scoring rules. Why is it that under the APKD rule, being in the same center as the donor awards slightly more points than being 17, while being 17 awards four times as many points as being in the same center under the OPTN rule? The fact that the scoring rules can largely differ is a telltale sign that we are dealing with fluid, controversial moral tradeoffs. And, again, the palatability of these tradeoffs is likely to influence people's decisions to participate as donors. Moral psychology can assess the public perception of these tradeoffs through experimentation (Freedman, Borg, Sinnott-Armstrong, Dickerson, & Conitzer, 2018), without the need for simplifying the problem to the extent it had to simplify AV ethics into trolley problems.

2.3 Algorithmic sentencing

There are other application domains, though, in which the ethical tradeoffs are not only hard to explain, but also hard to stylize—and these domains will likely prove the most difficult to investigate with the methods of moral psychology. This is especially the case with algorithmic sentencing. Many courts in the U.S. now offer judges the option of using an algorithm that provides a risk score for the defendant—for example, the risk that the defendant will not show up at trial (which can lead a judge to decide that the defendant should await trial in jail), or the risk of recidivism or violent

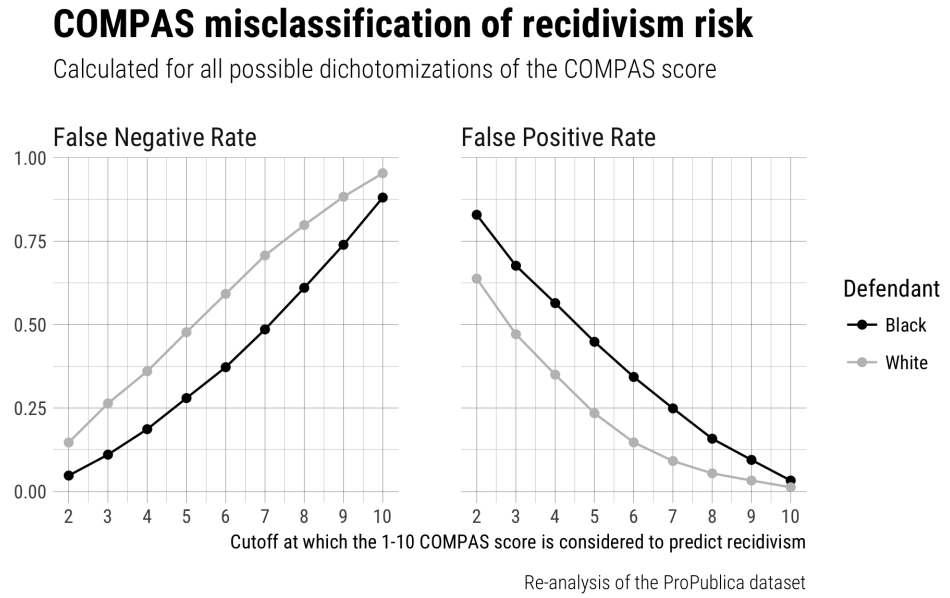

Figure 1: A re-analysis of the ProPublica dataset shows that the main result of Angwin et al. (2016) holds for all dichotomizations of the COMPAS score, assuaging concerns that the result was linked to the arbitrary choice of cutoff in the original article.

crime (which can lead to a longer jail sentence, or a sentence in a higher security prison). While there are dozens of such algorithms, some of them created by nonprofit organizations, the best-known exemplars are proprietary algorithms created by for-profit organizations, such as the COMPAS tool created by Northpointe (now Equivant). The opacity of these proprietary algorithms obviously imposes limits on the realism of any experimental vignette—if we do not even know which parameters the algorithm uses, we cannot investigate the public perception of the tradeoffs between these parameters.

There are some ethical tradeoffs we can experimentally investigate, though, even without knowing the specific implementation of the risk assessment algorithms—but these tradeoffs hardly lend themselves to a one-sentence explanation, or to a trolley-like stylized dilemma. To illustrate, let us unpack the controversy that arose about the potential racial biases of the COMPAS tool.

In May 2016, the investigative news organization ProPublica published a

story titled 'Machine Bias,' which argued that COMPAS was biased against African-American defendants (Angwin, Larson, Mattu, & Kirchner, 2016). ProPublica analyzed a dataset containing the identity of thousands of defendants, together with their COMPAS score for risk of recidivism, and whether they were actually arrested during the two years that followed the COMPAS assessment.[1]

The key result of the analysis, as well as the cornerstone of the story, was that COMPAS erred differently for black and white defendants. Angwin et al. (2016) reported that the false positive rate (i.e., the rate at which defendants were predicted to recidivize, but did not) was 38% for black defendants, compared to 18% for white defendants. Conversely, the false negative rate (i.e., the rate at which defendants were predicted not to recidivize, but did) was 38% for black defendants, compared to 63% for white defendants. In other words, overestimation of risk seemed more likely for black defendants, and underestimation of risk seemed more likely for white defendants. One concern with this result is that COMPAS does not predict recidivism as a binary variable, but delivers instead a risk score from 1 to 10. In order to compute false negative and false positive rates, it is necessary to choose an arbitrary cutoff above which COMPAS is considered to predict recidivism. The results of Angwin et al. (2016) are based on a cutoff of 5, and some critics argued that this arbitrary choice discredited the main finding of the report (Flores, Bechtel, & Lowenkamp, 2016). However, a re-analysis of the ProPublica data assuages this concern by showing that the main finding of the report holds for *any* choice of cutoff (Figure 1).

An algorithm whose mistakes are unfair to black defendants clearly raises ethical issues, but does it reflect an ethical *tradeoff?* In this specific case, the answer appears to be yes, because two conceptions of fairness can apply, whose simultaneous satisfaction is mathematically impossible (Chouldechova, 2017; Kleinberg, Mullainathan, & Raghavan, 2016; Pleiss, Raghavan, Wu, Kleinberg, & Weinberger, 2017). In essence, the algorithm can be equally predictive for both groups, or equally wrong for both groups, but not both. The algorithm is equally predictive for both groups when the probability of recidivism is the same for two individuals who have the same score, regardless of their group. The algorithm is equally wrong for both groups when it yields the same rate of false positives and false negatives for both groups. However, and this is the critical point, these two properties

[1] A detailed presentation of the analysis is available at `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`. Data and code can be downloaded from `https://github.com/propublica/compas-analysis`.

cannot be simultaneously satisfied if the two groups do not have the same baseline probability of recidivism. As soon as one group has a greater recidivism rate, one must decide where to put the cursor between equal predictive power and equal mistakes. It is obvious that unequal mistakes are unfair. And yet, if they are equalized, the risk score must be interpreted differently for black and white defendants. A score of 6 could denote a high risk for a white defendant, and a low risk for a black defendant—which means that judges using the algorithm would necessarily factor race into their sentences, something that they are currently forbidden to do.

We do not intend to explore the legal ramifications of such a transformation of judicial practices. Rather, our goal in unpacking the COMPAS controversy was to show that some ethical tradeoffs will be much harder than others to stylize for laypersons, and thus much harder to study with the standard methods of moral psychology. The problem, though, is that these tradeoffs may be the ones most in need of psychological investigation. We can venture that the number of persons who heard about the ProPublica story is orders of magnitude larger than the number of persons who know about the tradeoff it reflects—and we can imagine as a result that many people believe that sentencing algorithms are intrinsically racist. If we are to gauge the social acceptability of sentencing algorithms, behavioral scientists will have to uncover an appropriate method to make their ethical tradeoffs as understandable as the trolley problems made the ethical tradeoffs of AVs understandable to a general audience.

3 Second-order ethical challenges

Even if we can develop appropriate methods to measure social preferences and expectations about machine ethics, and even if we perceive the benefits of doing so, we need to be careful about the unintended negative consequences of such experiments. In other words, we need to be mindful of the second-order ethical challenges involved in conducting psychological studies of machine ethics. Here we consider two such concerns: that studies of machine ethics may lead to a waste of resources, and that studies of machine ethics may unduly scare the public. It is important to note right away that these concerns are proportional to the media attention that studies receive, for reasons that will be apparent shortly.

3.1 Wasteful studies

Many speakers who have given talks on autonomous vehicles to popular audiences have had the same experience: Whatever their specific topic was, they got a question about trolley problems. That is, not only did one specific method (trolley problems) capture the attention of the media and the public to the point of becoming synonymous with AV ethics, but it threatened to dominate the conversation on AVs to the detriment of more central questions such as overall safety and environmental efficiency.

The concern that we have often heard is that such a fixation may lead car companies and policymakers to make wasteful decisions. For car companies, a wasteful decision would be to commit too many resources to addressing trolley-like dilemmas (which companies are ill-equipped to deal with anyway, for the lack of staff trained in ethics), and not enough resources to improving safety and avoiding such dilemmas in the first place. While there may be some theoretical point at which spending on ethics becomes a wasteful extravagance, we argue that we are not yet close to approaching this point. Though we are not privy to the financial decisions made by car companies, the fraction of resources that these companies devote to ethical issues is most likely an infinitesimal portion of the resources that they devote to engineering issues. Being thrifty about any aspect of safety (absolute or relative) would be a suicidal move for an AV company, which suggests that we should not be overly concerned about ethical teams absorbing the resources of engineering teams.

When it comes to our other examples, kidney paired donation and sentencing algorithms, the situation is quite different, because these algorithms are already in place, and already raising ethical questions or concerns. Here it seems that devoting *more* resources to these ethical issues would be a *good* move, especially in the case of sentencing algorithms—and even if it means that some resources might be diverted from the technological refinement of the algorithms. Overall, it would seem that market forces are more than enough to counter any tendency to overspend on ethics and underspend on performance. Furthermore, the risk of Ethical Opt-Out means that money spent on ethics is not *wasted,* since performance without adoption is useless.

Policymakers, though, may find themselves under pressure to act too fast or too strongly in order to assuage the fears of their constituencies, if these constituencies identify ethics as the sole or most pressing issue regarding the use of AI. The antidote, though, is to conduct more psychological studies, not fewer—as long as these studies can appropriately inform policy-making. The faster we can inform policymakers of what citizens are willing or unwilling

to accept, the lower the risk that policymakers make hasty decisions that hamper the development of AI for no good reason. In sum, the toothpaste is out of the tube now that the general public is aware of the challenges of machine ethics; there is no going back. Psychological studies of machine ethics will not cause wasteful decisions, but the lack of such studies surely will.

3.2 Scary studies

A related but different concern with studies on machine ethics is that we can adversely affect public attitude toward AI by the process of measuring it. Consider again the focus on trolley problems in studies of AV ethics. Trolley-like situations are very aversive while being (in their literal and simplified form) extremely rare. Drawing attention to these situations, then, may adversely and irrationally affect the subjective perception of the safety of AVs.

When thinking of small probability events, people are prone to several biases that include the availability heuristic (risks are subjectively higher when they come to mind easily; Tversky & Kahneman, 1973) and the affect heuristic (risks are subjectively higher when they evoke a vivid emotional reaction; Finucane, Alhakami, Slovic, & Johnson, 2000). Because AV trolley situations can be easily imagined (whatever their actual probability of occurrence), and because they plausibly trigger a strong emotional reaction, the danger is that their likelihood may be overestimated, with downstream consequences on the acceptability of AVs in general. Worse, this impact may be compounded by algorithmic aversion (people lose confidence in erring algorithms more easily than for erring humans; Dietvorst, Simmons, & Massey, 2015). This is an important problem, but once more, it will not be solved by keeping ethical dilemmas out of public sight. In June 2016, the first fatality involving a car in self-driving mode drew more media attention than the some 15,000 human-driven car accidents that occurred in the US on that same day. We can only imagine the coverage of the first fatality that will occur when an AV faces something akin to a trolley dilemma. Before it comes, the public should have had discussed it openly and had a voice in how the AV was programmed to act, rather than been kept in the dark.

In any case, whether people are deterred by AV trolleys is an empirical question deserving of actual research. To explore this question, we conducted a survey on the Amazon Mechanical Turk platform, recruiting 400 participants from the U.S., of which 369 completed the full survey. Participants were randomly assigned to either a condition in which they were first

exposed to three AV trolley dilemmas, and then to four questions about their attitudes toward AVs (the *dilemma first* treatment), or the reversed-order–responding first to the four questions about their attitudes, and only then being exposed to the three AV dilemmas (the *control* treatment). In addition, all participants gave information at the end of the survey on their prior exposure to AV dilemmas,[2] their driving habits, their demographics, and their love of technology (7-item scale). The four questions about attitudes were:

- How excited are you about a future in which autonomous (self-driving) cars are an everyday part of our lives? (7-point scale from 1 = *Not at all,* to 7 = *Very Much*)
- How afraid are you about a future in which autonomous (self-driving) cars are an everyday part of our lives? (7-point scale from 1 = *Not at all,* to 7 = *Very Much*, reverse-coded so that higher scores reflect more comfort with AVs)
- Should they become commercially available by the time you are next purchasing a new car, how likely would you be to choose an autonomous vehicle? (7-point scale from 1 = *Not at all likely: I would rather buy a car without self-driving capabilities,* to 7 = *Extremely likely: I would definitely choose to buy a self-driving car*)
- Compared to current human driven cars, how safe do you expect self-driving cars to be? (7-point scale from 1 = *Much less safe,* to *7* = *Much more safe*)

As shown in Table 2, reading about the ethical dilemmas of AVs had no discernible impact on any measure of participants' attitude towards AVs (the analysis was restricted to the 264 participants who had never heard about the dilemmas before taking the survey; the results are even stronger when the analysis is conducted on the full sample). In particular, reading about ethical dilemmas did not impact participants' perception of their safety, and did not impact their willingness to acquire one. A Bayesian analysis (Morey & Rouder, 2015) showed that the Bayes factors $\frac{\Pr(H_0|D)}{\Pr(H_1|D)}$ ranged from 2.2 to 7.4, offering positive to substantial evidence for the null hypothesis.

[2]First, participants were asked: "Prior to doing this survey, had you heard any discussion about self-driving cars having to make ethical choices such as deciding who should live and die in an accident?" (yes/no). Participants who responded 'yes' were then asked: "You indicated that you had heard about self-driving car ethical issues before. How much thought have you given them?" (5-point scale, from 1 = *None,* to 5 = *A Great Deal).*

	Dilemmas first $N = 132$	Control $N = 132$	t	p	$\frac{\Pr(H_0\|D)}{\Pr(H_1\|D)}$
Excited with AVs	3.4–4.2	3.9–4.5	-1.6	.11	2.2
Will purchase	2.7–3.3	2.8–3.4	-0.4	.65	6.7
Feels safe	3.6–4.3	3.4–4.1	0.1	.92	7.4
Feels no fear	3.3–3.9	3.5–4.1	-0.7	.48	5.8

Table 2: Attitude towards autonomous vehicles (95% confidence interval) for participants who read about ethical dilemmas first, and for control participants who read about ethical dilemmas after they expressed their attitudes about AVs. This analysis is restricted to participants who had never heard about the ethical dilemmas of AVs before taking the survey.

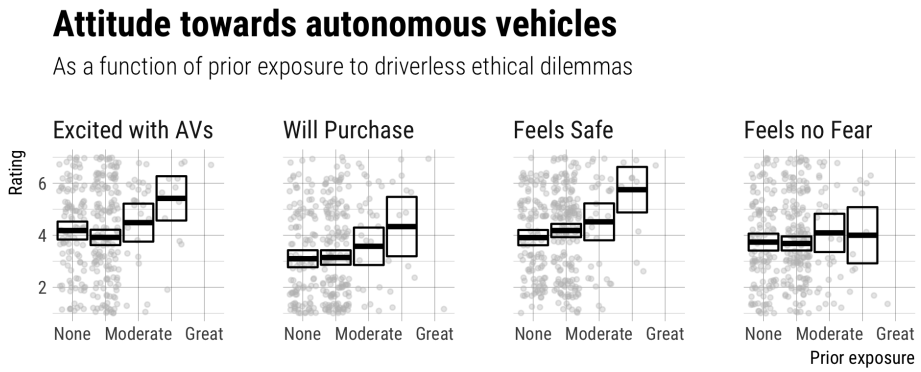

Figure 2: Attitude about AVs as a function of the level of prior exposure to the trolley-like dilemmas of AVs. Boxes show the 95% confidence interval of the mean for each level of exposure, except for 'great' exposure, for which not enough data points were available.

Since participants informed us of their level of exposure to the ethical dilemmas of self-driving cars before taking the survey, we could estimate the impact of this prior exposure on their attitude. Prior exposure to the dilemmas was measured on a 5-point scale (No exposure, little exposure, moderate exposure, lot of exposure, great deal of exposure). For the purpose of this analysis, we reclassified participants who had no prior exposure to the dilemmas but who read about the dilemmas first in the study as having 'a little' exposure. Figure 2 shows the effect of prior exposure on participants' attitudes about AVs. Visual inspection does not suggest that prior exposure

	Excited	Feels Unafraid	Will Purchase	Feels Safe
Prior Exposure	−0.0004	−0.02	0.04	0.10*
	(0.05)	(0.05)	(0.05)	(0.05)
Women	−0.19	−0.45***	−0.19*	−0.35***
	(0.10)	(0.10)	(0.10)	(0.10)
Age	−0.08	0.01	−0.11*	−0.06
	(0.05)	(0.05)	(0.05)	(0.05)
Usually Driver	−0.51*	−0.59*	−0.59**	−0.60**
	(0.23)	(0.23)	(0.22)	(0.23)
Usually Passenger	0.02	−0.27	−0.06	−0.40
	(0.26)	(0.27)	(0.26)	(0.26)
Old Kids	0.29	0.03	0.33	0.04
	(0.18)	(0.18)	(0.18)	(0.18)
Young Kids	0.10	−0.02	0.07	0.01
	(0.11)	(0.11)	(0.11)	(0.11)
Income	0.11*	0.11*	0.09	0.09
	(0.05)	(0.05)	(0.05)	(0.05)
Liberals	0.18***	0.15**	0.17***	0.20***
	(0.05)	(0.05)	(0.05)	(0.05)
Love for Tech	0.35***	0.25***	0.37***	0.27***
	(0.05)	(0.05)	(0.05)	(0.05)
Constant	0.47*	0.77***	0.55*	0.72**
	(0.22)	(0.23)	(0.22)	(0.22)
Observations	369	369	369	369
R^2	0.24	0.19	0.26	0.22
Note:				*p<.05 **p<.01 ***p<.001

Table 3: Attitude toward AVs, as a function of prior exposure to their ethical dilemmas, controlling for demographic characteristics. All continuous variables were standardized before analysis.

has any adverse affect—actually, the trend is positive, suggesting a positive effect of exposure. This trend, though, appears to result from a statistical confound: respondents with a high level of exposure are also the ones with the highest appreciation of technology (see Kramer, Borg, Conitzer, & Sinnott-Armstrong, 2018, for related results). Controlling for this variable (as well as demographic variables), the net effect of prior exposure on attitudes is essentially zero, as shown by regression analyses summarized in Table 3.

In sum, we did not find any evidence that the mere exposure to trolley-like dilemmas had any adverse impact on attitudes toward AVs, or on their safety in particular. People do not seem to be intrinsically scared by ethical dilemmas, which suggests that we might not have to worry too much about the affect heuristic. They may not like all possible *solutions* to these dilemmas, and they are likely to opt out of buying AVs if the solutions they do not like are implemented (Bonnefon et al., 2016)—but merely discussing these solutions is unlikely to sow fear and distrust in the public mind. As a result, there is reason to feel comfortable in continuing with experiments and surveys without fear of, as a byproduct, adversely influencing the attitudes that they measure.

It is unclear whether we should be concerned that exposing people to the ethical tradeoffs embedded in organ transplants algorithms, or sentencing algorithms, might generate some indiscriminate mistrust of all algorithms in these domains. In the case of sentencing algorithms, the question is probably moot. News media and popular books have already exposed a great many citizens to instances in which these algorithms behaved erratically or unfairly (O'Neil, 2017). Exposing study participants and study readers to the *tradeoffs* that the algorithms must face is unlikely to lead to any further generalized negativity than has the asymmetric focus on their mistakes or objectionable predictions. In the case of organ transplants, the notion that donors and recipients must be compatible is so deeply rooted in the public mind, that it would seem hard for people to object, in general, to any algorithm that would seek to maximize compatibility—even though they may object to other criteria introduced in the optimization function. Overall, it would seem that behavioral scientists are on safe ethical grounds for measuring people's preferences about machine ethics.

4 Conclusion

AI-driven technology is extending to domains where algorithms will make or inform decisions with tremendous consequences on people's lives and well-being. Machines may decide who survives a traffic accident; who receives a life-saving organ; or how long one will stay in jail. The promise of AI is to improve on human decisions and save more lives, be it by avoiding accidents, optimizing organ donation chains, or preventing violent crime—but this promise can only come true if people accept that AI may handle the kind of moral tradeoffs that were, until now, the reserved grounds of humans. If people are unhappy with the way moral machines are programmed, they can make them irrelevant by opting out of their use. People can refuse to buy self-driving cars, can opt out of being organ donors, and can vote out judges or politicians who allow the use of algorithms in court. To avoid this Ethical Opt-Out, behavioral scientists must give people a voice, by using the methods of moral psychology to assess citizens' preferences about the ways machines should handle ethical tradeoffs. This is a challenging task, for behavioral scientists will have to find a way to adapt the methods of moral psychology in order to tackle complex technical domains, which are likely to elicit complex moral preferences. Furthermore, behavioral scientists will have to tread carefully, and be mindful of second-order ethical challenges. But as we showed in this chapter, none of these challenges are intractable—and the stakes are great. Moral psychology has traditionally kept an eye on the past, be it the evolutionary past that shaped our moral intuitions, or the work of the great philosophers that formalized ethical theories. It is now time to turn an eye to the future, and to investigate the moral psychology of the newly possible.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*, 1573–1576.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (in press). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars. *Proceedings of the IEEE*.

Bostrom, N. (2014). *Superintelligence: Path, dangers, strategies*. Oxford: Oxford University Press.

Casey, M. J., Wen, X., Rehman, S., Santos, A. H., & Andreoni, K. A. (2015). Rethinking the advantage of zero-HLA mismatches in unrelated living donor kidney transplantation: implications on kidney paired donation. *Transplant International*, *28*, 401–409.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*, 114–126.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, *13*, 1–17.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Federal Probation*, *80*, 38–46.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15.

Freedman, R., Borg, J. S., Sinnott-Armstrong, W., Dickerson, J. P., & Conitzer, V. (2018). Adapting a kidney exchange algorithm to align with human values. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Gintis, H., Bowles, S., Boyd, R. T., Fehr, E., et al. (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life*. MIT press.

Goodall, N. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, *2424*, 58–65.

Greene, J. D. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Haidt, J. (2007). The new synthesis in moral psychology. *science*, *316*, 998–1002.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kramer, M. F., Borg, J. S., Conitzer, V., & Sinnott-Armstrong, W. (2018). When do people want AI to make decisions? *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: computation of

Bayes factors for common designs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-2)

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in neural information processing systems* (pp. 5684–5693).

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, *20*, 5–14.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*, 694–696.

Thaler, R., & Sunstein, C. S. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale: Yale University Press.

Trust in technology (Tech. Rep.). (2017). HSBC.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Walton, D. (2007). *Media argumentation: dialectic, persuasion and rhetoric*. Cambridge University Press.