# *From Public Preferences to Ethical Policy*

Julian Savulescu*, Guy Kahane, Christopher Gyngell

Julian Savulescu is Director of the Oxford Uehiro Centre for Practical Ethics
Co-Director Wellcome Centre for Ethics and Humanities, University of Oxford
Group leader, Biomedical Ethics Research Group, Murdoch Children's Research Institute
Visiting professor, Melbourne Law School, University of Melbourne

Guy Kahane is an Associate Professor at the Faculty of Philosophy, University of Oxford, Director of Studies at the Oxford Uehiro Centre for Practical Ethics, and Fellow and Tutor in Philosophy at Pembroke College, Oxford.

Christopher Gyngell is a Research Fellow in Biomedical Ethics at the University of Melbourne and Murdoch Children's Research Institute.

*Corresponding author email address: julian.savulescu@philosophy.ox.ac.uk

Standfirst: *Studies have provided rich data on global preferences for how autonomous vehicles (AVs) should act in collisions. We describe a framework for incorporating such preferences in policy. Preferences should inform the design of AVs only after being screened for bias and to the degree to which they match major ethical theories.*

Rapid advances in science and technology—think of gene editing, 'big data', or autonomous vehicles—raise urgent ethical dilemmas that policy-makers must address. One development that promises a possible solution is our growing ability to collect enormous amounts of data about people's moral preferences. An impressive recent example is Awad et al.'s 'Moral Machine' study[1], which collected 40 million moral decisions from people from 233 regions about dilemmas faced by autonomous vehicles. Awad et al. were able to identify robust global preferences—for example, to program autonomous vehicles to prioritise human over animal lives, to save more rather than fewer lives, and to save the young over the old. They were also able to identify cross-cultural ethical variation—people from southern countries, for example, tended to prioritise the lives of females and the physically fit. Such studies offer an unprecedented 'snapshot' of global moral preferences. What is unclear is how such data about people's preferences—or what philosophers call 'moral intuitions'—could be used to inform policy.

## *Basing policy on public views*

Policy-makers are interested in public views, but the way public opinion feeds into practice is inconsistent. In some cases, strong public opposition is used to justify forbidding a policy. For example, many policy bodies argue that technology enabling couples to select the sex of their child should be banned, not on principled ethical grounds, but because the public finds it repugnant.[2] In other cases, policy ignores public opinion. For example, voluntary assisted

dying has been largely banned in the UK, Australia and USA despite evidence that large majorities approve of its legalization.[3]

But policy should neither blindly follow people's moral intuitions nor entirely ignore them. It shouldn't follow them blindly because what people *want* done is one thing, and what morally *should* be done is another. Public views on moral questions can be deeply mistaken—there used to be broad support for racist and sexist policies. Or to pick a contemporary example, in many countries there is low support for organ donation. Yet this is rightly seen as a problem to overcome.[4] Moreover, psychologists have documented numerous ways in our intuitions can be systematically biased by morally irrelevant influences.

But neither should we just ignore the moral intuitions of the public. The alternative would be to let policy makers decide, perhaps aided by ethicists. But history shows that relying purely on abstract theory while ignoring common sentiments can also lead to disastrous decisions. Moreover, ethicists propose moral theories that would give conflicting answers to many ethical questions.

**An ethical framework for policy making**

The question then is how to draw on widely held moral intuitions while recognising that such intuitions can be inconsistent and biased. Intuitions can be an essential input to policy-making, but they should not be given the last word. Drawing on ideas from the political philosopher John Rawls, we suggest that two things need to be done before we can give intuitions a say.

(1) We first need to identify those intuitions that deserve to be taken seriously by making a conscientious effort to overcome bias. Rawls called the intuitions that survive such screening 'considered judgments': robust moral responses to situations that are based on careful reflection and clear understanding of the issues. While professional ethicists may seem best positioned to form such carefully considered intuitions, they are a tiny portion of the population, and highly idiosyncratic in their background and life experience. They may lack what Rawls described as the "sympathetic knowledge ... of those human interests which, by conflicting in particular cases, give rise to the need to make a moral decision."[5] The solution is to rely on both refined expert intuitions and widespread public responses. When both align we have convergence—and therefore further support to the view converged on. When the intuitions of ethicists (or other experts) and ordinary people diverge, we need to carefully scrutinise both sides: is the public response due to mere instinct, or failure to fully grasp the options? Or are the experts missing something that ordinary people spot?

(2) Critically, identifying a core set of intuitions isn't enough. That's still just data. We need to see if it hangs together, if there actually are good underlying reasons for our intuitions. After all, even our most considered and widespread judgments can be mistaken. So this intuitive data needs to be matched onto our more general ethical values and theories. When our considered intuitions are consistent with multiple robust ethical theories, this convergence is evidence that we are getting at something genuine. It also means that policy

based on these intuitions will enjoy what Rawls called 'overlapping consensus', endowing it with greater legitimacy.[6]

**(INSERT FIGURE 1)**

Caption: In our framework data on public and expert intuitions form the first step of a deliberative process. Intuitions are initially screened for bias and prejudice before coherence is sought between intuitions and ethical theory. Screened intuitions that best cohere with ethical theory form the basis of policy.

**Public intuitions about driverless cars**

Let's return to the example of driverless cars. How might the global preferences collected by Awad et al. inform policy in this area?

Awad et al. studied people's responses to what are known as 'Trolley Dilemmas'. In such dilemmas, a vehicle will hit one group of people if no action is taken, but it is possible to prevent the harm to the first group only by diverting the vehicle so that it would hit a second group. Awad et al. tested responses to 40 million in choices of injury or death to passengers of driverless cars versus to pedestrians. While they identified several factors which made people more likely to spare one group over another, we will consider three preferences: to save the greater number (a strong preference), to save younger lives over older lives (a moderate preference), and to save females over males. How should these figure in public policy?

We should first ask whether these preferences are unjustifiably discriminatory—treating individuals differently on the basis of differences that are morally irrelevant. An obvious example is the preference for females over males: sex is not a morally relevant factor, as it makes no difference to the value of someone's life; as we shall see, the major ethical theories would reject such a preference. It is possible that participants thought that women were more likely to be responsible for caring for others. This might be indirectly relevant but could not be accurately assessed by driverless cars. By contrast, there is a prima facie reason to save more people: human life is valuable, and saving more lives is better than saving fewer lives. Number can be directly assessed by AI. There is also a potential reason to save the younger: they have had less life and have more life ahead of them. The preferences that survive this initial screening stage can be called 'laundered preferences' and can serve as input for further ethical deliberation.

Next, we need to gather relevant ethical theories, concepts, principles, as well as professional guidelines and laws. For example, the German Federal Ministry of Transport and Digital Infrastructure's Ethics Commission has formed guidelines for driverless cars.[7] These guidelines represent an extreme form of *egalitarianism*: equal treatment for all. The German Federal Ministry's report states:

> *"In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical, or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice non-involved parties."*

On this egalitarian approach, no personal characteristics can be used when considering collisions between driverless cars and pedestrians. Strict egalitarianism denies a role to numbers—every individual must be given an equal chance of being saved. The German guidelines do allow for cars to be designed so as to generally minimise the number of people hurt, but it strictly forbids diverting a car in order to save the greater number.

We will compare the egalitarian view to two other ethical theories: *utilitarianism* and *contractualism*. In ethical deliberation, we try to match these candidate normative frameworks to 'laundered' preferences through a process of iterated discussion and reflection. Let's see how this works.

Utilitarianism is an ethical theory which defines the right act as the act which maximises utility, or the good. As Jeremy Bentham famously described it, the greatest good to the greatest number. Utilitarians would therefore save the greater number as well as younger people, since their future contains more utility.

Rawls himself was associated with a contractualist approach. Rawls argued that a fair choice is the one we would make behind a 'veil of ignorance', not knowing whether we would be the one advantaged or disadvantaged by the choice. Applied here, such a contractualist view would support saving the greater number: we would have a greater chance of surviving since we don't know whether we would be pedestrians or passengers. Contractualism would also reject sex as a criterion: if you didn't know whether you would be a man or a woman, you would want an equal chance for each.

What about age? In our own research, we found that when the differences in life expectancy are small (40 vs 41 years), the majority of people adopt an egalitarian approach – preferring to decide via a coin toss.[8] However, at more significant age differences, they become utilitarian and prioritise the young. This suggests that degree of difference is relevant in a contractualist framework. If people don't know whether they are going to be a 10-year-old or a 12-year-old in a crash, they will prefer policies that make no distinction between these groups. If the choice is between a 10-year-old and an 80-year-old, people prefer policies that favour the 10-year-old.

**(INSERT FIGURE 2)**

Caption: Different ethical theories either endorse or reject public preferences for driverless cars to take number, age, and sex into consideration in collisions. In general, we see broad coherence between ethical theories and the public preference to a take number into account. In contrast, all ethical theories reject sex as a relevant consideration. Age is in an intermediate position, enjoying support from some theories but not others.

Which policy should we form? Even though there is a public preference for saving females, all theories reject sex as a relevant factor. Age and numbers are features on which ethical theories disagree. Egalitarianism says age is never relevant, and rejects diverting a car to save the greater number. By contrast, and in line with strong public preferences, both utilitarianism and contractualism see numbers as relevant. They also see age as morally

relevant, but in different ways. In the face of this dispute, data about people's preference may support a policy that only takes significant age differences into account.

Evidence about public preferences can inform policy when coupled with ethical theory in this way. Ethical theory can also guide further research by identify potentially important factors for public deliberation. Given that there are factors that are always likely to be ruled out morally (e.g. preference based on race and sex), and some seen as morally crucial (e.g. number), data on public intuitions is likely to be most useful on topics where moral theories disagree (such as age). As a further example, Awad et al. did not report on whether and how responsibility for risk—e.g. does it matter whether the tragic choice is due to a mistake made by the autonomous vehicle—ought to be considered? The German guideline strictly forbid harm to third parties, but utilitarians are likely to disagree. Robust data on public intuitions about such matters will be especially valuable.

The policies and laws we adopt are ethical choices. Data about global preferences can contribute to ethical deliberation but it is no substitute. It is important to develop arguments that are both "convincing and able to convince".[9] In the public and political realm, this may take the form of presenting a policy or decision as one of common sense—even *global* common sense. But nonetheless it must also have robust ethical defence.

Ethics is important in a further way. Many real-life decisions involving AI will often be opaque to human observers. It is important that every such outcome be collected as a data point for ethical audit. Ethical evaluation will be necessary to give a plausible independent justification of these outcomes. If we cannot make such decisions ethically intelligible, we should conclude that the algorithm that produced them is ethically faulty. In this way, ethics is essential both to what we put into AI, and to assessing what we get out.

## Acknowledgments

JS and CG through their involvement with the Murdoch Children's Research Institute, received funding through from the Victorian State Government through the Operational Infrastructure Support (OIS) Program. JS was supported by the Wellcome Trust [WT 104848/Z/14/Z] and [WT203132/Z/16/Z]. All the funding bodies provided support for research on themes developed within this paper. The funders had no role in the conceptualization, decision to publish, or preparation of the manuscript.

## References

1. Awad, E. *et al. Nature* **563**, 59 (2018).

2. Kanellopoulou, N. *SCRIPT-ed* **1**, 217–223 (2004).

3. Sikora, J. & Lewins, F. *Health Sociology Review* **16**, 68–78 (2007).

4. Nordfalk, F., Olejaz, M., Jensen, A. M. B., Skovgaard, L. L. & Hoeyer, K. *Transplantation Research* **5**, (2016).

5. Rawls, J. *The Philosophical Review* **60**, 177–197 (1951).

6. Rawls, J. *Oxford Journal of Legal Studies* **7**, 1–25 (1987).

7. Luetge, C.. *Philosophy & Technology* **30**, 547–558 (2017).

8. Arora, C., Savulescu, J., Maslen, H., Selgelid, M. & Wilkinson, D. *BMC Medical Ethics* **17**, 69 (2016).

9. Dunn, M., Sheehan, M., Hope, T. & Parker, M. *Cambridge Quarterly of Healthcare Ethics* **21**, 466–480 (2012).

**Competing interests**

The authors declare no competing interests