Disagreement, AI alignment, and bargaining

Harry R. Lloyd

AFFILIATIONS: (1) Yale University, Department of Philosophy

(2) Center for AI Safety, Research Affiliate

ABSTRACT: New AI technologies have the potential to cause unintended harms in diverse domains including warfare, judicial sentencing, biomedicine and governance. One strategy for realising the benefits of AI whilst avoiding its potential dangers is to ensure that new AIs are properly 'aligned' with some form of 'alignment target.' One danger of this strategy is that – dependent on the alignment target chosen – our AIs might optimise for objectives that reflect the values only of a certain subset of society, and that do not take into account alternative views about what constitutes desirable and safe behaviour for AI agents. In response to this problem, several AI ethicists have suggested alignment targets that are designed to be sensitive to widespread normative disagreement amongst the relevant stakeholders. Authors inspired by voting theory have suggested that AIs should be aligned with the verdicts of actual or simulated '*moral parliaments*' whose members represent the normative views of the relevant stakeholders. Other authors inspired by decision theory and the philosophical literature on moral uncertainty have suggested that AIs should *maximise socially expected choiceworthiness*. In this paper, I argue that both of these proposals face several important problems. In particular, they fail to select attractive 'compromise options' in cases where such options are available. I go on to propose and defend an alternative, bargaining-theoretic alignment target, which avoids the problems associated with the voting- and decision-theoretic approaches.

1: Introduction[1]

In the spring of 2020, a Kargu-2 lethal autonomous drone hunted down and attacked a human target in Libya. This was perhaps the first time in history that a human being was autonomously targeted by a lethal autonomous weapon (LAW). Recent advances in machine learning have created the possibility of much more sophisticated, artificially intelligent LAWs being deployed in the near future (Klare 2023).

Warfare is one of several domains in which new AI technologies have the potential to cause significant and unintended harms to human beings (and other sentient creatures). Some commentators worry that LAWs increase the risks of civilian collateral damage in cases where AIs target adversaries in densely populated areas (Scharre 2016, chapter 5; Marijan 2022); that judicial AI sentencing tools discriminate against ethnic minorities (Angwin et al. 2016); that biomedical AI assistants will reduce the barriers to entry for bad actors wishing to synthesize dangerous (and

[1] This introduction partially overlaps with §1 of Lloyd 2024, a companion piece to this paper.

potentially novel) pathogens and chemical weapons (D'Alessandro, Lloyd and Sharadin 2023); and that even more sophisticated future AIs might engage in deceptive or 'power-seeking' behaviours (Hendrycks, Mazeika and Woodside 2023, §5; Bales, D'Alessandro and Kirk-Giannini 2024).

What we want is to realise these potential benefits of AI whilst avoiding potential dangers. One strategy for promoting this objective is ensuring that new AIs are properly *aligned* with human values. Roughly speaking, a 'well aligned' AI is an AI that "pursue[s] goals that match with human values or interests rather than unintended or undesirable goals" (Ngo et al. 2023, p. 1).

The alignment project can be split into two problems: (1) deciding what target one will try to align AIs with – call this the Normative Problem; and (2) technically implementing this chosen alignment proposal – call this the Implementation Problem (Gabriel 2020, pp. 412-3). The Implementation Problem has been the subject of considerable technical research, and is much more difficult to solve than one might at first expect. I will not discuss implementation difficulties in this paper. Instead, I will focus exclusively on the Normative Problem.

One important danger concerning the Normative Problem of alignment is that our AIs will optimise for objectives that only reflect the values of a certain subset of society, and that do not take into account alternative views about what constitutes desirable and safe behaviour (Himmelreich 2018; 2020; Peterson 2019; Gabriel 2020; Robinson forthcoming). For instance, imagine that an LAW has the chance to neutralise several enemy combatants by means of destroying a site of considerable cultural importance. Reasonable people can and will disagree about how the LAW should act in this choice situation. Some will argue that killing the enemy combatants is much more important than preserving the cultural site. Others will argue that there is a strong obligation to preserve sites that are important to people's identities and cultures. In the face of this kind of moral disagreement, there would be something objectionable about an AI whose objectives only reflected one or the other of these two extremal positions.[2] It would be more appropriate for the AI's machine ethic to reflect a median or compromise position on the value of the cultural site.[3]

Several AI ethicists have already suggested targets for AI alignment that are designed to be sensitive to widespread normative disagreement amongst the relevant stakeholders. Authors inspired by voting theory have suggested that AIs should be aligned with the verdicts of actual or simulated '*moral parliaments*' whose members represent the normative views of the AI's stakeholders (Conitzer et al. 2017; Noothigattu et al. 2018; Lee et al. 2019; Prasad 2019; Gabriel 2020, §4.4; Hendrycks and Mazeika 2022, p. 18; Mayhew et al. 2022; see also Koster et al. 2022; Conitzer et al. 2024). Other authors inspired by decision theory and the philosophical literature on moral uncertainty have suggested that AIs should *maximise socially expected choiceworthiness*

[2] Even committed moral objectivists should think that the correct theory of morality would probably require us to do some compromising if there is widespread moral disagreement, for either intrinsic or instrumental reasons (cf. Wong 1992, and §2 of the present paper). I thank an anonymous reviewer for pressing me to include this remark.

[3] For further examples of tensions between competing values in AI ethics, see Whittlestone et al. 2019.

(MSEC) (Bogosian 2017; Ecoffet and Lehman 2021; Martinho et al. 2021; Thomsen 2022; Takeshita et al. 2023; see also Bhargava and Kim 2017).[4]

I discuss both of these proposals in more detail in §§3-4 of this paper, where I argue that they both face several important problems. Then, in §5, I present and defend an alternative bargaining-theoretic alignment target; in §6, I conclude. Before all of that, in §2 I say a little more to motivate the idea that AI alignment targets should be sensitive to normative disagreements amongst stakeholders.

I also want to mention that the Normative Problem of AI alignment can be understood as one subcase of the more general political problem of deciding how social disagreements should be resolved under conditions of widespread moral dissensus. Thus, many of my criticisms of moral parliaments and MSEC are also applicable more generally to voting- and decision-theoretic approaches in other social domains. Likewise, many of my arguments in favour of a bargaining-theoretic approach can be similarly generalised. The particular, AI-facilitated version of inter-stakeholder bargaining that I defend in §5 below can be understood as a form of '*algorithmic governance*' (cf. Conitzer et al. 2016; Danaher et al. 2017; Lee et al. 2019; Grisenko and Wood 2022). Future research could investigate the topic-specific advantages and disadvantages of this form of social decision making in other possible domains of application. In this paper, however, I focus exclusively on its advantages in the AI alignment domain.

2: The importance of disagreement

Much of the literature on AI ethics focuses on *fairness* concerns. For instance, considerable attention has been devoted to racial and gender biases in AI tools designed for domains including criminal sentencing, recruitment, and image classification.

What does it mean to treat disadvantaged groups fairly in this context? One important requirement is to ensure that the *effects* of these AI systems are fair on those disadvantaged groups. However, another arguably important requirement for fairness here is to ensure that these disadvantaged groups have a *fair say* in determining the criteria under which the effects of AI systems should be judged as 'fair' or 'unfair.' In other words: fair AIs should optimise for objectives that at least partially reflect the values of all of the communities who those AIs will affect.[5]

Moreover, the potential for normative disagreement is particularly salient in the context of AI fairness. AI ethicists have proposed numerous *prima facie* plausible 'fairness criteria' for AI systems. Unfortunately, several impossibility theorems have recently demonstrated that no single

[4] Note that the boundary between these two different approaches is quite blurry. For instance, although Ecofett and Lehman (2021) are primarily inspired by the literature on moral uncertainty, and endorse a version of MSEC, they nonetheless couch their proposals in terms of voting procedures. Still other authors have proposed *ad hoc* approaches tailored to particular AI alignment contexts (e.g. Conitzer et al. 2016; Lera-Leri et al. 2022). I will not discuss those *ad hoc* proposals in this paper.

[5] This requirement is particularly important given that several survey studies suggest that AI professionals' normative judgements differ significantly from the normative judgements of other stakeholders (Pierson 2018; Jakesch et al. 2022).

AI system can jointly satisfy all of these criteria (Kleinberg et al. 2016; Chouldechova 2017; Corbett-Davies et al. 2017; Miconi 2017).[6] In the face of these impossibility theorems, it is inevitable that there will be social disagreement about what is required for fairness in AI systems – as several studies have confirmed (Grgić-Hlača et al. 2018; Pierson 2018).

A second strand in the AI literature focuses on *safety* concerns. This strand in the literature is focused on the danger of AIs behaving in ways that are uncontroversially 'catastrophic,' or at least highly undesirable. For instance, some authors worry about biomedical AIs assisting in the synthesis of novel pathogens that might kill millions of people (D'Alessandro, Lloyd and Sharadin 2023). Other authors worry about superintelligent future AIs resisting shutdown by their human creators (Hendrycks, Mazeika and Woodside 2023, §5).

At first glance, it might be unclear why those who are primarily worried about these kinds of catastrophic outcomes should be at all concerned about whether or not AIs will optimise for objectives that only reflect the values of a certain subset of society. After all, *uncontroversially* catastrophic outcomes will be disvalued by *any* AI that is aligned with an at least minimally reasonable system of human values.[7]

However, there are at least two reasons why AI safety theorists should be concerned with normative disagreements in the context of AI alignment. Firstly, I argue in §4 below that alignment with MSEC at least in principle creates a danger of AIs behaving in ways that *many* of us regard as highly undesirable or even unsafe. This might be almost as bad as AIs behaving in ways that are uncontroversially catastrophic. Thus, AI safety theorists should be concerned about MSEC, and should want to find better alternatives.

Secondly, there might also be instrumental reasons for AI safety advocates to care about finding an attractive alignment target that in some sense reflects the values of all relevant stakeholders. The project of aligning AIs with human values is arguably more likely to succeed if it can command a broad base of support. But the alignment project is unlikely to command a broad base of support if its intended alignment target only reflects the values of a certain subset of society (likewise Bogosian 2017; Robinson forthcoming, §3.1.1).[8] It is just bad politics for AI safety proponents to advocate alignment with potentially controversial conceptions of desirable AI behaviour such as total welfare maximisation (cf. Russell 2019, pp. 217-21; Walker 2019).

In summary: both AI fairness and AI safety advocates should support the project of finding a tenable AI alignment target that is sensitive to normative disagreements between the relevant stakeholders. I now turn my attention to that project.

3: Parliaments

[6] Although cf. Hedden 2021 for an argument that "among the statistical criteria of fairness discussed in the literature, none except perhaps calibration is a genuine necessary condition on fairness for predictive algorithms."

[7] Although for some important complications, see D'Alessandro forthcoming; Sharadin forthcoming.

[8] According to Baum et al. (2022), one of the key historical "lessons for artificial intelligence from other global risks" is that it is particularly important to achieve broad buy-in from a wide range of stakeholders.

3.1: Simulated and unsimulated versions

There are two quite different versions of the parliamentary approach to alignment. According to the *unsimulated* version of this approach, AI alignment targets should be chosen democratically by actual human voters (Gabriel 2020, §4.4; see also Koster et al. 2022). One way to implement this idea would be to convene special-purpose citizens' assemblies, tasked with deliberating on and then selecting alignment targets for AIs.[9] Another way to implement the idea would be for existing democratic institutions to appoint new commissions or subcommittees.

The second possible version of the parliamentary approach to alignment is the *simulation* version. On this version of the approach, any given AI's ethical objectives are determined by an AI simulation of a parliament composed of or at least representing that AI's community of stakeholders (Conitzer et al. 2017; Noothigattu et al. 2018; Lee et al. 2019; Hendrycks and Mazeika 2022, p. 18; Mayhew et al. 2022). For instance, Noothigattu et al. and Lee et al. have both developed systems that learn a model of the normative views of every stakeholder on the basis of each stakeholder's moral judgements in a relatively small number of cases (Noothigattu et al. 2018; Lee et al. 2019; see also Kim et al. 2018; Freedman et al. 2020; Martinho et al. 2021). "At runtime, when encountering an ethical dilemma involving a specific subset of alternatives," an AI should estimate the preferences of every stakeholder "over this particular subset, and [then] apply a voting rule to aggregate these preferences into a collective decision" (Noothigattu et al. 2018). Similarly, Lee et al. (2019) use the Borda rule voting method to determine how an AI's estimates of its stakeholders' preferences should be aggregated into a decision about how the AI should behave.

3.2: Advantages and disadvantages

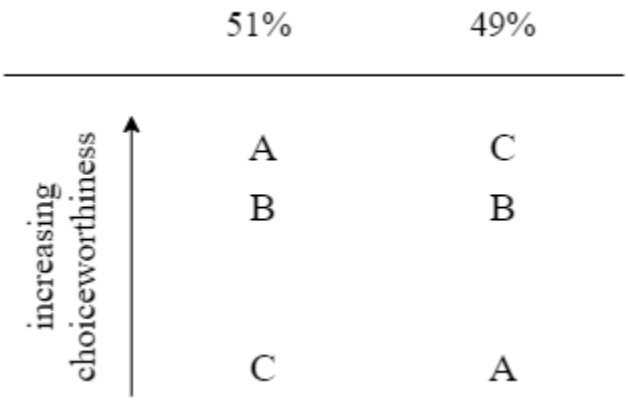
Some of the advantages and disadvantages of the unsimulated version of the parliamentary approach differ from those of the simulation version. For instance, many of us might regard the decisions of unsimulated democratic assemblies as carrying greater legitimacy than the decisions of mere simulations.[10] On the other hand, unsimulated assemblies will not be able to issue decisions as fine-grained as those that simulated parliaments allow for. (Recall that Noothigattu et al. suggest that the simulated parliament should vote every time the AI faces a new choice situation.) Similarly, unsimulated assemblies will be much slower than simulations in reaching their verdicts.

Both versions of the parliamentary approach also have some disadvantages in common. One such disadvantage is the 'tyranny of the majority': in plurality- and majority-rule voting systems, a coalition of 51% of the stakeholders can legislate in favour of any outcomes they like, regardless of the preferences of the other 49%. This means that parliamentary approaches can fail to select attractive compromise options (see also Ecoffet and Lehman 2021). For instance, consider the following scenario:

[9] On deliberative citizens' assemblies, see Gastil and Richards 2013; Vandamme and Verret-Hamelin 2017.
[10] However, Lee et al.'s survey of participants' perceptions of their simulated parliament suggested that "the framework successfully enabled participants to build models that they felt confident represented their own beliefs" (2019, p. 1).

> **Jackson:** some moral parliament (unsimulated or simulated) faces a choice between three options, A, B, and C. 51% of parliamentarians think that A is the best, B is almost as good, but C is terrible. 49% of parliamentarians think that C is the best, B is almost as good, but A is terrible (illustrated in figure #1).[11]

In this situation, a plurality- or majority-rule voting system will select option A. Yet many of us intuit, to the contrary, that it would be better for our alignment approach to select option B in this choice situation. After all, every stakeholder agrees that option B is almost as good as the best possible option. By contrast, 49% of stakeholders think that option A is terrible. In this situation, plurality- and majority-rule voting can fail to select an attractive compromise option.[12]



(Figure #1)

Another disadvantage of the parliamentary approach is closely related to the majority tyranny problem. This second disadvantage is hypersensitivity to small differences in stakeholder opinion, in a way that seem unnecessarily severe. For instance, consider the following scenario:

> **Biorisk:** some research scientists are using a biomedical AI to assist them in gain-of-function research. The biomedical AI has a certain fixed amount of compute at its disposal, each unit of which it can use for one of two purposes: (1) assisting the scientists in their gain-of-function research, and (2) performing simulations to determine how safe the research is likely to be. Furthermore, imagine that the stakeholders for this biomedical AI disagree on how its compute should be divided between these two possible uses. Roughly half of the stakeholders believe that we should trust academic scientists to determine for themselves whether their research is safe. Hence, these stakeholders believe that all of the

[11] This case is inspired by Jackson 1991. (A case with the same structure also appears in Regan 1980.)

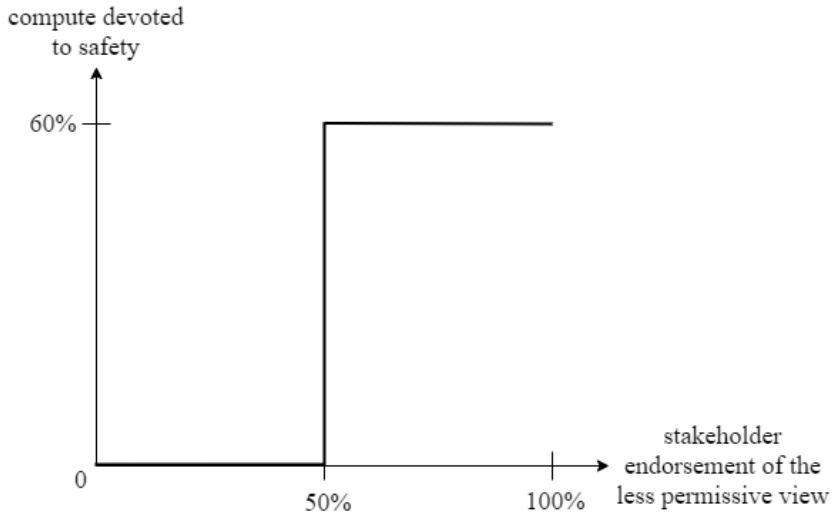
[12] Feffer et al. (2023) argue that Noothigattu et al.'s (2018) more complicated voting system also suffers from a tyranny of the majority problem. One might wonder whether the parliamentary approach could avoid this problem by adopting a supermajority-rule voting system. Unfortunately, supermajority requirements (1) still allow for tyrannies of the supermajority; and (2) create problems in cases where no option can secure supermajority support.

An anonymous referee also points out that the Borda Count voting method likewise results in a tyranny of the majority in **Jackson**, even though the Borda Count was introduced to avoid this flaw.

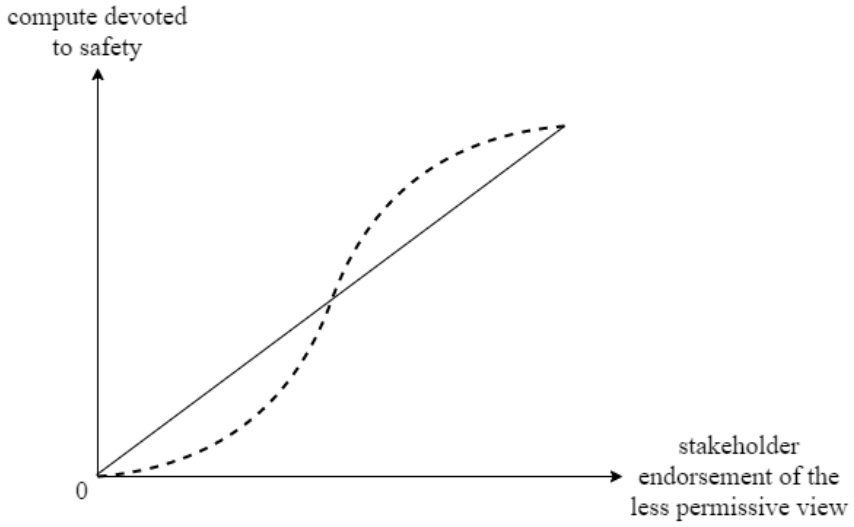
AI's compute should be spent on assisting the gain-of-function research. Call this the 'more permissive view.' By contrast, the rest of the stakeholders believe that a significant proportion – let's say 60% – of the AI's compute should be spent on determining how safe the research is likely to be before the AI decides whether or not it should assist the scientists. Call this the 'less permissive view.' Finally, suppose for sake of simplicity that gain-of-function and safety research both have constant returns to scale in compute.

Imagine a survey of the stakeholders suggests that they are split 51% to 49% between the more and less permissive views. Thus, this survey suggests that in a representative stakeholder parliament, the more permissive faction should have 51% of the delegates. Under a plurality- or majority-rule voting system, this faction can then legislate for the AI to spend all of its compute on assisting the scientists' gain-of-function research. However, now imagine we learn that the original survey was slightly mistaken. A more reliable survey now suggests that 49% of stakeholders endorse the more permissive view, and 51% of stakeholders endorse the less permissive view. Thus, this survey suggests that in a representative stakeholder parliament, the less permissive faction should have 51% of the delegates. This faction can then legislate for the AI to spend 60% of its compute on determining how safe the gain-of-function research is likely to be before deciding whether or not it should assist the scientists.

This example reveals that under plurality- or majority-rule voting, how much compute the biomedical AI should spend on safety testing depends *discontinuously* upon how many stakeholders endorse the less as opposed to the more permissive view (illustrated in figure #2a). This means that the parliament's verdict is extremely sensitive to small differences in stakeholder opinion, such as the difference between 49% and 51% endorsement. Moreover, this hypersensitivity strikes me as entirely unnecessary. In place of the discontinuous function illustrated in figure #2a, we should instead adopt a *continuous* function (two examples of which are illustrated in figure #2b). If the stakeholders are divided roughly 50-50 between the more and less permissive views, then the fraction of the AI's compute devoted to safety testing should plausibly be somewhere *in between* 0% and 60%. I discuss how the exact value of this fraction should be determined in §5.1 below.



(Figure #2a)



(Figure #2b)

Of course, this example is extremely simplified, and perhaps unrealistic. Nonetheless, it illustrates an unattractive feature of the parliamentary approach that might manifest itself even in more complex and more realistic scenarios.

4: Maximising socially expected choiceworthiness

4.1: The basic idea[13]

The MSEC approach to AI alignment is inspired by the *maximise expected choiceworthiness* (MEC) approach to the decision problems faced by *morally uncertain* individual agents. An agent is said to be morally uncertain iff she is at least somewhat uncertain about which moral theory is correct. Many philosophers believe that there is some notion of 'appropriate' action that is sensitive to moral uncertainty. For instance, suppose I have high credence in the view that animal suffering does not matter morally, but also some credence in the view that animal suffering is intrinsically disvaluable. If animal suffering matters morally, then I am required to eat tofu rather than foie gras, whereas if animal suffering does not matter morally, then I am permitted to eat either tofu or foie gras (MacAskill, Bykvist and Ord 2020, p. 15). Many philosophers believe that there is some sense in which eating the tofu is *more appropriate* for me than eating the foie gras.

Philosophers have proposed several possible theories of appropriate action under conditions of moral uncertainty, perhaps the most popular of which is MEC. According to MEC,

> some option A is appropriate under conditions of moral uncertainty iff choosing A maximises expected choiceworthiness (Oddie 1994; Lockhart 2000; Sepielli 2009; 2010; Wedgwood 2013; 2017; MacAskill, Bykvist and Ord 2020).

[13] This subsection partially overlaps with §3 of Lloyd 2024, a companion piece to this paper.

The *choiceworthiness* of some option A according to some moral theory T is a measure of the strength of the decision maker's moral reasons in favour of choosing A according to T (MacAskill, Bykvist and Ord 2020, p. 4). The *expected choiceworthiness* of some action is a weighted average of its choiceworthinesses according to each of the theories in which the decision maker has credence, where each theory's weight is the decision maker's credence in that theory.

Thus, MEC says that we should handle moral uncertainty in the same way as expected utility theory says that we should handle empirical uncertainty. In fact, several advocates of MEC regard this analogy with standard decision theory as a reason to endorse MEC. For instance, MacAskill, Bykvist and Ord (2020, §2.III) claim that since "expected utility theory is the standard way that we should handle empirical uncertainty … maximising expected choiceworthiness should be the standard account of how to handle moral uncertainty."[14]

The MSEC approach to AI alignment is closely related to MEC.[15] One way to motivate MSEC is to conceive of the set of relevant stakeholders as consisting in – or, at least, as being analogous to – some kind of *group agent*.[16] A group agent composed of all relevant stakeholders would have its own credence distributions over the various normative properties of potential AI alignment targets. This group credence distribution would in some way aggregate the stakeholders' own individual credences. For instance, if the overwhelming majority of stakeholders believe that A is more choiceworthy than B, then the stakeholder group agent ought to believe likewise that A is more choiceworthy than B. By contrast, if the stakeholders disagree about whether A is more choiceworthy than B, then the group agent ought to be uncertain about whether A is more choiceworthy than B. According to MSEC,

> some AI ought to be aligned with some alignment target X iff aligning that AI with X would maximise expected choiceworthiness according to the group agent that represents the AI's stakeholders.

Advocates of MSEC are often quite noncommittal on the question of how the stakeholder group's credence distribution should be aggregated from the stakeholders' own individual credence distributions (e.g. Bogosian 2017, §7.2; Ecoffet and Lehman 2021, §3; Thomsen 2022). Martinho et al. (2021, p. 219) suggest that the group's credence in any particular moral theory should be "the share of the population that adheres to that theory." By contrast, Bogosian (2017, p. 601) suggests that we should assign greater weight to the credences of experts in machine ethics relative to the weights assigned to non-expert stakeholders. Progress on questions like this might be advanced by drawing upon the existing philosophical and statistical literature on *group opinion pooling* (Dietrich and List 2016).[17]

4.2: Advantages and disadvantages

[14] Similarly, Christian Tarsney (2021, p. 172) claims that treating normative and empirical uncertainty "differently when we are not forced to is at least prima facie inelegant and undermotivated."
[15] Cf. also Barrett and Schmidt's (2024) application of the MEC idea to the political philosophy of public justification.
[16] On group agents, see e.g. List and Pettit 2011; Tollefsen 2015.
[17] I discuss this connection in more detail in Lloyd 2024, §3.

I discuss the advantages and disadvantages of MSEC in detail in a companion piece to this paper (Lloyd 2024). I now summarise the most important of these advantages and disadvantages.

- One advantage of MSEC is that it avoids the 'tyranny of the majority' problem faced by parliamentary approaches (recall §3.2 above). For instance, in **Jackson** MSEC implies that B is preferrable to A or C.
- Unfortunately, however, MSEC also shares one disadvantage with the parliamentary approach, *viz.* the disadvantage of being hypersensitive to small differences in stakeholder opinion (in ways that seem unnecessarily severe). In **Biorisk**, MSEC implies that how much compute should be spent on safety testing depends discontinuously upon the stakeholder group's credences. If the group's credence in the 'more permissive' view is above a certain threshold, then the biomedical AI should devote none of its compute to safety research. By contrast, if the group's credence in the more permissive view is below that same threshold, then 60% of the AI's compute should be devoted to safety research (Lloyd 2024, §7).
- To make matters worse, MSEC also suffers from several further problems of its own. One example is the problem of 'fanaticism,' which prevents MSEC from selecting the best available compromise options in certain choice situations.[18] For instance, imagine that the stakeholder group has 99.9% credence in the moral theory M_1, and 0.01% credence in the moral theory M_2. Our AI now faces a choice between two options, G and H. The choiceworthiness values according to M_1 and M_2 of the options G and H are given in table #1:

CHOICEWORTHINESS	M_1: 0.999 credence	M_2: 0.001 credence
G	10	10
H	-100	1,000,000

(Table #1)

  In this choice situation, MSEC implies that H is preferrable to G. Yet many of us intuit, to the contrary, that it would be better for our AI to select option G in this choice situation. After all, the stakeholder group is 99.9% sure that H is highly unchoiceworthy, and 100% certain that G is moderately choiceworthy. Under these circumstances, it would be reckless and uncompromising for an AI to prefer H over G.
- Finally, MSEC suffers from a problem of *intertheoretic unit comparisons*. MEC and MSEC are only applicable in cases in which differences between the choiceworthinesses of the options available according to every moral theory in which the agent or group has credence can be measured on some common scale of value. Intertheoretic expected choiceworthiness is simply undefined in cases where unit comparisons are impossible.

  Many philosophers of moral uncertainty have been sceptical about the possibility of intertheoretic unit comparisons. For instance, imagine trying to compare absolutist deontology against scalar utilitarianism. These two different moral theories don't even use

[18] 'Fanatical' decision theories prefer lotteries with tiny probabilities of sufficiently high payoffs over guarantees of modest payoffs.

> the same deontic categories: absolutist deontology sees the world only in terms of permissions and prohibitions, whereas scalar utilitarianism sees the world only in terms of betterness and worseness of outcomes in terms of aggregate well-being. It strikes many of us as implausible to say that there is some value of *k* such that the choiceworthiness difference between murdering someone and refraining from doing so according to absolutist deontology is *k* times the size of the choiceworthiness difference between buying a Mother's Day present and refraining from doing so according to scalar utilitarianism. Unfortunately, however, MEC and MSEC rely upon these kind of intertheoretic unit comparisons.[19]

In summary: MSEC shares some of the problems of the parliamentary approach to AI alignment, and also suffers from a couple of new problems of its own.

Despite MSEC's popularity, it is not particularly surprising that it faces these problems as a procedure for decision making under conditions of moral disagreement. MEC and MSEC are inspired by standard decision theory, which treats decision making under uncertainty as a problem of selecting the *gamble* that has the highest possible payoff in expectation. This framework is neither designed nor intended to select *compromise* options around which one can build some measure of social consensus. In the remainder of this paper, I introduce a bargaining-theoretic approach which I suggest is better suited to the challenges of AI alignment under moral disagreement.

My presentation of the bargaining approach to AI alignment will necessarily be in broad strokes. I focus on the general contours and prospects of the approach, since I lack the space to develop all of the details in this paper. My aim is only to argue that the bargaining-based approach is a more attractive prospect for development than the voting and MSEC approaches.

5: Bargaining

5.1: Resource division

According to the bargaining-theoretic approach, we should conceptualise the problem of AI alignment in the face of moral disagreement as a kind of bargaining problem, in which the stakeholders – or their simulated surrogates – may negotiate with each other over what the alignment target should be.[20]

---

[19] For further discussion – including on 'statistical normalization' responses to the problem of intertheoretic comparisons – see Lloyd 2024, §§8-9.

[20] For two analogous approaches to the problem of intrapersonal moral uncertainty, see Greaves and Cotton-Barratt 2023; Kaczmarek, Lloyd and Plant forthcoming.

Another approach to moral uncertainty that incorporates the idea of bargaining is Newberry and Ord's (2021) 'moral parliaments' approach (inspired by Bostrom 2009). Newberry and Ord (2021, p. 1) propose that "the appropriate choice under moral uncertainty is the one that would be reached by a parliament comprised of delegates representing the interests of each moral theory, who number in proportion to [the decision maker's] credence in that theory." Furthermore, there should be "a period during

In order to model the alignment problem bargaining-theoretically, we must designate some option in each choice situation as the 'disagreement point' – the option that the AI will choose if the stakeholders cannot together agree upon some alternative option. Our rule for selecting this disagreement point should be as fair as possible to each of the AI's stakeholders.

First of all, consider **Biorisk**. In this scenario, the stakeholders disagree about how the AI should utilise some continuously divisible resource (in this case, its compute). In scenarios of this kind, one outcome in particular strikes me as procedurally fair to all of the stakeholders, *viz.* an outcome in which the AI's resource (in this case, its compute) is divided by the total number of stakeholders and each stakeholder is then entitled to decide how 'her' share of the AI's resource should be used.[21] For instance, if the stakeholders are split 50-50 between the more and less permissive views in **Biorisk**, then the AI's compute should be split 50-50 between safety and gain of function research in the disagreement outcome.

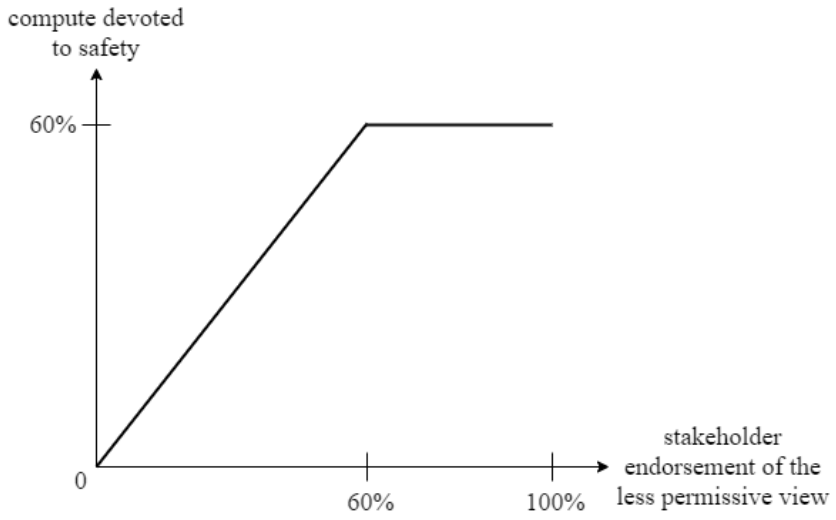
What if the stakeholders are split 20% to 80% between the more and less permissive views? Recall that according to the less permissive view, 60% of the AI's compute should be devoted to safety research. By contrast, according to the more permissive view, none of the AI's compute should be devoted to safety research. Under the disagreement outcome, the 20% of stakeholders who endorse the more permissive view will assign all of their compute to gain of function research. Given these circumstances, the 80% of stakeholders who endorse the less permissive view will be able to realize what they regard as the best of all possible allocations of compute in this choice situation (60% on safety, and 40% on gain-of-function) by assigning the majority (75%) of their allotted compute to safety, with the remainder of their allotted compute being assigned to gain of function research.

More generally, figure #3 illustrates the percentage of compute that should be assigned to safety testing in the disagreement outcome as a function of the percentage of stakeholders who endorse the less permissive view.

which delegates may negotiate: this could include trading votes on one motion for votes on another, introducing novel options for consideration within a given motion, or forming deals with others to vote for a compromise option that both consider to be acceptable" (Newberry and Ord 2021, p. 8).

Newberry and Ord's approach to moral uncertainty incorporates elements of both voting- and bargaining-based approaches. However, unlike Greaves and Cotton-Barratt (2023) and Kaczmarek, Lloyd and Plant (forthcoming), Newberry and Ord do not discuss what formal model of the parliamentary bargaining process would be most appropriate. This is an important lacuna, which leaves seriously underdetermined the implications of Newberry and Ord's approach to moral uncertainty.

[21] I assume, for sake of simplicity, that each stakeholder's claim to influence the AI's alignment target is of the same *pro tanto* strength. If some stakeholders have stronger *pro tanto* claims than their peers, then these stakeholders should have control over a larger share of the AI's resources. I hope to discuss this complication in more detail in future research.



(Figure #3)

Under any possible distribution of stakeholder opinion, more-permissive stakeholders will assign all of their allotted compute to gain of function research. On the other hand, the less-permissive stakeholders will assign all of their allotted compute to safety research only if they number less than or equal to 60% of the total stakeholders; if they number more than 60%, then the less-permissive stakeholders will split their allotted compute between safety and gain of function research, in order to realize what they regard as the best of all possible allocations of compute in **Biorisk**.

Figure #3 also illustrates how much compute the bargaining-theoretic approach to alignment implies *should* in fact be devoted to safety research, as a function of stakeholder endorsement of the less permissive view (assuming my preferred 'equal division' disagreement point). This is because the stakeholders in **Biorisk** have nothing to gain from bargaining with each other as opposed to simply using their share of the AI's compute in the way that they would most prefer. Less-permissive stakeholders think that the share of the AI's compute devoted to safety testing should be as close as possible to 60%. By contrast, more-permissive stakeholders think it should be as close as possible to 0%. Given that the disagreement value of this percentage will always be between 0 and 60 (see figure #3), there are never any 'gains from trade' to be realised by bargaining between these two groups of stakeholders.

Figure #3 strikes me as a desirable resolution of **Biorisk**. According to the bargaining-theoretic approach, how much compute the biomedical AI should spend on safety testing is a *continuous* function of how many stakeholders endorse the more as opposed to the less permissive view. Thus, the bargaining-theoretic approach's verdicts are not particularly sensitive to small differences in stakeholder opinion – unlike the verdicts of the parliamentary and MSEC approaches to alignment. This is one attractive feature of the bargaining-theoretic approach.

In this discussion of **Biorisk**, I have assumed that the 'choice situation' over which stakeholders are bargaining is fairly narrowly individuated, in the sense that this choice situation ends once the stakeholders decide how the AI should behave in **Biorisk**. However, I make this

assumption only for simplicity of exposition. In reality, many AIs will be able to forecast the moral choices they are likely to face much further into the future. Under these circumstances, the stakeholders should be allowed to make longer-term bargains with each other.[22] For instance, one group of stakeholders could trade away their endowments of compute in **Biorisk**, in return for receiving some other stakeholders' endowments of compute at some time in the future. Of course, there is a trade-off to be settled here, because broadening the time horizon requires us to devote more compute to running the bargaining simulation.[23]

5.2: Discrete-choice situations

The stakeholders in **Biorisk** disagree about how an AI should allocate a certain continuously divisible resource (in this case, its compute). However, moral disagreements between an AI's stakeholders can also occur in cases where the AI faces a choice between several discrete options. For instance, in **Jackson** our AI faces a choice between three options, A, B, and C. What should the disagreement outcome be in a discrete-choice scenario like **Jackson**?

It would be easiest to specify a fair disagreement point if we knew that our AI were about to face an arbitrarily large (but finite) sequence of identical instances of the **Jackson** choice situation. In that case, we could say that 'the right to decide what the AI does in one instance of the **Jackson** choice situation' is like one indivisible unit of a certain resource; and the disagreement outcome could then be one in which these decision rights are divided up equally between all of the stakeholders. For instance, if our AI had only 100 stakeholders, and were about to face a sequence of 1,000 instances of the **Jackson** choice situation, then in our disagreement outcome the first stakeholder would decide how the AI should behave in the first 10 of these choice situations; and the second would decide for the next 10 situations; and so on.

In **Biorisk**, the AI's stakeholders had nothing to gain from bargaining to reach some alternative to the disagreement point; there were no possible 'gains from trade' to be realised through bargaining. By contrast, in a sequence of 1,000 instances of **Jackson**, it is safe to assume that significant gains from trade stand to be realised through bargaining.

Recall that according to the 51% of stakeholders who endorse the moral theory T_1 in **Jackson**, A is the best option available, B is almost as good as A, and C is terrible. By contrast, according to the 49% of stakeholders who endorse the moral theory T_2, C is the best option available, B is almost as good as C, and A is terrible. Under these conditions, it is safe to assume that both sets of stakeholders will prefer an outcome in which the AI chooses B all or almost all of the time over the disagreement outcome in which the AI chooses A 51% of the time, and C 49%

[22] Otherwise, the bargaining approach will be insensitive to differences in relative stakes across sequences of choices (Greaves and Cotton-Barratt 2023, §10), and subject to diachronic inconsistencies. I thank an anonymous reviewer for pressing me to clarify this point.

[23] What is the best way of resolving this trade-off? Although I won't try to settle that question here, I will tentatively suggest that this second-order question could itself be resolving by some process of bargaining between the stakeholders (antecedent to, and setting the terms for, any subsequent bargaining concerning first-order ethical problems).

of the time. Thus, the bargaining-theoretic approach to machine ethics implies that choosing B all or almost all of the time is preferrable to the disagreement outcome.[24]

What about in cases where (at least: for all we know) our AI is about to face only one instance of a discrete-choice situation like **Jackson**? One possible option here is to stipulate that the disagreement outcome should be a *lottery* in which each stakeholder has an equal chance to decide how the AI should behave (Newberry and Ord 2021; Greaves and Cotton-Barratt 2023, §4.2). For instance, the disagreement point in **Jackson** would involve performing option A with 51% probability, and option C with 49% probability. This lottery approach in effect converts a single, indivisible 'right to decide what the AI does in **Jackson**' into a continuously divisible resource, *viz.* the *chance* to decide what the AI does in **Jackson**.

This lottery proposal is likely to yield an attractive result in the one-off **Jackson** case. It is safe to assume that every stakeholder will prefer an outcome in which the AI chooses B with high probability over the disagreement outcome in which the AI chooses A with 51% probability, and C with 49% probability. Thus, this proposal implies that choosing B with high probability is preferrable to the disagreement outcome.

Unfortunately, the lottery proposal also suffers from a couple of significant defects. Firstly, although there is an *ex ante* sense in which the lottery is fair to all of the stakeholders, it also has the potential to create outcomes that strike me as highly unfair *ex post*. For instance, consider the following discrete-choice situation:

> **Binary:** our AI faces a choice between two options, D and E. 99% of the AI's stakeholders think that D is better than E, whereas 1% think that E is better than D.

The random lottery disagreement outcome involves performing D with 99% probability, and E with 1% probability. Under these conditions, there are no possible 'gains from trade' to be realised through bargaining. Hence, our bargaining-theoretic approach implies that our AI should randomise 99-1 over D and E. On the 1% chance that E is randomly selected, our bargaining approach implies that our AI should choose an option that 99% of stakeholders think is the worst possible option available in this choice situation. This result strikes me as highly unattractive (likewise Newberry and Ord 2021).

Secondly, the lottery proposal has the potential to unfairly disadvantage risk-averse stakeholders. If the stakeholders are trying to bargain towards a resolution in which some determinate outcome is chosen with certainty, but the disagreement point is a risky lottery, then risk-averse stakeholders might be more willing than risk-neutral stakeholders to make concessions, giving up more of what they want in the negotiations in order to avoid defaulting to the risky

[24] Exactly what percentage of the time the AI should choose B will depend on which formal bargaining solution we adopt (see §5.3 below).

disagreement lottery.[25] However, I can see no reason why a stakeholder's degree of influence over our AI should depend in this way on her moral theory's degree of risk aversion.

Fortunately, a modified version of the lottery proposal can avoid these two problems. As in the original lottery proposal, the disagreement outcome will be a lottery in which each stakeholder has an equal chance to decide how the AI should behave. And, as before, the stakeholders can bargain and make contracts with each other before the lottery winner is selected. (For instance, in cases like **Jackson** the stakeholders could negotiate a contract under which everyone or almost everyone would choose option B if they won the decision lottery.) In addition, however, we now stipulate that if some stakeholder wins the lottery in any discrete-choice situation X, then although she is still permitted to choose whichever option O she wishes, her choice is subject to the requirement that the stakeholders who favoured option O will be required to *compensate* the stakeholders who opposed O. This compensation will come in the form of promises to resolve future choice situations in ways favoured by the compensated stakeholders.[26]

What would be a fair scheme of compensation between the stakeholders after a lottery winner chooses option O? Recall that there is a highly attractive disagreement point in cases where our AI is about to face an arbitrarily large sequence of identical instances of some discrete-choice situation X, *viz.* the decision rights for these instances of X being divided up equally between the stakeholders. I suggest that we can use each stakeholder's level of satisfaction with this disagreement point as a baseline to determine how much compensation should flow to the stakeholders who opposed option O. Where possible, we should choose a compensation scheme such that, as N tends towards infinity, each stakeholder tends towards being indifferent between

> (1) a world in which X is repeated N times over, and where in each of these N instances of X, O is chosen and compensation flows between the stakeholders in accordance with our scheme

and

> (2) a world in which X is repeated N times over, and where the decision rights for these N instances of X are initially divided equally between all of the stakeholders. (As always, after this initial distribution of decision rights, the stakeholders will then bargain with each other to determine how the AI should behave.)

To illustrate my new compensation proposal, let's now reconsider the **Binary** choice situation. First of all, suppose that the modified lottery is won by a stakeholder who chooses D

[25] In the words of Volij and Winter (2002), "that increasing risk aversion reduces a player's share in the bargaining outcome and increases that of his opponent" is "one of the results most frequently quoted in the bargaining literature," and appears in many "different variations including both the cooperative and the non-cooperative frameworks."

[26] In other words, the default 'currency' or 'medium' of compensation will be, in effect, 'probability of controlling future choice situations.' For practical purposes, it is safe to assume that compensation in this currency will always or nearly always be possible, even though in principle this might be impossible, for instance if we knew for certain that we were confronting the last choice situation that humanity (or its AI agents) will ever face. I thank an anonymous reviewer for pressing me to clarify this point.

instead of E. We wish to find a compensation scheme under which each of the stakeholders is indifferent between

> (3) a world in which **Binary** is repeated N times over, and in each of these N instances of **Binary**, D is chosen and compensation flows between the stakeholders in accordance with our scheme

and

> (4) a world in which **Binary** is repeated N times over, and where the decision rights for these N instances of **Binary** are initially divided equally between all of the stakeholders. (As always, after this initial distribution of decision rights, the stakeholders will then bargain with each other to determine how the AI should behave.)

In (4), D would be chosen in 99% of the N instances of **Binary**, whereas E would be chosen in 1% of these instances.[27]

Any compensation scheme that satisfies this new fairness criterion would involve each of the stakeholders who think that D is better than E paying a small amount of compensation to the 1% of stakeholders who think that E is better than D.[28] If no compensation were to change hands, then the D-favouring stakeholders would marginally prefer (3) over (4) (all else being equal, D being chosen 100% of the time is marginally better than D being chosen 99% of the time), and the E-favouring stakeholders would marginally prefer (4) over (3) (all else being equal, E being chosen 1% of the time is marginally better than E being chosen 0% of the time). However, in this case the amount of compensation owed by each D-favouring stakeholder will be rather small – partly because each E-favouring stakeholder will receive compensation from 99 D-favouring stakeholders, and partly because there is not a dramatic difference between E being chosen 1% versus 0% of the time.

On the other hand, what if the modified lottery is won by a stakeholder who chooses E instead of D? In this case, we wish to find a compensation scheme under which each of the stakeholders is indifferent between (4) and

[27] Recall that in **Binary** as well as in **Biorisk**, there are no possible gains from trade to be realised through bargaining, and so each stakeholder will just use her initial endowment of decision rights in the manner that she herself most prefers. By contrast, however, recall that in a counterfactual world in which a scenario like **Jackson** is repeated N times over – with the N decision rights being initially divided between all of the stakeholders – significant gains from trade could be realised through a post-endowment bargain under which the stakeholders agree to choose option B in most or all of the N **Jackson** cases. This example illustrates the important role that the bargaining mechanism can play even in my modified lottery approach. (I am grateful to one of the guest editors of this special issue for pressing me to clarify this point.)

[28] More precisely, stakeholders should pay compensation *to the extent that* they think D is better than E, and should receive compensation *to the extent that* they think E is better than D. (I thank an anonymous reviewer for pressing me to clarify this point.)

(5) a world in which **Binary** is repeated N times over, and in each of these N instances of **Binary**, E is chosen and compensation flows between the stakeholders in accordance with our scheme.

Any compensation scheme that satisfies this criterion would involve each of the stakeholders who think that E is better than D paying a large amount of compensation to the 99% of stakeholders who think that D is better than E.[29] If no compensation were to change hands, then the D-favouring stakeholders would greatly prefer (4) over (5) (all else being equal, D being chosen 99% of the time is much better than D being chosen 0% of the time), and the E-favouring stakeholders would greatly prefer (5) over (4) (all else being equal, E being chosen 100% of the time is much better than E being chosen 1% of the time). The compensation owed by each of the E-favouring stakeholders will be very large – partly because the D-favouring stakeholders outnumber the E-favourers 99-to-1, and partly because there is a dramatic difference between D being chosen 99% versus 0% of the time.

Hence, if one of the E-favouring stakeholders wins the modified lottery in **Binary**, although she is still permitted to choose option E if she wishes, she may do so only at the cost of committing the E-favouring stakeholders to paying a large amount of compensation to the D-favouring stakeholders. If the E-favourers get their way in **Binary**, then the D-favourers will have to get their way with certainty in a large number of future choice situations. This strikes me as an attractive resolution of discrete-choice situations like **Binary** in which no good compromise options are available.[30]

Before I move on, I want to address one complication concerning my compensation proposal.[31] (Readers interested only in following the main line of argument may wish to skip over this last part of the current subsection.)

My new compensation proposal always asks us to imagine a counterfactual world in which our AI faces a sequence of N identical instances of some discrete choice situation X. By 'identical,' I mean here: identical in every respect that any of our stakeholders regard as at least potentially normatively relevant to how our AI should behave.

However, for at least some choice situations paired with at least certain collections of stakeholders, it might be impossible to imagine this sequence of N repetitions of X, identical in all normatively salient respects. For example, imagine that our AI has to decide whether or not to make the duck-billed platypus permanently extinct. Clearly, this is not a choice situation that one could imagine repeating multiple times over, since no species can be made permanently extinct multiple times over!

[29] More precisely, stakeholders should pay compensation *to the extent that* they think E is better than D, and should receive compensation *to the extent that* they think D is better than E. (Once again, I thank an anonymous reviewer for pressing me to clarify this point.)

[30] Although I have motivated this modified lottery procedure using a **Binary** discrete choice, I also intend for the procedure to be applied in discrete-choice situations like **Jackson** where there are more than two options. I thank an anonymous referee for pressing me to clarify this point.

[31] I thank an anonymous reviewer for pressing me to discuss this complication.

In fact, we could imagine confronting this problem even in some rather more mundane choice situations. For instance, imagine that our AI has to decide whether or not to kill someone, having never killed anyone before. Furthermore, imagine that some of the AI's stakeholders think (for whatever reason) that it killing someone after having already killed another person before would be even worse than killing someone having never killed anyone before. Under this assumption, the first and second situations in a sequence of choices between killing and not killing would *not* be identical to each other in all normatively relevant respects according to the stakeholders who think that killing for a second time would be morally worse than the AI's first act of killing.

Fortunately, I think there is simple fix to my compensation proposal that will allow me to avoid this problem. This fix involves conceptualising 'choice situations' in far more abstract terms than one might at first have expected. For instance, imagine that our AI faces a choice between killing and not killing an enemy combatant. Some of the AI's stakeholders are absolutist deontologists, who think that killing in these circumstances would be 'IMPERMISSIBLE,' whereas not killing would be 'REQUIRED.' By contrast, some other stakeholders are utilitarians, who think that killing would have 'UTILITY +10,' whereas not killing would have 'UTILITY -5.'

I want to suggest that only some of the facts about this scenario should be understood as essential to the '*choice situation*' X that our AI confronts here. More specifically, I suggest that the following description captures everything that we should regard as essential to the choice situation X:

(i) There are two options available.
(ii) The deontological stakeholders think that the first option is IMPERMISSIBLE, but the second option is REQUIRED.
(iii) The utilitarian stakeholders think that the first option has UTILITY +10, but the second option has UTILITY -5.

Thus, to imagine a world in which the *choice situation X* is repeated N times over would just be to imagine a world in which our AI confronts N choices between two options, the first of which is IMPERMISSIBLE for the deontologists but has UTILITY +10, and the second of which is REQUIRED for the deontologists but has UTILITY -5. In other words, there's no need to get bogged down in actually imagining any details about whether the first action is an act of 'killing,' or anything like that. Rather, all we need to ask ourselves are questions like: how would our stakeholders compare

> (1*) a world in which a deontologically IMPERMISSIBLE but UTILITY +10 option is chosen N times over, with compensation flowing between the stakeholders

against

> (2*) a world in which a deontologically IMPERMISSIBLE but UTILITY +10 option is chosen N/2 times over, and yet a deontologically REQUIRED yet UTILITY -5 option is also chosen N/2 times over.

Thus, my compensation proposal need not be threatened by any concerns about 'unrepeatable' choice situations.

5.3: Bargaining solutions

All of the choice situations that I have discussed in §§5.1-5.2 have been cases in which the rough outcomes of the stakeholder bargaining process are intuitively uncontroversial. However, in other more complicated choice situations it will be much less obvious which outcomes will be agreed to by reasonable stakeholder bargainers. In order to handle scenarios of this kind, we will need to specify a determinate 'solution procedure' for calculating which contracts our simulated bargainers will agree to.

One potential response to this problem would be to train our AI to predict how human stakeholders would actually bargain with each other to resolve ethically-charged disputes. We can then stipulate that the outcome of any simulated alignment bargaining problem should be given by the AI's prediction of how its stakeholders would actually resolve that bargaining situation. Call this the 'simulated folk bargaining' solution procedure.

In my view, the simulated folk bargaining procedure suffers from several important disadvantages. One such disadvantage is that unless and until significant progress is made on the problem of 'interpretability,' it will be difficult to understand why the AI bargaining simulator selects particular outcomes.[32] The machine learning models that have been at the forefront of recent advances in AI capabilities essentially function as 'black boxes;' even the developers of these new models have little insight into the particular procedures for producing outputs that they have learnt from their training data. Several studies suggest that this feature of machine learning models reduces public trust in their outputs (Ashoori and Weisz 2019; Lai and Tan 2019; Zerilli et al. 2022; see also von Eschenbach 2021). Thus, the simulated folk bargaining procedure threatens to undermine public trust in the bargaining-theoretic AI alignment project.

Another potential disadvantage of the simulated folk bargaining approach is that real-world cases of bargaining over morally contentious issues might have undesirable features that we should not wish to emulate in our bargaining solution. For instance, a subgroup of stakeholders who strongly dislike some of the other stakeholders might be motivated by spite or schadenfreude to thwart the bargaining objectives of those other stakeholders. Furthermore, even good-faith stakeholders might struggle to deal with all of the complexities of particularly complicated bargaining situations that involve multiple moral viewpoints and multiple potential options faced by some AI. In these kinds of complicated choice situations, real-world stakeholders might resort to imperfect, 'quick and dirty' heuristics to inform their decision making in the bargaining process. Clearly, our alignment procedure should not attempt to emulate these imperfect heuristics.

[32] Regarding interpretability, see Mittelstadt forthcoming.

In place of the simulated folk bargaining approach to the solution procedure, I suggest that we should adopt a precisely-specified 'normative solution concept,' of the kind that has been developed in the economic literature on bargaining.[33]

The most well-known such solution concept is the *Nash bargaining solution* (NBS). The NBS is applicable in cases where all of the bargainers can be modelled as expected utility maximisers. John Nash's groundbreaking treatment of this class of bargaining problems lays out four plausible axioms on the outcomes of good-faith (referred to as 'cooperative') bargaining procedures (Nash 1950):

1. *Scale invariance:* any positive affine rescaling of any bargainers' utility functions should not alter the bargaining solution.
2. *Pareto optimality:* no feasible alternatives should Pareto dominate the bargaining solution.
3. *Symmetry:* if every bargainer has the same utility function and disagreement utility, then every bargainer should have the same utility in the bargaining solution.
4. *Independence of irrelevant alternatives:* eliminating an element from the set of feasible outcomes should only make a difference to the bargaining solution if the eliminated outcome would itself have been selected as the bargaining solution had it not been eliminated.

The NBS uniquely satisfies all four of these axioms (Nash 1950).[34] Before stating the NBS, I first introduce the necessary notation. Firstly, let $\{1,\ldots,n\}$ denote the set of bargainers. For each bargainer $i \in \{1, \ldots, n\}$, $u_i(a)$ denotes i's utility under the possible outcome or lottery over outcomes a. And let d denote the disagreement outcome or lottery. Some outcome or lottery over outcomes A is an NBS of this situation iff (1) $u_i(\mathrm{A}) \geq u_i(d)$ for every bargainer $i \in \{1, \ldots, n\}$, and (2) setting a=A maximises the Nash maximand

$$\prod_{i=1}^{n} \big(u_i(a) - u_i(d)\big)$$

One attractive feature of the NBS is that (all else being equal) it favours equal division of gains from trade between the bargainers. For example, suppose that two bargainers are choosing between (i) an option A that gives each bargainer a utility gain of 4 over the disagreement point, and (ii) another option B that gives the two bargainers utility gains over the disagreement point of 2 and 6 respectively. Under option A, the value of the Nash maximand is $4 \times 4 = 16$, whereas under option B the value of the Nash maximand is $2 \times 6 = 12$. Hence, as desired, the NBS prefers option A over option B.

One set of potential problems with the NBS lies in its assumption that all of the bargainers can be modelled as expected utility maximisers.[35] Admittedly, the orthodox view amongst decision theorists is that expected utility maximisation is a requirement of practical rationality. Several

[33] For a useful overview, see Thomson 1994.
[34] The NBS can also be justified as the limiting outcome of several diachronic 'non-cooperative' models of the bargaining process (Binmore, Rubinstein and Wolinsky 1986).
[35] One might also worry that the NBS inappropriately disadvantages risk-averse bargainers (Lloyd 2022).

important representation theorems have demonstrated that any decision maker whose preferences satisfy certain attractive axioms can be modelled as maximising expected utility; and one can also argue in favour of expected utility maximisation using the Law of Large Numbers (Steele and Stefánsson 2020; Briggs 2023).

Especially in recent years, however, several decision theorists have proposed alternatives to expected utility maximisation (see, *inter alia*, Buchak 2013; 2022; Tenenbaum 2017).[36] If one or more of our AI's stakeholders endorse one of these rivals to expected utility maximisation, then this creates a potential problem for a Nash bargaining approach to AI alignment. This is because the NBS is only applicable in cases where all of the bargainers can be modelled as expected utility maximisers. (I explain the reasons for this restriction in a slightly technical footnote.)[37]

One potential response to this problem would be to adopt expected utility maximising representations that *best approximate* the preferences of our AI's stakeholders. Of course, these representations will sometimes deviate from the preferences of stakeholders who cannot be modelled perfectly as expected utility maximisers. Nonetheless, it might be possible to find approximations for which these deviations are usually small. It is difficult to predict *a priori* how accurate these approximations could be for the kinds of moral views about AI behaviour held by real-world stakeholders. This is in large part an empirical question, which could be explored in future research.

Another potential response to this problem is to shift from the NBS to an 'ordinal' bargaining solution that is applicable even in cases where some of the bargainers cannot be represented as expected utility maximisers. Although the literature on ordinal bargaining solutions is significantly less developed than the literature on 'cardinal' solutions like the NBS, there are nonetheless some promising proposals. In an appendix to this paper, I develop an ordinal bargaining solution that builds on several of the proposals in this literature.

5.4: Practicalities

The complicated bargaining proposal that I have developed in §§5.2-5.3 would need to be implemented by a *simulated* version of the bargaining approach to AI alignment. If we had unlimited resources and compute at our disposal, then our AI could learn models of the alignment preferences of each and every one of its stakeholders – and could then simulate bargaining between them in any moral choice situation. However, under real-world limitations on resources and compute, we will likely have to adopt a less ambitious procedure. First, it should be sufficient for

[36] See also the literature on the 'precautionary principle' (Rechnitzer 2020).

[37] The NBS is only 'scale invariant' up to positive *affine* transformations of the bargainers' utility functions. Hence, for the NBS to be applicable to some bargaining problem, (1) the class of affine transformations of the chosen functional representations of the bargainers' utilities must be somehow privileged over any other possible class of representations of those orderings. In other words: the bargainers' utilities must be *cardinal*. Furthermore, (2) the set of feasible utility vectors must be convex in order for the NBS to satisfy its characteristic axioms. In many cases where one or more of the bargainers are not expected utility maximisers, conditions (1) or (2) are violated. Restricting the applicability of the NBS to cases in which all of the bargainers can be represented as expected utility maximisers is the generally accepted way to ensure that conditions (1) and (2) will both be satisfied.

our AI to learn the preferences of some representative subsample of the stakeholder population. Second, our AI is also likely to be able to group the preference orderings of the stakeholders within this subsample into a manageable number of latent classes. The AI can then compute the solution for bargaining between the representative members of these latent classes, rather than between each and every one of the stakeholders. These two simplifications should significantly reduce the computational demands of simulated bargaining.

Several studies already demonstrate the feasibility of learning models of stakeholders' alignment preferences based on their judgements in a small number of cases (Kim et al. 2018; Noothigattu et al. 2018; Lee et al. 2019; Freedman et al. 2020; Martinho et al. 2021). And a few of these studies also demonstrate the feasibility of grouping stakeholders' preference orderings into a manageable number of latent classes (Awad et al. 2018; Martinho et al. 2021). Furthermore, Lee et al.'s 2019 study provides some encouraging evidence that their "framework enabled participants to build models that they felt confident represented their own beliefs." Stakeholders' trust in AI models of their moral views might improve even further if proposals (such as those of Sinnott-Armstrong, Skorburg, Giublini and Savulescu) for personalised AI moral advisors tailored to their advisee's own values achieve widespread uptake in the near future (Giublini and Savulescu 2018; Sinnott-Amstrong and Skorburg 2021). Developing this form of trust is likely to be an important precondition for the adoption of a simulated version of the bargaining-theoretic approach to alignment.

6: Conclusion

I have now discussed three potential responses to the Normative Problem of AI alignment under conditions of stakeholder disagreement. I have argued (in §§3-4) that the voting- and decision-theoretic responses to this problem both suffer from several important disadvantages. By contrast, a simulated version of the bargaining-theoretic approach need not suffer from any of these disadvantages. Unlike the voting- and decision-theoretic approaches, the bargaining approach is specifically designed to select attractive *compromise* options around which one can build some measure of social consensus. Although it has not yet been discussed or defended in the alignment literature, I suggest that the bargaining approach is superior to its voting- and decision-theoretic alternatives.

One important question that I have not discussed in this paper is that of who should count as a 'stakeholder' for any particular AI system (see Martin 2017; Baum 2020).[38] This is a particularly important topic for future alignment research, since it has the potential to significantly affect the verdicts of all three of the voting-, decision-, and bargaining-theoretic approaches to alignment. Future research might investigate, for instance:

1. Whether it can be legitimate to exclude some stakeholders from the decision-making process in cases where almost all of the other stakeholders agree that their views about alignment are completely 'unreasonable.'

[38] For business ethics responses to the 'stakeholder identification' problem, see Kaler 2002; Miles 2017.

2. Whether and how we should attempt to represent the preferences of future generations in present-day alignment decision making-processes (cf. Karnein 2016; Zwarthoed 2018).

These kinds of questions strike me as highly productive topics for future research.

Appendix

In this appendix I present an ordinal bargaining solution that builds on several existing proposals in the literature: the *Position Maximin Solution* (PMS).[39] Although I suggest the PMS as a solution concept in terms of which the bargaining approach to AI alignment could be developed, I do not necessarily wish to claim that PMS is the best solution concept for this purpose.[40] All I suggest is that it is 'proof of concept' for the idea of using an ordinal bargaining solution.

I begin by introducing some terminology:

- An agent *weakly prefers* X over Y iff she thinks that X is at least as good as Y.
- An agent *strictly prefers* X over Y iff she thinks that X is better than Y.
- An outcome X is *Pareto optimal* iff there does not exist any alternative outcome Y that every agent weakly prefers, and at least one agent strictly prefers, to X.
- X is *individually rational* iff every agent weakly prefers X to the disagreement point.
- X is an *imputation* iff X is both Pareto optimal and individually rational.

In the remainder of this subsection, we will be concerned exclusively with imputations; outcomes that are Pareto suboptimal or are individually irrational are unattractive as potential bargaining solutions. In the context of PMS, 'outcomes' (and hence 'imputations') should be construed as ways of using or planning to use the resources or lottery tickets that our stakeholder-bargainers are endowed with.[41]

We may define an agent's *position-measured satisfaction* with some imputation X as the fraction of imputations that the agent weakly prefers X over.[42] Thus, an agent's position-measured satisfaction with some imputation X is an ordinal measure of the desirability to that agent of X

[39] PMS generalises the *Imputational Compromise* solution discussed by Kıbrıs and Sertel (2007) and Conley and Wilkie (2012). It is also closely related to Sprumont's *Rawlsian Arbitration* (1993), Hurwicz and Sertel's *Kant-Rawls Compromise* (1999), Brams and Kilgour's *Fallback Bargaining* (2001), and Congar and Merlin's *Maximin Rule* (2012). For an alternative approach to bargaining problems without vNM preferences, see Nicolò and Perea 2005.

[40] For instance, many will regard Position *Leximin* as superior to PMS. But since PMS is already complicated enough, I won't complicate things any further by presenting Position Leximin in this paper.

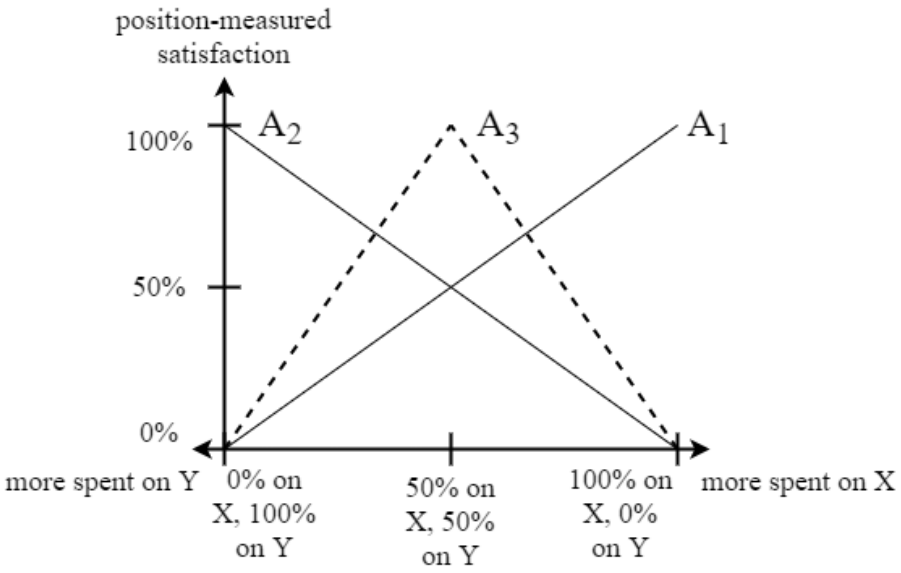
[41] Hence, in this context the 'outcomes' can include *lotteries* over determinate choices, as well as the determinate choices themselves. (Thanks to an anonymous reviewer for pressing me to clarify this point.)

[42] This definition assumes that there is some privileged measure over the space of imputations. In the context of moral disagreement about AI alignment, this requires us to assume that there are privileged interval-scale measures of any resources that our AI might be endowed with, unique up to positive affine transformation. For instance, I assume that measuring computer memory by bytes, megabytes, or gigabytes is privileged over measuring it by bytes squared or log bytes. And in discrete-choice situations, I assume that measuring the chance of some option being chosen in a decision lottery by probability of selection is privileged over measuring this chance by probability squared or log probability.

(relative to the other available imputations). An agent is *positionally worst-off* under some imputation X iff her position-measured satisfaction with X is no greater than any other agent's. We may now define PMS: some outcome X is a PMS to some bargaining problem B iff X is an imputation of B, and the position-measured satisfaction of the positionally-worst off agent or agents in B under X is no lower than the position-measured satisfaction of the positionally-worst off agent or agent in B under any other imputation.

It might be helpful to illustrate PMS using a simple scenario with continuous space of imputations.[43] For example, suppose that there are three agents, A_1, A_2, and A_3, and a single resource that can be divided in any proportion between two possible uses, X and Y, to form the space of imputations. The three agents' preferences over this space of imputations are summarised as follows, and illustrated more completely in figure #4:

- A_1 prefers for as much as possible to be spent on X, and as little as possible to be spent on Y.
- A_2 prefers for as much as possible to be spent on Y, and as little as possible to be spent on X.
- A_3 prefers a split between X and Y is as close as possible to 50-50.



(Figure #4)

Clearly, a 50-50 division between X and Y is the point on figure #4 that maximises the position-measured satisfaction of the positionally worst-off agent. Under the 50-50 division, A_1 and A_3 are the joint positionally worst-off agents, with both of them having a position-measured satisfaction of 50%. But under any division that allocates less than 50% to X, A_1 is the positionally worst-off agent, with a position-measured satisfaction of less than 50%. And under any division that allocates more than 50% to X, A_3 is the positionally worst-off agent, with a position-measured

[43] Cases this simple are unlikely to occur in the AI alignment context. I use this simplified example only for ease of exposition.

satisfaction of less than 50%. Hence, a 50-50 division between X and Y is the unique PMS to this bargaining problem.

Unlike the NBS, the Position Maximin Solution is defined over the set of possible outcomes, rather than the set of possible utility vectors. That is why the PMS can handle agents who cannot be modelled as expected utility maximisers (Sakovics 2004). Having introduced the notion of position-measured satisfaction, we could in principle use it to define other potential ordinal bargaining solutions. I focus on the PMS in this paper in large part because it belongs to a family of ordinal solution concepts that have already been studied by bargaining theorists.

PMS can be understood as the result of an attractive 'fallback' cooperative procedure for finding a compromise in any bargaining problem. At the start of this procedure, each agent 'reports' the set of imputations S_0 that she weakly prefers to every other imputation – in other words: her (joint-)favourite imputation(s). *n*% of the way through the time allotted for this fallback procedure, each agent reports the set S_n of imputations that she weakly prefers to at least (100-*n*)% of all possible imputations. The procedure halts as soon as one or more imputations are being reported by all of the agents. That imputation or set of imputations is the PMS to the bargaining problem.

ACKNOWLEDGEMENTS: For helpful comments and conversations, I wish to thank David Bloom, Bill D'Alessandro, James Evershed, Paul Forrester, Frank Hong, Simon Goldstein, Dan Greco, Cameron Domenico Kirk-Giannini, Nate Sharadin, Christian Tarsney, and the attendees at the May 23rd, 2024 meeting of the Global Priorities Institute's AI Work-In-Progress Group. I also wish to thank the Center for AI Safety for their financial support.

References

Angwin, J. et al. (2023, December 20). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused Decision-Making processes. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1912.02675

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature, 563*(7729), 59–64.

Bales, A., D'Alessandro, W., & Kirk-Giannini, C. D. (2024). Artificial intelligence: arguments for catastrophic risk. *Philosophy Compass*, *19*(2). https://doi.org/10.1111/phc3.12964

Barrett, J., & Schmidt, A. T. (2024). Moral uncertainty and public justification. *Philosophers Imprint*, *24*(1).

Baum, S. D. (2017). Social choice ethics in artificial intelligence. *AI & Society*, *35*(1), 165–176.

Baum, S. et al. (2022). Lessons for artificial intelligence from other global risks. In M. Tinnirello (Ed.), *The Global Politics of Artificial Intelligence* (pp. 103–131). CRC Press.

Bhargava, V., & Kim, T. W. (2017). Autonomous vehicles and moral uncertainty. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (pp. 5–19). Oxford University Press.

Binmore, K., Rubinstein, A., & Wolinsky, A. (1986). The Nash bargaining solution in economic modelling. *The RAND Journal of Economics*, *17*(2), 176. https://doi.org/10.2307/2555382

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

Bostrom, N. (2009, January 1). Moral uncertainty – towards a solution? *Overcoming Bias*. https://www.overcomingbias.com/p/moral-uncertainty-towards-a-solutionhtml

Brams, S. J., & Kilgour, D. M. (2001). Fallback bargaining. *Group Decision and Negotiation*, *10*(4), 287–316. https://doi.org/10.1023/a:1011252808608

Briggs, R. A. (2023). Normative theories of rational choice: Expected utility. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2023). Retrieved from https://plato.stanford.edu/archives/fall2023/entries/rationality-normative-utility/

Buchak, L. (2013). *Risk and rationality*. Oxford University Press.

Buchak, L. (2022). Normative theories of rational choice: Rivals to expected utility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022). Retrieved from https://plato.stanford.edu/archives/sum2022/entries/rationality-normative-nonutility/

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163.

Congar, R., & Merlin, V. (2012). A characterization of the maximin rule in the context of voting. *Theory and Decision*, *72*(1), 131–147.

Conitzer, V., et al. (2016). Rules for choosing societal tradeoffs. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (pp. 460–467).

Conitzer, V., et al. (2017). Moral decision making frameworks for artificial intelligence. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (pp. 4831–4835).

Conitzer, V., et al. (2024). Social choice for AI alignment: Dealing with diverse human feedback. *Unpublished paper*. Retrieved from https://arxiv.org/abs/2404.10271

Conley, J. P., & Wilkie, S. (2012). The ordinal egalitarian bargaining solution for finite choice sets. *Social Choice and Welfare*, *38*(1), 23–42.

Corbett-Davies, S., et al. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806).

D'Alessandro, W. (forthcoming). Deontology and safe artificial intelligence. *Philosophical Studies*. https://doi.org/10.1007/s11098-024-02174-y

D'Alessandro, W., Lloyd, H. R., & Sharadin, N. (2023). Large language models and biorisk. *The American Journal of Bioethics*, *23*(10), 115–118. https://doi.org/10.1080/15265161.2023.2250333.

Danaher, J., et al. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data and Society, 4*(2).

Dietrich, F., & List, C. (2016). Probabilistic opinion pooling. In A. Hájek & C. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy* (pp. 519–542). Oxford University Press.

Ecoffet, A., & Lehman, J. (2021). Reinforcement learning under moral uncertainty. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 2926–2936). Proceedings of Machine Learning Research.

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy and Technology, 34*(4), 1607–1622.

Feffer, M., Heidari, H., & Lipton, Z. C. (2023, May 27). Moral machine or tyranny of the majority? *Unpublished paper*. Retrieved from https://arxiv.org/abs/2305.17319

Freedman, R., et al. (2020). Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence, 283*, 103261.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437.

Gastil, J., & Richards, R. (2013). Making direct democracy deliberative through random assemblies. *Politics and Society, 41*(2), 253–281.

Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor: The "ideal observer" meets artificial intelligence. *Philosophy and Technology, 31*(2), 169–188.

Greaves, H., & Cotton-Barratt, O. (2023). A Bargaining-Theoretic approach to moral uncertainty. *Journal of Moral Philosophy, 21*(1–2), 127–169.

Grgić-Hlača, N., et al. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference* (pp. 903–912).

Gritsenko, D., & Wood, M. (2022). Algorithmic governance: A modes of governance approach. *Regulation and Governance, 16*(1), 45–62.

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs, 49*(2), 209–231.

Hendrycks, D., & Mazeika, M. (2022, September 20). X-risk analysis for AI research. *Unpublished paper*. Retrieved from https://arxiv.org/abs/2206.05862

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. *Unpublished paper*. Retrieved from https://arxiv.org/abs/2306.12001

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, *21*(3), 669–684.

Himmelreich, J. (2020). Ethics of technology needs more political philosophy. *Communications of the ACM, 63*(1), 33–35.

Hurwicz, L., & Sertel, M. R. (1999). Designing mechanisms, in particular for electoral systems: The majoritarian compromise. In M. R. Sertel (Ed.), *Economic behaviour and design* (Vol. 4, pp. 69–88). Palgrave Macmillan.

Jackson, F. (1991). Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics, 101*(3), 461–482.

Jakesch, M., et al. (2022). How different groups prioritize ethical values for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 310–323).

Kaczmarek, P., Lloyd, H. R., & Plant, M. (forthcoming). Moral uncertainty, proportionality, and bargaining. *Ergo*.

Kaler, J. (2002). Morality and strategy in stakeholder identification. *Journal of Business Ethics, 39*(1–2), 91–99.

Karnein, A. (2016). Can we represent future generations? In I. González-Ricoy & A. Gosseries (Eds.), *Institutions for future generations* (pp. 83–97). Oxford University Press.

Kıbrıs, Ö., & Sertel, M. R. (2007). Bargaining over a finite set of alternatives. *Social Choice and Welfare, 28*(3), 421–437.

Kim, R., et al. (2018). A computational model of commonsense moral decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 197–203).

Klare, M. (2023). Pentagon seeks to facilitate autonomous weapons deployment. *Arms Control Today, 53*(2), 32–33.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (p. 43).

Koster, R., et al. (2022). Human-centred mechanism design with democratic AI. *Nature Human Behaviour, 6*(10), 1398–1407.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability and Transparency*.

Lee, M. K., et al. (2019). WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 181.

Lera-Leri, R., et al. (2022). Towards pluralistic value alignment: Aggregating value systems through ℓp-regression. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (pp. 780–788).

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Lloyd, H. R. (2022). *The property rights approach to moral uncertainty*. Happier Lives Institute Working Paper. Retrieved from https://www.happierlivesinstitute.org/wp-content/uploads/2022/10/The-property-rights-approach-to-moral-uncertainty-MASTER.docx.pdf

Lloyd, H. R. (2024). The maximise socially expected choiceworthiness approach to machine ethics. Unpublished manuscript.

Lockhart, T. (2000). *Moral uncertainty and its consequences*. Oxford University Press.

MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral uncertainty*. Oxford University Press.

Marijan, B. (2022, November 28). Autonomous weapons: The false promise of civilian protection. Centre for International Governance Innovation. Retrieved from https://www.cigionline.org/articles/autonomous-weapons-the-false-promise-of-civilian-protection/

Martin, D. (2017). Who should decide how machines make morally laden decisions? *Science and Engineering Ethics, 23*(4), 951–967.

Martinho, A., Kroesen, M., & Chorus, C. (2021). Computer says I don't know: An empirical approach to capture moral uncertainty in artificial intelligence. *Minds and Machines, 31*(2), 215–237.

Mayhew, A., et al. (2022). Envisioning ethical mass influence systems. *Proceedings of the Association for Information Science and Technology, 59*(1), 756–758.

Miconi, T. (2017). The impossibility of "fairness": A generalized impossibility result for decisions. Unpublished manuscript. Retrieved from https://arxiv.org/abs/1707.01195

Miles, S. (2017). Stakeholder theory classification: A theoretical and empirical evaluation of definitions. *Journal of Business Ethics, 142*(3), 437–459.

Mittelstadt, B. (forthcoming). Interpretability and transparency in artificial intelligence. In C. Véliz (Ed.), *The Oxford handbook of digital ethics*. Oxford University Press.

Nash, J. F., Jr. (1950). The bargaining problem. *Econometrica, 18*(2), 155–162.

Newberry, T., & Ord, T. (2021). The parliamentary approach to moral uncertainty. *Future of Humanity Institute*, technical report 2021-2.

Ngo, R., Chan, L., & Mindermann, S. (2023, February 22). The alignment problem from a deep learning perspective. Unpublished manuscript. Retrieved from https://arxiv.org/abs/2209.00626

Nicolò, A., & Perea, A. (2005). Monotonicity and equal-opportunity equivalence in bargaining. *Mathematical Social Sciences, 49*(2), 221–243.

Noothigattu, R., et al. (2018). A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (pp. 1587–1594).

Oddie, G. (1994). Moral uncertainty and human embryo experimentation. In K. W. M. Fulford, G. Gillett, & J. M. Soskice (Eds.), *Medicine and moral reasoning* (pp. 144–161). Cambridge University Press.

Peterson, M. (2018). The value alignment problem: a geometric approach. *Ethics and Information Technology*, *21*(1), 19–28.

Pierson, E. (2018). Demographics and discussion influence views on algorithmic fairness. Unpublished manuscript. Retrieved from https://arxiv.org/abs/1712.09124

Prasad, M. (2019). Social choice and the value alignment problem. In R. V. Yampolskiy (Ed.), *Artificial intelligence safety and security* (pp. 291–314). CRC Press.

Rechnitzer, T. (2020). Precautionary principles. *The Internet Encyclopedia of Philosophy*. Retrieved from https://iep.utm.edu/pre-caut/

Regan, D. (1980). *Utilitarianism and cooperation*. Clarendon Press.

Robinson, P. (Forthcoming). Moral disagreement and artificial intelligence. *AI and Society*.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Sakovics, J. (2004). A meaningful two-person bargaining solution based on ordinal preferences. *Economics Bulletin, 3*(26), 1–6.

Scharre, P. (2016). *Autonomous weapons and operational risk*. Center for a New American Security.

Sepielli, A. (2009). What to do when you don't know what to do. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics* (Vol. 4, pp. 5–28). Oxford University Press.

Sepielli, A. (2010). *'Along an imperfectly lighted path': Practical rationality and normative uncertainty* (PhD dissertation). Department of Philosophy, Rutgers University.

Sharadin, N. (forthcoming). Morality first? *AI & Society*. https://doi.org/10.1007/s00146-024-01926-y

Sinnott-Armstrong, W., & Skorburg, J. A. (2021). How AI can aid bioethics. *Journal of Practical Ethics, 9*(1), 1–22.

Sprumont, Y. (1993). Intermediate preferences and Rawlsian arbitration rules. *Social Choice and Welfare, 10*(1), 1–15.

Steele, K., & Stefánsson, H. O. (2020). Decision theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Retrieved from https://plato.stanford.edu/archives/win2020/entries/decision-theory/

Takeshita, M., Rafal, R., & Araki, K. (2023, June 20). Towards theory-based moral AI: Moral AI with aggregating models based on normative ethical theory. Unpublished manuscript. https://doi.org/10.48550/arXiv.2306.11432

Tarsney, C. J. (2021). Vive la différence? Structural diversity as a challenge for metanormative theories. *Ethics, 131*(2), 151–182.

Tenenbaum, S. (2017). Action, deontology, and risk: Against the multiplicative model. *Ethics, 127*(3), 674–707.

Thomsen, F. K. (2022). Iudicium ex machinae: The ethical challenges of ADM at sentencing. In J. Ryberg & J. V. Roberts (Eds.), *Sentencing and artificial intelligence* (pp. 252–276). Oxford University Press.

Thomson, W. (1994). Cooperative models of bargaining. In R. Aumann & S. Hart (Eds.), *Handbook of game theory with economic applications* (Vol. 2, pp. 1237–1284). Elsevier.

Tollefsen, D. P. (2015). *Groups as agents*. Polity.

Vandamme, P.-E., & Verret-Hamelin, A. (2017). A randomly selected chamber: Promises and challenges. *Journal of Public Deliberation, 13*(1), 5.

Volij, O., & Winter, E. (2002). On risk aversion and bargaining outcomes. *Games and Economic Behavior, 41*(1), 120–140.

Walker, M. (2019). Consequentialism, deontology, and artificial intelligence safety. In R. V. Yampolskiy (Ed.), *Artificial Intelligence Safety and Security* (pp. 411–421). CRC Press.

Wedgwood, R. (2013). Akrasia and uncertainty. *Organon F, 20*(4), 484–506.

Wedgwood, R. (2017). Must rational intentions maximise utility? *Philosophical Explorations, 20*(S2), 73–92.

Whittlestone, J. et al. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200.

Wong, D. B. (1992). Coping with moral conflict and ambiguity. *Ethics, 102*(4), 763–784.

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns, 3*(4), 100455.

Zwarthoed, D. (2018). Political representation of future generations. In M. Düwell, G. Bos, & N. van Steenburgen (Eds.), *Towards the Ethics of a Green Future: The Theory and Practice of Human Rights for Future People* (pp. 79–109). Routledge.