AI & SOCIETY
https://doi.org/10.1007/s00146-023-01697-y

OPEN FORUM

Moral disagreement and artificial intelligence

Pamela Robinson[1]

Received: 10 May 2022 / Accepted: 15 May 2023

Abstract
Artificially intelligent systems will be used to make increasingly important decisions about us. Many of these decisions will have to be made without universal agreement about the relevant moral facts. For other kinds of disagreement, it is at least usually obvious what kind of solution is called for. What makes moral disagreement especially challenging is that there are three different ways of handling it. *Moral solutions* apply a moral theory or related principles and largely ignore the details of the disagreement. *Compromise solutions* apply a method of finding a compromise and taking information about the disagreement as input. *Epistemic solutions* apply an evidential rule that treats the details of the disagreement as evidence of moral truth. Proposals for all three kinds of solutions can be found in the AI ethics and value alignment literature, but little has been said to justify choosing one over the other. I argue that the choice is best framed in terms of *moral risk*.

Keywords moral disagreement · moral decision-making · value alignment

1 Why moral disagreement matters

Moral disagreement is commonly cited as an obstacle to building moral and value-aligned artificial intelligence (e.g., Bostrom 2014; Brundage 2014; Etzioni and Etzioni 2017; Formosa and Ryan 2020; Gabriel 2020.) It poses two main problems.

The *methodological problem*: How should we design artificially intelligent systems that align with morality or our values when neither the designers nor those affected by these systems can agree about what is moral or valuable? While we are sometimes in agreement about these things, we still inevitably find plenty to disagree about.[1] To respond to the methodological problem, we need a more foundational decision-making rule—that is, one more foundational than the moral theories or decision-making principles we disagree about. And the deeper our disagreement, the worse this problem becomes. In the worst case, it can interfere with our ability to find a methodology for decision-making in the face of moral disagreement that we can all accept.

The *skeptical problem* is that moral disagreement can—and perhaps should (e.g., Mackie 1977)—make us doubt the existence or standing of moral facts, and can make it irrational to maintain a high level of confidence in moral claims. This threatens to pull the rug out from under the whole enterprise of building *moral* AI.

The methodological problems raised by moral disagreement have traditionally been taken up by moral and political philosophers. Their aim (e.g., Rawls 2005; Gutmann and Thompson 1990, 1996; Wong 1992; Muldoon 2017; Mulligan 2020; O'Flynn and Setälä 2020) has been to find a just and fair methodology for decision-making in the face of deep and reasonable moral disagreement and to consider how the skeptical problems bear on this project (e.g., Enoch 2017; Kappel 2018; Carlson 2018; Edenberg 2021; van Wietmarschen 2018). The skeptical problems raised by moral disagreement have been discussed by metaethicists and epistemologists. Metaethicists have investigated whether moral disagreement is a threat to moral realism, moral knowledge, and objective moral truth (e.g., Brink 1984; Tolhurst 1987). Epistemologists have considered how rational people should adjust their confidence in moral claims in response to moral disagreement (e.g., McGrath 2008; Skipper and Steglich-Petersen 2020).

In this paper, I aim to make progress on the methodological problem posed by moral disagreement regarding AI ethics and value alignment. To simplify matters, I will frame the methodological problem as follows. Imagine that

✉ Pamela Robinson
pamela.robinson@anu.edu.au

[1] School of Philosophy, Australian National University, Level 6, 146 Ellery Crescent ANU, Canberra, ACT 0200, Australia

[1] Alexander (1999: 531–2) gives a compelling description of how significant moral disagreement can arise even among people with similar moral views and strong motivations to be moral.

Published online: 03 June 2023

we have built an 'AI Decider': an all-purpose, artificially intelligent decision-making system whose decisions affect people in morally significant ways. We can imagine giving the AI Decider any decision in any scenario with any amount of information and asking how, in general terms, it should decide. Thinking about the AI Decider allows us to abstract away from the limitations of any particular artificially intelligent decision-making system. It is just a way of implementing the best solution.

I will call those affected by the AI Decider's decision in a morally significant way its 'decision subjects'. Decision subjects might be directly affected (e.g., the AI Decider grants me entry to a country), indirectly affected (e.g., the AI Decider injures a friend in an unavoidable crash scenario), or more subtly affected by the AI Decider (e.g., it is a system that I designed, or it makes decisions in the stock market that slightly change the economic situation of my country). The questions of who counts as an AI Decider's decision subjects, how much weight each should get in the decision, and whether they are the only people whose input might matter are all interesting but are also beyond the scope of this paper. Here I will assume, for simplicity, that we know who all the decision subjects are, that they all get the same weight in the decision, and that no one else need to be taken into account in decision-making.

The methodological problem can now be put as: *how should the AI Decider decide in cases where its decision subjects have a (relevant) moral disagreement?* Or, if you prefer: *how should the AI Decider be designed to decide in these cases?* I do not assume that the AI Decider has moral or rational obligations in the same way that we do. I only assume that there are better and worse ways that the AI Decider can make its decision and that this, at least, matters to the moral and other normative evaluations of *our* choices about how to design artificially intelligent decision-making systems.

This way of setting things up puts many important questions to the side. One is: *under what conditions should we use artificial intelligence to make decisions in the face of moral disagreement*? If we try to avoid the obstacle posed by moral disagreement by *never* having artificial agents make these sorts of decisions, then we would give up most of what we stand to gain from artificial intelligence. For example, we can expect to benefit from the speed at which an AI Decider can make decisions, its ability to process complex data in ways that are difficult for us (e.g., Freedman et al. 2020), and the potential predictability of its decisions (e.g., Brennan-Marquez and Chiao 2021). Even the decisions of an 'Oracle AI', which only interacts with the world by offering advice (e.g., Bostrom 2014), can affect people in morally significant ways.[2] On the other hand, there are many good reasons not to use artificially intelligent systems to make every possible decision, and these range from considerations of autonomy and responsibility to transparency and safety.

Here I am only focusing on the kinds of decisions that we might actually want to design AI Deciders to make for us. One example is the decisions we are considering allowing autonomous vehicles make—like how to navigate unavoidable crash situations. These decisions directly affect everyone in these crashes and indirectly affect everyone in society. Assuming that we can develop autonomous vehicles that are safer than human drivers in most cases, we may want to allow the AIs in these vehicles to make decisions for us. And yet, there is plenty to morally disagree about when it comes to matters like who should be prioritized in an unavoidable crash.[3]

2 Why moral disagreement is especially challenging

To address the methodological problem, we might start by considering how the AI Decider should handle other kinds of disagreement. First, the AI Decider might encounter *descriptive disagreement,* in which decision subjects disagree about 'descriptive' facts as opposed to moral (or other normative) facts.[4] For example, Alice and Bob both want to go to the cinema, but Alice thinks it is open and Bob thinks it is closed. They have a descriptive disagreement about whether the cinema is open. Second, the AI Decider might encounter *preference disagreement*, in which the decision subjects do not disagree about any facts—they just like or want different things.[5] For example, Carla and Dan want to go on a holiday together, but Carla wants to go cross-country skiing and Dan wants to go snorkeling. Their preferences conflict in a way that makes it difficult to plan their trip.

[2] If it couldn't, it wouldn't be useful.

[3] E.g., Awad et al. 2018.

[4] By a 'moral fact' or 'normative fact', I just mean a moral or normative truth. This rules out moral or normative error theory, but does not presuppose realism or objectivism. However, as I use the term 'moral disagreement' in this paper, there is actual disagreement about some moral fact or other, and so some degree of objectivity is required (and I will sometimes talk as if objectivism is true). For example, given a version of cultural relativism according to which what is moral depends on our culture's practices, you and I could morally disagree about the moral facts if you and I belong to the same culture, but you and I might not morally disagree if we belong to different cultures and are simply both speaking truly about what is moral for each of us. I do not take a stand on whether this last case would be true disagreement or not; what is important is that it is not parallel to a standard case of descriptive disagreement about descriptive facts, and so is not what I mean by 'moral disagreement' in this paper.

[5] I use 'preference disagreement' here broadly, to cover all sorts of disagreements or clashes of goals or utility functions that aren't about matters of fact, regardless of whether they actually involve preferences.

Descriptive disagreements usually call for *epistemic solutions*. Epistemic solutions aim at the truth, and information about the disagreement is used as evidence about the truth of the matter being disagreed about. For example, if the cinema is open, then it is best if Alice and Bob go to the cinema. If it is closed, then it is best if they stay home. If the AI Decider does not have independent access to the truth of the matter, then it can take the fact that Alice thinks the cinema is open and Bob thinks it is closed as evidence about how likely the cinema is to be open. The AI Decider might, for example, assign a probability of 0.5 to the cinema's being open and decide that Alice and Bob should stay home, or that they should investigate further, etc. If the AI Decider knows in advance that the cinema is closed, then there is no need for it to respect or otherwise give weight to Alice's mistaken belief by deciding, for example, that Alice and Bob should go to the cinema anyway and try the door.

In contrast, preference disagreements call for *compromise solutions*. They aim at compromise, and information about the disagreement is not used as evidence about the truth of the matter being disagreed about, because there is no truth of the matter. Instead, information about the disagreement is used to ensure a fair or acceptable outcome. For example, if there is no fact about which of skiing or snorkeling is better, then the AI Decider would be making a mistake if it were to decide that Carla and Dan should go skiing because it is superior to snorkeling.

A *moral disagreement*, at least for the purpose of this paper, is different from both kinds of disagreement. It is not a case in which people simply 'have different values', but is one in which they disagree about what is of moral value, or about what is morally permissible, etc., where there is a fact of the matter about these things. For example, Emma and Fred disagree about whether they should stop going on vacations and instead donate their vacation money to charity. Emma thinks they should because those with the means to do so have a strong moral obligation to help those in need. Fred thinks they should not because there is no such obligation. This is not merely a case of preference disagreement, for we can imagine that both Emma and Fred equally desire to continue vacationing, and also equally desire to do so if, and only if, it is morally permissible.

How should the AI Decider handle cases of moral disagreement? It may seem that moral disagreement calls for a third kind of solution, which I will call a '*moral solution*'. If an epistemic solution uses information about a disagreement as evidence about the truth of the matter being disagreed about, and if a compromise solution uses information about a disagreement as input to a mechanism for finding a fair decision, a 'moral solution' in the sense intended here does not need to use this information about a disagreement at all. It just involves applying a moral theory or other related principles to the case to determine what ought to be done. Moral disagreement is often cited to explain why we cannot simply design AI systems to act in accordance with a specific moral theory like utilitarianism, Kantianism, virtue ethics, etc. The main difficulty is not that this approach could not produce a moral AI. (Though that is also a serious worry; see, e.g., Brundage 2014.) It is that *just choosing some moral theory* is not a good *method* for handling moral disagreement when the disagreement is over which theory is true. But it is a third kind of solution, and I will discuss some more sophisticated versions of it.

Moral disagreement is uniquely challenging because, unlike descriptive disagreement and preference disagreement, it is not obvious whether it calls for a moral solution, a compromise solution, or an epistemic solution.

2.1 Proposed versions of each solution

Some AI ethics and value alignment researchers have suggested that moral disagreement calls for a moral solution, some have suggested that it calls for a compromise solution, and others have suggested that it calls for an epistemic solution. I will give examples of some of the claims each group is inclined to make, and briefly describe the kinds of solutions that have been proposed.

2.1.1 Moral solutions

Here are some examples of suggestions for moral solutions:

> The reason we have moral philosophy is that there is more than one person on Earth. The approach that is most relevant for understanding how AI systems should be designed is often called *consequentialism*: the idea that choices should be judged according to expected consequences. The other two principal approaches are *deontological ethics* and *virtue ethics*, which are, very roughly, concerned with the moral character of actions and individuals… Absent any evidence of self-awareness on the part of machines, I think it makes little sense to build machines that are virtuous or that choose actions in accordance with moral rules if the consequences are highly undesirable for humanity. (Russell 2019: 217)
>
> The idea of human rights-congruent AI … has much to recommend it. If there is a global overlapping consensus concerning human rights, then AI can be aligned with human rights doctrine while avoiding the problems of domination and value imposition. (Gabriel 2020: 427)
>
> [One approach] focuses not on the values people already agree on, but rather on the principles they would agree upon if they were placed in a position where no one could impose their view on anyone else.

> To understand what principles would be chosen in this kind of situation, Rawls proposes a thought experiment in which parties select principles from behind a 'veil of ignorance'—a device that prevents them from knowing their own moral beliefs or the position they will occupy in society. … The outcome of deliberation under these conditions is principles that do not unduly favor some over others. Such principles are, therefore, *ex hypothesi*, fair. (Gabriel 2020: 429)

There are roughly three kinds of moral solutions. The first, and least sophisticated, is simply to choose one of the moral theories that people disagree about and then apply it to the situation despite the disagreement. Call this the *true moral theory approach*. If the true moral theory is chosen, then this is the best possible solution to the problem in one way, since the objectively moral choice is made. But any real-life application of this version of the solution would involve choosing the moral theory that one *thinks* is true and applying it to the situation despite the disagreement. And this approach is not guaranteed to arrive at an objectively moral choice. It is risky and does not offer a principled methodology, and it is perhaps for these reasons that it is difficult to find anyone who recommends it.

The second kind of moral solution is illustrated by the first two quotes. The idea is to find some general moral theory or principles that we can all agree on, at least for the decisions of the AI Decider. Call this *the agreement approach*. Russell, in the first quote, is proposing that we design AI to be consequentialist, though not specifically as a way of solving the problem of moral disagreement. However, what he says makes it clear that he is aware that there is some disagreement between moral theorists about the correct moral theory. He argues that the motivations for adopting one of the other moral theories do not apply to the question of how to design an AI Decider (at least if the AI Decider is not self-aware). And, assuming he is right about this, we may be able to agree that the AI Decider should be consequentialist. Gabriel, in the second quote, points out the benefits of appealing to principles of human rights if a 'global overlapping consensus' can be found about them. If agreement can be reached about some fundamental principles of human rights, then might be able to agree that an AI Decider should act in accordance with these.

The third kind of moral solution is described by Gabriel in the third quote. Call it the *hypothetical agreement approach*. It is very similar to the second version of the moral solution. But instead of looking for principles that we actually agree on, it aims instead for principles that we would hypothetically agree on under certain assumptions, like those defining Rawls' 'veil of ignorance'. Leben (2017) also proposes a 'Rawlsian algorithm' for autonomous vehicles, designed to predict what self-interested people would agree to in an unavoidable crash if they were blind to information about how they would actually expect to fare.

It may seem strange to count this as a version of a moral solution as opposed to a compromise solution since Rawls' main claim was that decisions made under the veil of ignorance would be *fair*. But one of the things that distinguishes what I call 'moral solutions' from what I call 'compromise solutions' is their relative insensitivity to the actual facts about the disagreement. In applying a Rawlsian solution like the one Gabriel or Leben suggest, the AI Decider would first consider what principles its decision subjects would agree to under certain idealized assumptions, and then go on to use those principles to determine which decision to make about the situation at hand. This could end up being a decision that is completely insensitive to any of the AI Decider's decision subjects' actual views or positions in the disagreement. The agreement approach is more sensitive to the details of the moral disagreement than the other two versions of the moral solution, but since it aims to find agreement at a more fundamental level (about moral theories or principles), it can also be indifferent to the specific conflicting moral views the decision subjects in ways that compromise solutions and epistemic solutions are not.

Each of the last two versions of moral solutions aim to find principles that we can agree on and are, therefore, more sophisticated and plausible than the first. However, they are also entirely limited by what we can (or would) agree on. And finding this kind of agreement will not always be possible. While actual agreement about some of the moral facts is somewhat common, finding agreement about *all* the relevant moral facts is much less likely. For example, even if Russell is right and we both should, and do, all agree that consequentialism is the right approach for designing ethical AI, it is unlikely that we will all agree about which version of consequentialism is right. Hypothetical agreement is much easier to find since we can idealize away anything that would cause disagreement. (We can do things like assume that everyone is rational and self-interested and unaware of who they will be, for example.) But this benefit does not come for free. We can now disagree about which specific hypothetical agreement approach is the right way to design an AI Decider or whether it correctly picks out the moral decisions. And this can leave us back with an actual disagreement problem where the hypothetical agreement approach is treated like another competing moral theory or set of moral principles that can be disagreed about. For example, one might claim that self-interested people behind a veil of ignorance would choose harm-minimizing self-driving vehicles that give no special weight to the interests of their passengers. I draw attention here to two further points of potential disagreement: (i) we can disagree about whether this is really what people would rationally choose, and this may depend on how we understand 'rational' or the veil of ignorance, or

the vehicles' harm-minimization algorithms, and (ii) we can disagree about how much this fact of hypothetical agreement bears on what we morally ought to do, supposing, for example, that most people in fact would prefer to have self-driving vehicles that give special weight to their passengers. My intention here is not to argue against Rawls or applications of his ideas, but only to point out that the hypothetical agreement approach does not obviously or completely solve the problem of moral disagreement.

For this reason, I think it is best to treat the agreement and hypothetical agreement approaches as initial, ground-clearing, steps in any full solution to the problem of moral disagreement. We can first see if any of the disagreement can be *dissolved* by looking for potential agreement, but then we need a different method to solve the often-unavoidable problem of remaining disagreement.

2.1.2 Compromise solutions

Here are some examples of calls for compromise solutions:

> …human beings hold a variety of reasonable but contrasting beliefs about value. … To avoid a situation in which some people simply impose their values on others, we need to ask a different question: In the absence of moral agreement, is there a fair way to decide what principles AI should align with? (Gabriel 2020: 425)
>
> [w]hat seems needed is a kind of compromise to overcome disagreements over issues of value. Insofar as we value the moral diversity of our political community, it should be recognized that autonomous vehicles pose primarily a political problem, not a moral one. (Himmelreich 2018: 676)
>
> [s]ince decisions on a 'fairness measure' and the related techniques for fair algorithms essentially involve choices between competing values, 'fairness' in algorithmic fairness should be conceptualized first and foremost as a *political* question and be resolved *politically*. (Wong 2020: 225)
>
> [g]iven that there is no universal agreement, even among humans on ethical values, social choice is a necessary tool to address the value alignment problem. (Prasad 2018: 291)

One version of a compromise solution is the *social choice approach*. (See, e.g., Prasad 2018; Gabriel 2020; Baum 2020.) Social choice theory offers methods for combining individual preferences or opinions into a single collective decision. Applying these methods might involve having the AI Decider ask its decision subjects to deliberate and vote,[6] or the AI Decider might elicit and aggregate their moral judgements on its own. An example of a social choice (aggregation) approach is Freedman et al.'s (2020) proposal for improving algorithms used to solve the kidney exchange matching problem. They elicit moral judgements about which characteristics should be used to prioritize people and how these characteristics trade off against one other. These judgements are combined and represented as single weights for each combination of characteristics, and information about these weights is used to improve existing matching algorithms. Skorburg et al. (2020) and Sinnott-Armstrong and Skorburg (2021) argue that this approach could be used to help improve ethical decision-making in many domains. This is not exactly a proposal that the AI Decider (in this case, the matching algorithm) do the aggregation itself. An example of one that does is Noothigattu et al.'s (2018) proposal for an AI Decider that would learn people's ethical opinions and then use a voting method to make a decision when there is disagreement.[7]

A similar approach is to treat cases of moral disagreement as multiobjective optimization problems, where the decision subjects' moral judgements are interpreted as information about morally valuable goals or 'objectives'. The AI Decider would then try to optimize for all of them at once. Multiobjective optimization problems can be, though do not have to be, solved with social choice methods. Vamplew et al. (2018) and Petersen (2020) propose *multiobjective optimization approaches* to the methodological problem posed by moral disagreement.

[6] One might wonder where democratic approaches involving deliberation fit in. (Such approaches might involve the AI Decider having its decision subjects deliberate on the matter first before it decides what to do.) I think they're best understood as versions of the compromise solution, but deliberation on its own cannot be the whole solution. It has the potential to do two things: (a) to create consensus, and (b) to create better-informed decision-subjects. But it won't always lead to consensus, and where it does not, some other approach will be needed to handle the remaining disagreement. And we'll also need to appeal to another approach to manage disagreement among better-informed decision-subjects. What I suggest is that deliberation may feature in the best solution to moral disagreement, but only as a first step in the hope of reducing the scope of the disagreement and improving the remaining disagreement problem in other ways. For some of the limitations of democratic deliberation as well as the prospects for integrating it with social choice, see List (2018).

[7] Another related approach with a much narrower scope could be inspired by Russell's (2019) proposal for 'provably beneficial AI'. His suggestion is that an inverse-reinforcement-learning-based AI Decider could learn our preferences and then use this information to make decisions that maximize our total degree of preference-satisfaction (or something similar). While Russell's suggestion is actually that we appeal to a moral theory like preference utilitarianism to handle cases where people have conflicting preferences (Russell 2019: 220), and he does not directly address the problem of *moral* disagreement, we can imagine this approach being used to handle disagreements about, e.g., how altruistic we wold prefer self-driving cars to be. If we interpret these disagreements as moral disagreements, then his approach would at least offer a way of finding a compromise.

2.1.3 Epistemic solutions

Here are some examples of the sorts of claims made by those who propose epistemic solutions to moral disagreements:

> the proper response to [moral disagreement] is to design machines to be fundamentally uncertain about morality. … [A] direct approach to overcoming [the problem of moral disagreement] is to assume that there is a correct moral theory which we are searching for, acknowledge that we are fundamentally uncertain about which moral theory is correct, and then act in such a way as to give some weight to the judgements of different theories. (Bogosian 2017: 591, 595)
>
> [there is] no consensus within moral philosophy as to which theory is correct, and [there are] divergent moral and ethical views across individuals in society. As a result, if an RL agent is to act as ethically as possible, it is reasonable that it should exhibit moral uncertainty. (Ecoffet and Lehman 2021)

An epistemic solution to moral disagreement involves having the AI Decider use the information about the decision subjects' disagreement as evidence about the moral facts. One way for this to go is for this evidence to inform the subjective probabilities the AI Decider assigns to various moral statements and theories. Then it can appeal to a rule for decision-making under moral uncertainty to determine what to do. Call this the *moral uncertainty approach*. Currently, the most popular rule is to *maximize expected choiceworthiness*, 'MEC', which would have the AI Decider maximize expected moral value (MacAskill 2014; MacAskill and Ord 2020; MacAskill et al. 2020).[8] If, for example, (a) it assigns a probability of 0.5 that Emma and Frank are obligated to donate their vacation money to charity and a probability of 0.5 that they are not, and (b) if it would be very morally bad for Emma and Frank to fail to meet this obligation if they have it, but only mildly bad for them to donate their vacation money even if they are not obligated to do so, then (c) an AI Decider following MEC might decide that they should donate the money—just to be safe.

Bogosian (2017) advocates for this approach, and Martinho et al. (2021) argue for it by comparing a system following MEC to one that has not been designed to be morally uncertain. Bhargava and Kim (2017) and Thomsen (2022) also favor the moral uncertainty approach.

A second kind of epistemic solution to moral disagreement might appeal to the method of reflective equilibrium. Call this the *reflective equilibrium approach*. The idea would be for the AI Decider to try to achieve an overarching and coherent moral view that incorporates as many of the ethical judgments of its decision subjects as possible. Anderson et al. (2006) MedEthEx involves a limited application of this approach. And Zhang and Conitzer (2019) describe a method of learning a single correct concept from a group of people who each have a 'noisy estimate of the correct concept', and suggest that this could be used to handle moral disagreement. This could also be interpreted as a reflective equilibrium approach.

Compromise solutions to moral disagreement are currently more popular than either moral solutions or epistemic solutions, at least among AI researchers. This is probably partly because moral disagreement is often treated as a kind of preference disagreement. Outside of moral philosophy, it is common to think that there are no objective moral facts, and so some of the authors I cite may favor compromise solutions for this reason. For these authors, it would be obvious that moral disagreements require compromise solutions because these are the only solutions it would make sense for them to have. But it would be a mistake to conclude that moral disagreements could not require compromise solutions if there really are objective moral facts. There are plausible reasons to think that moral disagreements could require compromise solutions either way.

I have argued that moral disagreements are especially challenging because, unlike descriptive disagreements and preference disagreements, it is not obvious which kind of solution they call for. However, before turning to try to figure this out, it is important to note that this is not the only reason that the methodological problem posed by moral disagreement is so difficult. Even if we know which kind of solution to use, we still face the daunting task of choosing between the different kinds of moral, compromise, or epistemic solutions and finding effective ways of implementing them. Furthermore, each of the seven approaches I have described—whether a version of a moral, compromise or epistemic solution—faces its own considerable theoretical difficulties.[9]

3 Initial comparison

I will divide grounds to adopt solutions into two groups: *pragmatic grounds*, which include considerations of acceptance, predictability, and safety; and *moral (and metaethical)*

[8] See also Lockhart (2000), Ross (2006), and Sepielli (2009).

[9] For obstacles to the social choice approach, see Baum (2020). Gabriel (2020) discusses limitations of both the version of the agreement approach that appeals to principles of human rights and the Rawlsian hypothetical agreement approach. For an explanation of the challenges to rules for decision-making under uncertainty and MEC in particular, see MacAskill et al. (2020). For a philosophical summary of some of the controversies about the method of reflective equilibrium in ethics, see Tersman (2018).

grounds, which include considerations of the decision-making process, proximity to moral truth, and metaethical disagreement.

3.1 Pragmatic grounds

3.1.1 Acceptance

One good reason to adopt a compromise solution is to ensure that the AI Decider's verdict is something that can be agreed to. Obvious benefits of ensuring that all decision subjects can accept the decision include things like avoiding physical conflict. Alexander (1999: 533–6), for example, argues that communities need to be able to solve moral disagreements "authoritatively" to avoid bad results, identifying three main benefits of authoritative decision-making: cooperation, reduction of error, and reduction in decision-making costs.

And consider, for example, this justification offered by Awad et al.:

> For consumers to switch from traditional human-driven cars to autonomous vehicles, and for the wider public to accept the proliferation of artificial intelligence-driven vehicles on their roads, both groups will need to understand the origins of the ethical principles that are programmed into these vehicles. In other words, even if ethicists were to agree on how autonomous vehicles should solve moral dilemmas, their work would be useless if citizens were to disagree with their solution, and thus opt out of the future that autonomous vehicles promise in lieu of the status quo. (Awad et al. 2018: 59)

The general point is that, if finding a consensus is required for decision subjects to receive the benefit of using the AI Decider, then that is what is most important. Getting at the truth is beside the point.

This consideration counts strongly in favor of some moral solutions and strongly against others. Where agreement can be found, the agreement approaches obviously do well. But where it cannot be found, they have nothing to offer—either to come to a decision or to enable all parties to the disagreement to accept it. And the true moral theory approach performs the worst since the true theory (or the theory one thinks is true) might not be one all can accept.[10]

[10] It might be argued here that the true moral theory would take facts about acceptance into consideration, and so would permit one to make a decision that would (e.g.) lead to conflict or other bad consequences due to lack of acceptance. However, while this may be true, it makes it even more apparent that this approach is incomplete if it is not supplemented with a clear method for choosing a correct moral theory. And when one searches for such a method, one may run up against candidates very similar to the other approaches I'm considering here.

The importance of acceptance may seem to be a good reason to reject epistemic solutions. But they might do an equally good job of ensuring that the AI Decider's decision is something that all could accept. Without knowing more about what it takes to be the sort of thing that all could accept, it seems that any solution that *gives every decision subject equal say*[11] could meet this criterion. Both the moral uncertainty approach and the reflective equilibrium approach could be implemented in a way that would ensure this. In addition, none of the compromise solutions considered can *guarantee* that the AI Decider's decision will be accepted by all.

3.1.2 Predictability

Some have stressed the importance of having predictable decisions. For example, Bostrom and Yudkowsky (2014: 317) draw an analogy between AI decision-making and the legal system, explaining that

> the job of the legal system is not necessarily to optimize society, but to provide a predictable environment within which citizens can optimize their own lives.

And Brennan-Marquez and Chiao (2021) claim that

> [t]here is value in the consistency of application even in contexts where there is fundamental disagreement about parameters. The promise of algorithmic decision-making in this type of context is not that it will resolve these fundamental disagreements, but that it will facilitate consistency and predictability.

If these authors are right, then it is not enough for the AI Decider to find a decision that its decision subjects can accept. Its decision subjects must also be able to predict its decisions in advance. This may ensure that they have some minimal understanding of the process the AI Decider uses, but the main concern here is that it provides reliability and stability.

It is plausible that predictability is a desirable property, at least for certain kinds of decisions. The question is, does this criterion give us more reason to favor one kind of solution over another? Every approach that has been considered will offer a certain amount of predictability since the AI Deciders using them will all make decisions on the basis of some general rule—a theory or set of principles, social choice method, rule for decision-making under moral uncertainty, etc.

There is a difference, however, between the moral solutions and the others. On compromise solutions and epistemic

[11] At least insofar as the decision subject is expected to be affected by the decision.

solutions, the AI Decider's decision will be determined in part by the actual, and conflicting, moral judgments of its decision subjects. Because of this, an AI Decider following one of these approaches will be expected to make different decisions depending on who its decision subjects are and what their moral views are. And so, in one sense, its decisions will be less predictable than those of an AI Decider following a single moral theory or set of principles. However, since predictability relies on so many factors, it is unclear how much of a measurable increase in predictability this would generate overall.

3.1.3 Safety

In ensuring that a highly intelligent AI Decider does what we want, designing it to follow any existing moral theory seems like a bad plan. One reason for this is that any plausible moral theory is extremely difficult, and perhaps impossible, to express with the kind of precision that may be required for this task. And even if we can rely on our highly intelligent AI Decider's grasp of English and common sense to bridge the gap and understand what we are after, any room for misunderstanding could be disastrous. Furthermore, all moral theories seem to say very wrong things about at least some cases.

For this reason, the true moral theory approach appears to be the least safe. The agreement approaches are plausibly much safer but are also incomplete. Compromise solutions also seem to have an advantage here. Because they ensure that the AI Decider's decisions will involve some combination of the moral views of its decision subjects, this could keep its decisions more tightly tethered to morality. While an AI Decider following total utilitarianism might decide to increase the world's population to maximum capacity in the pursuit of maximum total well-being, an AI Decider following a social choice approach, for example, would find the decision unpopular.

However, if this is a promising approach to AI safety, then the same can be said for epistemic solutions. They, too, involve some combination of (or hedging between) all the moral views of the AI Decider's decision subjects, and so might be expected to do just as well at keeping the AI Decider away from extremely unpopular decisions.[12]

It may be argued that it could be particularly dangerous to have an AI Decider that is uncertain between moral theories. Any way of implementing an epistemic solution will require decisions about what it considers as evidence, how it updates its probabilities in light of it, how it evaluates moral risk, and so on. Errors we make in deciding *these* things could have catastrophic consequences, and the current state of the research on decision-making under moral uncertainty has already pointed to some areas of concern (e.g., MacAskill et al. 2020 Ch. 6).

However, as Baum (2020: 167) points out, the social choice approach has parallel kinds of risks:

> There is no single aggregate ethical view of society. Instead, there are many aggregate views depending on how the views are aggregated. These different aggregations can have very different consequences, some of which could be considered pathological or even catastrophic.

3.2 Moral (and metaethical) grounds

3.2.1 The decision-making process

Some solutions may be ruled out if we must ensure that the *process* the AI Decider uses to make decisions in the face of moral disagreement is just, appropriate, responsive to reasons, or something to that effect.

This consideration might be used to rule out any version of the true moral theory approach according to which someone—e.g., the designer of the AI Decider—is allowed to choose her own favorite moral theory for the AI Decider to follow. But it does not clearly rule out any of the other approaches wholesale. Instead, what it most obviously rules out are specific versions of the other approaches.

For example, considerations of procedural justice might be used to rule out *predictive* versions of some of the approaches mentioned. Suppose that the AI Decider does not actually aggregate moral judgments, become morally uncertain, or use a method of reflective equilibrium. Instead, it simply predicts what the result of carrying out each process might be. It might be argued that the resulting decisions would not, e.g., be grounded in the right reasons. Or, we might imagine an AI Decider designed only to find decisions and rationalizations for those decisions based on what it predicts its decision subjects would be most likely to accept. Many will be rightly uneasy at the idea of being subject to an AI Decider like this.

3.2.2 Proximity to moral truth

Compromise solutions, it may be argued, have a better chance of ensuring that the AI Decider's decisions are actually or approximately moral than the 'method' of just choosing a moral theory and having the AI Decider follow it. One reason for this is that any such moral theory will likely have some counterintuitive implications. But there is also a considerable risk that we will just pick the wrong

[12] This first impression may not stand up to careful scrutiny, however. For example, both the problems of 'moral cluelessness' (e.g., Greaves 2016) and the 'infectiousness of nihilism' (e.g., MacAskill 2013) threaten to lead to the opposite result.

theory from the start if we attempt to follow the true moral theory approach. The social choice approach, for example, may be more likely to produce moral decisions than any single human decision-maker. Baum (2020: 167) suggests that the idea may be that

> better results are achieved when using the views of many individuals, as in the maxim 'wisdom of the crowd'… Thus, a market democracy could outperform a communist dictatorship because it empowers many people to contribute their unique insights.

While this has some plausibility, the 'wisdom of the crowd' may not apply as well to moral judgements. If finding the correct moral theory is more like finding the correct theory of physics, then relying on the votes of decision subjects to choose would result in an inaccurate 'folk morality', just as it might be expected to result in an inaccurate 'folk physics'.

However, insofar as compromise solutions do have a good chance of getting at or approximating the moral truth, it seems that epistemic solutions will have at least as good a chance. After all, that is what they are designed to do.[13]

3.2.3 Metaethical disagreement

Not everyone believes, or is certain, that there are moral facts. And confronting moral disagreement can itself cause metaethical disagreement and uncertainty (e.g., Beebe 2014). How should the AI Decider make decisions in light of this kind of disagreement? Here moral solutions seem to be of no use at all unless agreement can be found between moral skeptics' rational (non-moral) preferences and moral realists' moral preferences. Since the agreement solutions are already limited by the implausibility of always being able to find an agreement, this requirement would limit them further.

[13] One might argue that, if the true moral theory is a version of moral pluralism (e.g., following Ross 1930), it would be most accurately represented by a compromise solution that aggregates moral judgements or represents the plurality of values as a multiobjective decision problem, and that an epistemic approach like the moral uncertainty approach could never arrive at it. However, even if this is right, an AI Decider following a parallel moral uncertainty approach may produce exactly the same decisions, even if it represents the problem differently. For example, if you think it's important to eat food that's both healthy and delicious, then you will choose meals that rank highly on both measures. But if you think that only one of health or taste matters and you're not sure which, you'll still choose meals that rank highly on both measures, since it would be risky to choose meals that are slightly healthier but taste awful, or slightly tastier but terribly unhealthy. Further, moral pluralism may not be true, and an AI Decider following a moral uncertainty approach may have the advantage of being able to represent uncertainty about whether pluralism is true.

What of the other two kinds of solutions? Here is one potential argument for favoring compromise solutions over epistemic ones.

Metaethical disagreement favors compromise solutions

1. When making a decision in the face of moral disagreement, one cannot presuppose any contested moral claim (or the negation of any agreed-upon claim).
2. Most actual moral disagreements will be, in part, metaethical disagreements about the existence of moral facts.
3. Epistemic solutions presuppose that there are moral facts; compromise solutions do not.
4. Therefore, compromise solutions will be required in most cases.

One could even take this further and claim that, unless one is certain that a moral disagreement does not involve a metaethical disagreement (or is not one in which all decision subjects are moral anti-realists), one cannot presuppose the existence of moral facts.

But while some may find this argument compelling, there are two reasons for doubt. First, it is not easy to rationally adopt (1) while denying the stronger claim that *when making a decision in the face of any kind of disagreement, one cannot presuppose any contested claim*. And it seems possible that people who disagree about moral facts might also disagree about which kind of method should be used to make a decision in light of their disagreement. Either facts about which kind of method to use are moral facts, in which case (1) seems impossible to satisfy, or else no (legitimate) decision can be made in the face of this kind of disagreement, which does not seem right.

There is also a potential argument going the other direction:

Metaethical disagreement favors epistemic solutions

1. When there is any chance that there may be moral facts, one ought to presuppose that there are.
2. Unless all parties to a moral disagreement are certain that there are no moral facts, the decision to be made in the face of the disagreement should presuppose that there are moral facts.
3. Epistemic solutions presuppose that there are moral facts; compromise solutions do not.
4. Therefore, epistemic solutions will be required in most cases.

The reasoning behind (1) is that, if there are no moral facts, then it does not really matter what one does, and, so, one might as well assume that there are moral facts, since one has everything to gain and nothing to lose by doing so. However, if one can doubt the existence of moral facts

without doubting the existence of, say, facts about rationality, then it will still matter what one does if there are no moral facts—it will just matter rationally; not morally. So, this argument is equally unpersuasive.

Here is why thinking about metaethical disagreement cannot help us choose between compromise and epistemic solutions. If the goal is to make a *moral* decision in the face of a metaethical disagreement, then the fact that there is metaethical disagreement should not change anything. We have nothing to lose *morally* from presupposing that there are moral facts. But once we do, this still does not obviously favor epistemic solutions, since there may be moral reasons to favor compromise solutions. And if the goal is different—for example, that of making a *rational* decision, where one is self-interest is what matters as opposed to the interests of others—then we are either changing the topic, or the same problem can arise in a new form. We will still need to choose between compromise solutions and epistemic solutions.[14]

4 How to decide?

I can conclude at this point is that moral solutions are the weakest, as they are either implausible if they do not rely on finding agreement or incomplete if they do. But neither compromise solutions nor epistemic solutions are obviously better than the other on the pragmatic and moral grounds considered. What does this mean, then, for the prospects of solving the methodological problem? Can the AI Decider use either kind of solution to make decisions about decision subjects who morally disagree?

I am not satisfied with this answer. For one thing, there seems to be an important methodological difference between compromise solutions and epistemic solutions.

4.1 Different aims or different questions?

One possibility is that there are two different aims one might have in answering the methodological question:

> *Q1: How should the AI Decider decide in cases where its decision subjects have a (relevant) moral disagreement—given the aim of finding a solution all decision subjects can agree to?*
> *Q2: How should the AI Decider decide in cases where its decision subjects have a (relevant) moral disagreement—given the aim of ensuring or best approximating a moral decision?*

AI Deciders following compromise solutions use information from their decision subjects' moral disagreement to find some kind of fair compromise, so it is plausible to think that compromise solutions line up with a pragmatic aim and are best for those trying to answer Q1. AI Deciders following epistemic solutions use this information as evidence about the moral truth, so it is plausible to think that epistemic solutions line up with a moral-truth aim and are best for those trying to answer Q2. This cannot be *completely* right, since the arguments in the previous section show that both kinds of solutions might be justified on either pragmatic or moral (truth) grounds. But thinking of these two different aims could explain why there at least *appears* to be a more significant methodological difference between political and epistemic solutions.

However, this story is not completely satisfactory. To solve the methodological problem posed by moral disagreement, we need a solution that will ensure the AI Decider's decision is *both* morally acceptable *and* acceptable by its decision subjects. And epistemic and political solutions appear to be different because they offer two different approaches to this single problem—not because one does one thing and the other does the other thing.

A related possibility is that the methodological problem should be seen as raising two distinct questions:

> *Q3: How should the AI Decider decide in cases where its decision subjects have a (relevant) moral disagreement (and where there is no question of moral uncertainty)?*
> *Q4: How should the AI Decider decide in cases where its decision subjects have a (relevant) moral disagreement and this is grounds for moral uncertainty?*

Perhaps epistemic solutions are required when you are uncertain of the moral facts and compromise solutions are required otherwise. When moral disagreements are grounds for moral uncertainty, they call for epistemic solutions; when they do not, they call for compromise solutions. If this is right, proponents of each kind of solution might also have a deeper disagreement: those favoring epistemic solutions might think that moral disagreement is always grounds for moral uncertainty, while those favoring compromise solutions might think that moral disagreement often is not grounds for moral uncertainty.

This seems to promise a tidy explanation of the apparent methodological difference between epistemic and compromise solutions, and to identify exactly what our choice between them should turn on: whether or not moral disagreement is grounds for moral uncertainty. But is it right?

One thing that is unclear is why compromise solutions are not equally plausible in cases where moral disagreement is grounds for moral uncertainty. Consider the following case:

[14] The only difference is that these epistemic solutions will involve aiming at the truth about rationality instead of morality.

An AI Decider's decision subjects disagree morally about the truth of
M1: We should adopt a rule requiring the use of seatbelts.

Some think the rule is morally required because it would reduce the number of deaths in car accidents. Others find the rule morally repugnant because it infringes upon freedom and autonomy. It might be that this is grounds for moral uncertainty, and that the AI Decider should respond to this by becoming morally uncertain about M1. But this is compatible with the AI Decider remaining certain that

M2: The (morally) right way to make decisions in the face of moral disagreement is to use a compromise solution—e.g., a social choice approach.

M1 is a first-order moral claim. M2 is a second-order moral claim, since it is about the right way to handle disagreement about first-order moral claims.[15] The point here is that, even if disagreement about first-order claims is grounds to become uncertain about first-order claims, it need not be grounds for uncertainty about second-order moral claims. And so, it seems that M2 might be true, despite the fact that the disagreement is grounds for moral uncertainty about M1.

Epistemic solutions could also be appropriate ways of responding to moral disagreement when the disagreement is not grounds to be morally uncertain. This is because we can draw a distinction between cases in which the disagreement provides grounds for us to be morally uncertain, and ones where it provides grounds *for the AI Decider* to be morally uncertain. For example, one possible position to take on moral disagreement is that we do not need to become less confident about our moral beliefs when we find others with opposing beliefs. But, even if these views are right, we may still want an AI Decider to become morally uncertain in the face of moral disagreement. Humans, one might hold, have better access to the moral truth. An AI Decider's access is bound to be more indirect—through the evidence provided by its decision subjects. On this way of thinking, at least, epistemic solutions may be appropriate even if there are not grounds for us to be morally uncertain.

[15] Some prefer not to posit second-order moral claims like M2, and say that there could only be a *rationally* right way to make decisions in the face of moral disagreement. While I think it is perfectly acceptable to talk about the morally right way to make these decisions, M2 could also be treated as a claim about rationality. This should not affect my overall argument, but it would require more complex terminology. At the very least, we'd need 'normative disagreement' and 'normative uncertainty' to refer to disagreement and uncertainty about higher-order claims like M2.

4.2 Moral risk

I think the choice between epistemic and compromise solutions should ultimately come down to *moral risk*.[16] By 'moral risk', I mean the chance of getting things wrong morally and what you thereby risk. If a risk is *potential loss*, then moral risk is *potential moral loss*. For example, if you choose an option that, while it may not be best, is at least permitted by every plausible moral theory, you have taken a very small moral risk. On the other hand, if you take a chance and destroy the world, thinking that it would either be the morally best thing within your power, since it would end the suffering of all known sentient beings in the universe or would be the morally worst thing within your power, since it would end all the happiness and flourishing of every known sentient being in the universe, then you have taken a *massive* moral risk.

How might this apply to answering the methodological problem posed by moral disagreement? First, note that there can be higher 'layers' of moral disagreement and moral uncertainty than the ones I have considered so far. Return to the example of the AI Decider whose decision subjects morally disagree about M1. It becomes morally uncertain about M1 (that we should adopt a rule requiring the use of seatbelts) but is sure that M2 (that the right way to make decisions in the face of moral disagreement is to use a compromise solution). What happens if its decision subjects also disagree about M2? Perhaps different decision subjects favor each of all of the different approaches I have discussed. How should the AI Decider handle this second level of disagreement? Plausibly, this disagreement is grounds for moral uncertainty about M2. But it is less clear how this should affect the AI Decider's decision. Perhaps it can be certain of, or at least act in accordance with, the claim that:

M3a: The best way to handle this second level of disagreement is with a compromise solution.

Or perhaps it can be certain of, or at least act in accordance with, the claim that:

M3b: The best way to handle this second level of disagreement is with an epistemic solution.

This line of thinking can produce quite complex results from epistemic and political solutions. For example, here is how the AI Decider might implement M3b, the idea that the second level of disagreement calls for an epistemic solution:

[16] In this paper I'm assuming that there are moral facts. If it turns out that there aren't any facts about (e.g.) which moral theory is true or what's morally required, then compromise solutions may be the only ones that make sense.

Step 1. The AI Decider follows a version of an epistemic solution: the moral uncertainty approach. It uses all relevant information extracted from the decision subjects' moral disagreement about M2 as evidence about *which kind of approach for making decisions in the face of moral disagreement is correct*. It adjusts its subjective probabilities accordingly and applies a rule like MEC for decision-making under moral uncertainty. The output of this rule for making decisions under moral *uncertainty* is that the AI Decider should make decisions in the face of moral *disagreement* in accordance with a compromise solution: a social choice approach.

Step 2. The AI Decider follows the social choice approach, aggregating its decision subjects' moral judgements about different seatbelt policies (M1) to decide on a compromise policy—perhaps requiring only that a small fine is paid by anyone who fails to wear a seatbelt.

The point of drawing your attention to this extra kind of complexity is that, however deep the layers of disagreement or grounds for uncertainty go, it seems that the AI Decider will have to start somewhere, and either apply a compromise solution or an epistemic solution when it does.[17]

So, one choice between kinds of solutions is that between which kind of solution an AI Decider should start with—that is, whether the *first* rule it applies should be an implementation of an epistemic solution or whether the *first* rule it applies should be a compromise solution. This general kind of decision—about where to start or what to take for granted—is a crucial one in designing AI decision procedures more generally.

Consider a discussion that might take place between someone who favors the moral solution of just building an AI Decider to follow preference utilitarianism, and one who favors a moral uncertainty approach. The preference utilitarian might claim that the morally right way for the AI Decider to handle moral disagreement is for it to identify its decision subjects' preferences (which will reflect their different moral views to some extent) and then make a decision that best satisfies their collective preferences. The moral uncertaintist might protest that this is *risky* because it assumes so many contentious things. What if preference utilitarianism is wrong? And in implementing it, many morally relevant decisions will have to made—e.g., about how to elicit, represent, and aggregate preferences. Would not it be better to have the AI Decider be morally uncertain? Surely it would be safer to have it use its decision subjects' disagreement as evidence for how confident to be in different moral theories, and then to make a decision that hedges between them.

But is this approach really less risky? It may not seem so to someone who thinks that preference utilitarianism is very likely to be true. What's key, I think, is that *no matter what*, some important moral or morally relevant decisions are going to have to be made in designing the AI Decider, and *they will not be made with either complete agreement or with complete certainty*. A choice will have to be made between moral, compromise, and epistemic solutions, for one thing, and choices will need to be made within each category as well. Further choices will have to be made about how to implement these solutions. Call these unavoidable and contentious choices *morally risky design choices*.

Since we cannot escape making some morally risky design choices, the important question becomes how to choose between them.[18] Perhaps those who favor epistemic solutions think that the choices we need to make in choosing and implementing epistemic solutions are the least morally risky, and those who favor compromise solutions think that the choices we need to make in choosing and implementing compromise solutions are the least morally risky.

But which side is right? A lot will depend on the details of each case, and in particular, on the quality of the evidence or information being used and what it is used for. If the choice is between a compromise social choice approach where the AI Decider just gathers information about healthcare preferences from its decision subjects' Amazon purchases, this may be much riskier than an epistemic reflective equilibrium approach where the AI Decider elicits evidence about the right to free healthcare from the views of moral philosophers. But the same social choice approach combined with more relevant information may be less morally risky than the same reflective equilibrium approach combined with worse evidence (or a poor operationalization of the process of reflective equilibrium).

It might appear that, at least in general, epistemic solutions are better because they seem specifically designed to mitigate moral risk. But the question of whether they actually are any less risky deserves scrutiny. For one thing, there is an important difference between moral disagreement and purely descriptive disagreement. In simple descriptive disagreements, like Alice and Bob's, everyone can agree on what good evidence and proof looks like. Because this is often not so in cases of moral

[17] An even further layer of complexity is that some specific compromise and epistemic solutions might adopt or mimic other kinds of approaches. For example, one proposed for difficult cases of decision-making under moral uncertainty has us first represent the problem *as if* it's a social choice problem. See, e.g., MacAskill (2016).

[18] There are other methods for reducing moral risk that I'm bracketing here. One obvious example is the decision not to use an AI Decider for a particularly sensitive kind of decision-making (like when to launch nuclear missiles).

disagreement, it is a reason to worry that epistemic solutions might even carry *especially* morally risky design choices. They may make everything turn on how the AI Decider gathers and uses evidence about the moral facts, and we may be more, or less, sure about the right way to gather and use moral evidence than we are about the moral truths themselves.

I think these questions should be explored further. But what I propose here is that the overarching solution to the methodological problem of moral disagreement is one that aims to minimize moral risk. The answer to *How should the AI Decider decide in cases where its decision subjects have a (relevant) moral disagreement?* is *In the way that best minimizes moral risk*. While this may not be sufficient grounds for choosing any of the seven approaches or three kinds of solutions in particular, turning to questions of moral risk offers more fundamental grounds for weighing specific morally risky design choices.

5 Conclusion

This paper has been an investigation into the methodological problem posed by moral disagreement for AI ethics and value alignment. I have argued that moral disagreement is especially challenging because it is not clear whether it calls for a *moral*, *compromise* or *epistemic solution*, and examples of each solution can be found in the literature. I have argued that the best solution to managing moral disagreement is to treat it as a problem of managing moral risk. This, perhaps surprisingly, does not clearly favor any of the three kinds of solutions mentioned. It also raises a bunch of unanswered questions about how to identify and weigh morally risky design choices. But it can offer us an answer for how to decide between various solutions in specific cases, at least, and it offers a plausible explanation of what might divide proponents of each solution. It might even be a more fundamental answer to the methodological question that proponents of all three solutions could agree on.

Acknowledgements Research on this paper was supported by the ANU Humanising Machine Intelligence Grand Challenge project and ARC discovery grant DP170101394. The author also gratefully acknowledges Seth Lazar and participants at the 2021 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society and the 2021 Upstate Workshop on AI and Human Values for helpful feedback.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Alexander L (1999) "With Me, It's All er Nuthin'": formalism in law and morality. Univ Chicago Law Rev 66(3):530–565. https://doi.org/10.2307/1600416

Anderson M, Andersen SL, Armen C (2006) MedEthEx: a prototype medical ethics advisor. In: Proceedings of the 18th conference on innovative applications of artificial intelligence, vol 2, pp 1759–1765. AAAI Press, Boston

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon F, Rahwan I (2018) The moral machine experiment. Nature 563:59–64. https://doi.org/10.1038/s41586-018-0637-6

Baum S (2020) Social choice ethics in artificial intelligence. AI Soc 35(1):165–176. https://doi.org/10.1007/s00146-017-0760-1

Beebe JR (2014) How different kinds of disagreement impact folk metaethical judgments. In: Sarkissian H, Wright JC (eds) Advances in experimental moral psychology. Bloomsbury Academic, London, pp 167–187

Bhargava V, Kim TW (2017) Autonomous vehicles and moral uncertainty. In: Lin P, Abney K, Jenkins R (eds) Robot ethics 2.0: from autonomous cars to artificial intelligence. Oxford University Press, New York. https://doi.org/10.1093/oso/9780190652951.003.001

Bogosian K (2017) Implementation of moral uncertainty in intelligent machines. Mind Mach 27(4):591–608. https://doi.org/10.1007/s11023-017-9448-z

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

Bostrom N, Yudkowsky E (2014) The ethics of artificial intelligence. In: Frankish K, Ramsey WM (eds) The cambridge handbook of artificial intelligence. Cambridge University Press, Cambridge, pp 316–334. https://doi.org/10.1017/CBO9781139046855.020

Brennan-Marquez K, Chiao V (2021) Algorithmic decision-making when humans disagree on ends. New Crim Law Rev 24(3):275–300. https://doi.org/10.1525/nclr.2021.24.3.275

Brink O (1984) Moral realism and the sceptical arguments from disagreement and queerness. Australasian Journal of Philosophy 62(2):111–125. https://doi.org/10.1080/00048408412341311

Brundage M (2014) Limitations and risks of machine ethics. J Exp Theor Artif Intell 26(3):355–372. https://doi.org/10.1080/0952813X.2014.895108

Carlson J (2018) Epistemology of disagreement, bias, and political deliberation: the problems for a conciliary democracy. Topoi. https://doi.org/10.1007/s11245-018-9607-8

Ecoffet A, Lehman J (2021) Reinforcement learning under moral uncertainty. Proceedings of the 38th International Conference on Machine Learning, pp 2926–2936
Edenberg E (2021) Political disagreement: epistemic or civic peers? In: Hannon M, de Ridder J (eds) Routledge handbook of political epistemology. Routledge, London
Enoch D (2017) Political philosophy and epistemology: the case of public reason. In: Sobel D, Vallentyne P, Wall S (eds) Oxford studies in political philosophy, vol 3. Oxford University Press, Oxford, pp 132–165. https://doi.org/10.1093/oso/9780198801221.003.0007
Etzioni A, Etzioni O (2017) Incorporating ethics into artificial intelligence. J Ethics 21:403–418. https://doi.org/10.1007/s10892-017-9252-2
Formosa P, Ryan M (2020) Making moral machines: why we need artificial moral agents. AI Soc. https://doi.org/10.1007/s00146-020-01089-6
Freedman R, Schaich Borg J, Sinnott-Armstrong W, Dickerson JP, Conitzer V (2020) Adapting a kidney exchange algorithm to align with human values. Artif Intell. https://doi.org/10.1016/j.artint.2020.103261
Gabriel I (2020) Artificial intelligence, values, and alignment. Mind Mach 30:411–437. https://doi.org/10.1007/s11023-020-09539-2
Greaves H (2016) Cluelessness. Proc Aristot Soc 116(3):311–339. https://doi.org/10.1093/arisoc/aow018
Gutmann A, Thompson D (1990) Moral conflict and political consensus. Ethics 101(1):64–88. https://doi.org/10.1086/293260
Gutmann A, Thompson D (1996) Democracy and disagreement. Cambridge University Press, Cambridge
Himmelreich J (2018) Never mind the trolley: the ethics of autonomous vehicles in mundane situations. Ethical Theory Moral Pract 21:669–684. https://doi.org/10.1007/s10677-018-9896-4
Kappel K (2018) How moral disagreement may ground principled moral compromise. Politics Philos Econ 17(1):75–96. https://doi.org/10.1177/1470594X17729132
Leben D (2017) A rawlsian algorithm for autonomous vehicles. Ethics Inf Technol 19:107–115. https://doi.org/10.1007/s10676-017-9419-3
List C (2018) Democratic deliberation and social choice: a review. In: Bächtiger A et al (eds) Oxford handbook of deliberative democracy. Oxford University Press, Oxford
Lockhart T (2000) Moral uncertainty and its consequences. Oxford University Press, Oxford
MacAskill W (2013) The infectiousness of nihilism. Ethics 123(3):508–520. https://doi.org/10.1086/669564
MacAskill W (2014) Normative uncertainty. PhD dissertation, Department of Philosophy, Oxford University, Oxford
MacAskill W (2016) Normative uncertainty as a voting problem. Mind 125(500):967–1004. https://doi.org/10.1093/mind/fzv169
MacAskill W, Ord T (2020) Why maximize expected choice-worthiness? Noûs 54(2):327–353. https://doi.org/10.1111/nous.12264
MacAskill W, Bykvist K, Ord T (2020) Moral uncertainty. Oxford University Press, Oxford
Mackie JL (1977) Ethics: inventing right and wrong. Penguin Books, Harmondsworth
Martinho A, Kroesen M, Chorus C (2021) An empirical approach to capture moral uncertainty in AI. Minds & Machines 31:215–237. https://doi.org/10.1007/s11023-021-09556-9
McGrath S (2008) Moral disagreement and moral expertise. In: Shafer-Landau R (ed) Oxford studies in metaethics, vol 3. Oxford, New York, pp 87–108
Muldoon R (2017) Exploring tradeoffs in accommodating moral diversity. Philos Stud 174(7):1871–1883. https://doi.org/10.1007/s11098-016-0825-x
Mulligan T (2020) Social choice or collective decision-making: what is politics all about? In: Kaul V, Salvatore I (eds) What is pluralism? Routledge India, London
Noothigattu R, Gaikwad SS, Awad E, Dsouza S, Rahwan I, Ravikumar P, Procaccia A (2018) A voting-based system for ethical decision making. Paper presented at the thirty-second AAAI conference on artificial intelligence, New Orleans, Louisiana, February 2–8
O'Flynn I, Setälä M (2020) Deliberative disagreement and compromise. Crit Rev Int Soc Pol Phil. https://doi.org/10.1080/13698230.2020.1737475
Petersen S (2020) Machines learning values. In: Liao SM (ed) Ethics of artificial intelligence. Oxford, New York, pp 413–435
Prasad M (2018) Social choice and the value alignment problem. In: Yampolsky RV (ed) Artificial intelligence safety and security. Chapman and Hall, London, pp 291–314
Rawls J (2005) Political liberalism, Expanded. Columbia University Press, New York
Ross WD (1930) The right and the good. Oxford University Press, Oxford
Ross J (2006) Rejecting ethical deflationism. Ethics 116:742–768. https://doi.org/10.1086/505234
Russell S (2019) Human compatible: AI and the problem of control. Penguin
Sepielli A (2009) What to do when you don't know what to do. In: Shafer-Landau R (ed) Oxford studies in metaethics. Oxford University Press, Oxford
Sinnott-Armstrong W, Skorburg JA (2021) How AI can aid bioethics. Journal of Practical Ethics 9(1). https://doi.org/10.3998/jpe.1175
Skipper M, Steglich-Petersen A (2021) When conciliation frustrates the epistemic priorities of groups. In: Broncano-Berrocal F, Carter JA (eds) The epistemology of group disagreement. Routledge, New York
Skorburg JA, Sinnott-Armstrong W, Conitzer V (2020) AI methods in bioethics. AJOB Empirical Bioethics 11(1):37–39. https://doi.org/10.1080/23294515.2019.1706206
Tersman F (2018) Recent work on reflective equilibrium and method in ethics. Philos Compass 13(6):e12493. https://doi.org/10.1111/phc3.12493
Thomsen F (2022) Iudicium ex machinae: the ethical challenges of automated decision-making in criminal sentencing. In: Roberts J, Ryberg J (eds) Sentencing and artificial intelligence. Oxford University Press, Oxford
Tolhurst W (1987) The argument from moral disagreement. Ethics 97(3):610–621. https://doi.org/10.1086/292869
Vamplew P, Dazeley R, Foale C, Firmin S, Mummery J (2018) Human-aligned artificial intelligence is a multiobjective problem. Ethics Inf Technol 20:27–40. https://doi.org/10.1007/s10676-017-9440-6
van Wietmarschen H (2018) Reasonable citizens and epistemic peers: a skeptical problem for political liberalism. J Political Philos 26(4):486–507. https://doi.org/10.1111/jopp.12152
Wong DB (1992) Coping with moral conflict and ambiguity. Ethics 102(4):763–784. https://doi.org/10.1086/293447
Wong P-H (2020) Democratizing algorithmic fairness. Philos Technol 33:225–244. https://doi.org/10.1007/s13347-019-00355
Zhang H, Conitzer V (2019) A PAC framework for aggregating agents' judgments. Proc AAAI Conf Artif Intell 33(1):2237–2244. https://doi.org/10.1609/aaai.v33i01.33012237