Metanormative Methods

This is a supplement to Rethink Priorities' Worldview Investigation Team's CRAFT Sequence. The purpose of this supplement is to provide a survey introduction to some of the key components of our Moral Parliament Tool.

Contents

Introduction

Worldviews encompass sets of normative commitments that guide us in our moral decision-making. They contain information about what is valuable, what is moral, and how to respond to risk. However, how should we make decisions when we are uncertain between worldviews? A metanormative method takes worldviews and our credences in them as inputs and produces some action guidance as an output. Many proposed metanormative methods have taken inspiration from political processes involving agents who differ in their conceptions of the good and their decision-making strategies. Different political analogies suggest different conceptions of what a decision-making agent is like and predict different results about the effects of uncertainty. The number of potential analogies is as vast as the number of proposed political processes. Our goal here is to taxonomize some of the high-level political approaches to metanormative uncertainty (focusing on the methods in our **Moral Parliament Tool**) and to highlight some of their representational commitments and predictions.

An example of moral uncertainty

Take the following scenario. A rural village has a growing human population that it is struggling to feed, so it wants to expand its grazing territory into the adjacent countryside. However, the village abuts a forest that is home to an endangered endemic species of monkeys that doesn't have suitable habitat elsewhere. If the forest is razed, the monkeys will starve.[1] However, a greater number of humans will be fully nourished. If the forest is not razed, then many villagers will face nutritional deficiencies, leading to serious health problems and possible death.

[1] To make things simple, we'll only consider the effects on humans and monkeys. Razing a forest would, of course, affect many other sentient (or possibly sentient) creatures that would need to be considered in a full moral accounting.

You are tasked with deciding what should be done with the forest.[2] You are morally uncertain, assigning some credence to each of the following worldviews, which give very different recommendations about what you ought to do:

> **Species-neutral justice:** The welfare of all individuals matters equally, regardless of species. Justice requires that we secure a minimal amount of welfare for every individual, not that we maximize the overall or average welfare.
> *Recommendation:* preserve the monkeys' habitat because it is necessary for them to live.[3]

> **Species-neutral utilitarianism:** The welfare of all individuals matters equally, regardless of species. The correct action is the one that maximizes overall welfare, even if it requires sacrificing the interests of some individuals.
> *Recommendation:* raze the forest because it will result in greater overall welfare.

> **Humans-only prioritarianism:** Human welfare matters much more than monkey welfare. The correct action is the one that has the best overall consequences for welfare, where the welfare of the worst off is given extra weight.
> *Recommendation:* raze the forest because that will save humans, and the interests of the monkeys are not morally important in comparison.

Suppose your credences in these worldviews are .4, .3, and .3, respectively. What should you do in light of this uncertainty? A few possibilities suggest themselves.

---

[2] This scenario may be analogous to decisions about how to allocate limited funds among animal- and human-affecting charities.

[3] In some cases, it's not entirely clear what action a worldview recommends. For example, it may be impossible to secure a threshold of welfare for every individual, in which case, Species-Neutral Justice would likely recommend saving the monkeys and helping to secure human welfare by donating food or encouraging villagers to relocate to urban areas with less food insecurity. See Nussbaum (2022, Ch. 8) for a discussion of tragic conflicts of the kind sketched here.

Since two worldviews recommend razing the forest, perhaps that option should win out. But the worldview that you are most confident in tells you that razing the forest is wrong, so perhaps that should win. Maybe there are compromise positions (razing half of the forest, say) that would be best. Different metanormative models will represent the interactions of these worldviews in different ways, with different results about what you ought to do.

High-level distinctions

We can draw several distinctions among types of methods for navigating worldview uncertainty. A first distinction concerns the number of worldviews that make some difference to the final outcome (such that if their preferences were different, or if they were assigned a different credence, the final allocation would be different). Some methods are aggregative: they somehow synthesize the recommendations of multiple worldviews, and the action guidance that emerges will (probably) differ from any of the particular worldviews that they started with. Others are non-aggregative: they provide a method for selecting a particular worldview for action.

A second distinction concerns the method through which decisions are made. Some methods are democratic: the course of action is decided via a political process among hypothetical agents who each act to promote the interests of a specific worldview. Voting and bargaining are the two most prominent analogies here.

- In a voting process, actions are proposed to the voting body, and each delegate votes in a way that represents the interests of her worldview. Voting can be aggregative, where the selected action reflects some compromise position among worldviews. It can also be non-aggregative if instead of choosing an action, the voting method selects a single worldview that is given the authority to act.

- In a bargaining process, the distribution of resources emerges from negotiations and transactions that participants make with one another. For example, we might give each worldview a share of the budget proportional to its number of delegates to do with as it pleases, and worldviews can make deals with each other if they would be mutually beneficial (Kaczmarek, et al., ms). Bargaining is aggregative by default.[4]

Other methods are non-democratic. They operate as a dictator who observes the various worldviews that are part of its uncertainty set and makes a decision on their behalf. For example, a social welfare function operates as an aggregative dictator that takes the preferences of various worldviews into account and selects the optimal action given those preferences. A non-aggregative dictator, like My Favorite Theory, selects a single worldview and makes decisions on its preferences alone.

	Aggregative	Non-aggregative
Democratic	Bargaining, some voting	Some voting
Non-democratic	Social welfare functions	My Favorite Theory

We can evaluate these various approaches—and particular instantiations of voting, bargaining, and the like—in two main ways. First, we can evaluate an analogy for its *representational aptness*. A good model of decision-making under uncertainty need not be entirely descriptively accurate about the psychology of an individual or the deliberative process of a group. Nevertheless, it should capture something important about the normative structure of the decision. Note that the aptness of a particular political analogy might be very different depending on

[4] A mixed process involves both bargaining and voting. For example, theories could bargain with each other to make proposals which could then be voted on. Alternatively, proposals might have to pass an initial round of voting, and participants bargain to decide among the remaining options. Actual political procedures (such as passing a bill in the US Congress) often involve many iterated rounds of voting and bargaining.

whether we're modeling decision-making by individuals, group agents (e.g. charities), or groups of agents (e.g., a moral community like EA).

Second, we can evaluate an analogy for its *functional aptness*. Does the method yield intuitively correct results about what we ought to do? Does it highlight the right metanormative reasons for arriving at a particular answer? Does it get the normative structure of considerations correct, and does that normative structure correspond with the ways in which decisions should actually be made? Does it predict when and why compromises will occur? Does it help decision-makers design better ways of making decisions?

Non-democratic models

Non-democratic models portray the decision-maker as a dictator. The dictator can use methods that aggregate the views of its constituents, but the constituents do not have a direct say over the resulting decision. There is no requirement that the selected course of action is one recommended by any of the worldviews. A helpful analogy here might be the relationship between a CEO and an advisory board, where the CEO must be responsive to input of advisors but can integrate this input in the way she sees fit. Alternatively, we could model the agent as an investor seeking to maximize the amount of moral value from various sources.

Non-democratic methods seem psychologically realistic for individuals; our theories don't directly produce actions, they do so only when filtered through some central decision-making mechanism. On the other hand, this model pushes the hard explanatory work to the next question: what process does the dictator use to make decisions in light of her worldviews?

My Favorite Theory

My Favorite Theory (Gustafsson and Torpman 2014) is one prominent non-democratic, non-aggregative method in which the dictator adopts the worldview that has the highest credence and acts on its recommendation.[5]

My Favorite Theory seems to embody the common-sense idea that you should act on what you most strongly believe. One problem here is that you may not be very confident in your favorite theory. In our working example above, the favorite theory had a credence of .4. If you individuate theories in a fine-grained way, your highest credence may be far lower than that.[6] Further, in our working example, theories that recommend razing the forest collectively have a higher credence than the theory that recommends saving it. In one sense, the agent most strongly believes that she should save the forest (via the theory of which she's most certain), but in another sense, she most strongly believes she should raze it.

A second problem with My Favorite Theory is that uncertainty has no real effect on your decision-making; you should behave the same way whether you are completely certain of a worldview or whether you are just slightly more certain of it than an array of competing views. If you think that a good model of moral uncertainty should show that uncertainty matters, then MFT will fail to be functionally adequate.

Social welfare functions

Some non-democratic methods instead incorporate the preferences of diverse worldviews to arrive at some all-things-considered optimal action. They start by generating each worldview's utility or choiceworthiness function, a measure of the degree of normative support that the worldview assigns to each

---

[5] Note that in the Moral Parliament Tool, My Favorite Theory recommends a spread over multiple projects rather than a single project. This is because projects are modeled as having diminishing returns, so a worldview can do best by distributing its resources across multiple projects. If there were no diminishing returns (which can be done by changing the settings in the Allocations page), MFT would recommend the top project of the top worldview.

[6] See MacAskill and Ord (2020) on this point.

option. Then, they input these utility functions and credences over worldviews into an algorithm to find the action that optimally satisfies the aggregated preferences. Different social welfare functions reflect different normative commitments about how to navigate uncertainty.

Utilitarian social welfare function / Maximize Expected Choiceworthiness

One prominent approach to making decisions under normative uncertainty is to use the same tools as decisions under empirical uncertainty. The expected utility of an action is the sum of its payoffs in different states, weighted by our credences in those states. The utilitarian welfare function / expected choiceworthiness is the sum of the utilities assigned to an option by each worldview, weighted by the credence in each worldview. This can be interpreted as the distribution that achieves the highest weighted average of normative support, aggregated across delegates' utility functions. The Maximize Expected Choiceworthiness (MEC) framework states that the option with the highest expected choiceworthiness is the best overall option.[7]

A consequence of MEC is that theories that assign extreme utilities will come to dominate meta-normative decision-making, even when we assign them low credences. This threatens to lead to moral fanaticism, in which moral uncertainty is swamped by the extremely large stakes of implausible moral theories. For example, suppose you assign a miniscule non-zero credence to the view that cutting down a tree is gravely morally wrong, as wrong as killing many people. This theory will dominate your decision-making, causing you to save the forest, even if you deem the view very implausible.

Other social welfare functions

Other kinds of social welfare functions embody different views about how to navigate moral uncertainty. For example, Rawls's difference principle can be formulated by a maximin function that ranks alternatives by the utility of the

[7] MacAskill (2016) uses "appropriateness" to refer to the overall metanormative utility.

worst-off individual. In this context, maximin tells you to evaluate each option by the lowest score it receives across all worldviews. You should choose the option that has the highest low score. This embodies a kind of moral risk-aversion, declining to take an action that any of your worldviews deems to be extremely bad. For this reason, it is also susceptible to fanaticism, as you may assign some extremely low credences to worldviews that imply that seemingly good actions are, instead, morally horrific (e.g., the tree example above). Other options include:

- Prioritarian functions, which are like MEC but give greater weight to lower utility levels, prioritizing the theories on which an action is deemed morally worse.
- Headcount rules, which rank options by how many worldviews assign it a utility above some threshold value.[8]

Evaluating non-democratic methods

What makes a method non-democratic is that worldviews are not modeled as autonomous actors in the meta-normative process. Their preferences are combined or filtered by an algorithm that reflects normative commitments about how to act under normative uncertainty.

There are several advantages to non-democratic methods. First, even though actual decision-makers are probably not executing social welfare algorithms, the conception of the individual as a dictator who makes decisions with worldviews as inputs is plausible. Second, they make for a helpful continuity between formal methods of decision-making under empirical uncertainty and under normative uncertainty. Third, they are highly flexible, transparent, and predictable; it is relatively easy to see what the normative commitments of different functions are and how they will operate.

The chief disadvantage of aggregative non-democratic methods is the problem of intertheoretic comparisons. In order to aggregate the utility functions of different moral theories, we must be able to map the utilities assigned by various theories onto the same cardinal scale. For example, MEC depends crucially on the

[8] As we will see, this kind of social welfare function will function the same way as approval voting.

possibility of intertheoretic comparisons. A positive affine transformation of worldview A's utility function (leaving B's intact) will cause A to receive a much higher relative MEC. Unless we can pin each utility function to a common scale, MEC's recommendations may be a mere artifact of differences in scale.

While scale commensurability may be achievable for nearby worldviews (e.g., consequentialist theories with overlapping axiologies), it becomes increasingly difficult for worldviews that have very different kinds of value structures. When the nature of normative support differs significantly on different worldviews, philosophers have been highly skeptical that intertheoretic comparisons of this kind are possible.[9] For example, it is difficult to say whether the moral difference between razing the forest and saving it is greater or lesser on Species-Neutral Justice than Humans-Only Prioritarianism.

Democratic Methods

Democratic methods model worldviews as autonomous participants in a democratic political process.[10] One of the chief advantages of democratic methods is that they (arguably) do not require us to solve the problem of intertheoretic comparisons. The preferences of different worldviews are not aggregated directly, but only indirectly as the result of a democratic process. Political theory, especially in the tradition of political liberalism, is rife with theories about how we can best structure deliberative processes involving worldviews that have very different conceptions of the good. One significant cost is that there is no simple mathematical function from credences and utilities to all-things-considered choiceworthiness. Instead, we must model more complicated procedures. A second

[9] Though see Cotton-Barratt, *et al.* (2020) and Carr (2022) for discussions of attempts to resolve this problem. Carr argues that theories with merely ordinal rankings (e.g., A is ranked higher than B but the theory doesn't say by how much it is higher) pose a more serious problem of intertheoretic comparisons.
[10] We shouldn't overemphasize the difference between democratic and non-democratic methods. Indeed, certain voting methods will end up recapitulating certain social welfare functions; e.g. range voting will deliver the same results as MEC (Newberry & Ord 2021). As we have noted, headcount social welfare functions resemble approval voting.

cost is that there are many candidate democratic procedures we might use, all of which have virtues and drawbacks of their own.

Voting vs. Bargaining

In a voting process, worldviews are represented as delegates to a moral parliament, where each has a proportion of delegates equal to the credence assigned to it. In a bargaining process, worldviews control a share of an overall budget (most naturally, a share proportional to their credence). They can negotiate and make deals with other worldviews to spend their budget in ways that will further their worldview's interests.

These methods have different representational commitments about what a deciding entity is like. Voting methods funnel competing worldviews through a single deciding body that then acts. In bargaining, worldviews are much more autonomous. There is no unified entity that decides; the overall allocation of resources is a summary of the actions taken by entities within the deliberating body.[11]

The representational aptness of these two approaches depends on the kind of agent we are modeling. On the one hand, it can be somewhat odd to represent individual people via bargaining methods as if their beliefs (even those they assign low credence to) are able to independently control their actions. This is particularly odd if an individual assigns credence to worldviews that are at cross-purposes. On the other hand, bargaining might reflect a kind of respect that an individual gives to the worldviews of which she is uncertain, making gestures toward ends that each values. Bargaining is a more literal representation of group entities, such as charitable organizations or moral communities. These entities really are made up of autonomous agents who have some degree of local control over resources.

We should use information about the entities we are modeling when designing democratic methods. For example, when modeling individual decision-makers via a voting process, we can assume that theories have full

[11] We also shouldn't overemphasize the distinction between voting and bargaining. For example, we can make bargaining more like voting by putting bargainers in the same ship, say, by imposing penalties for failing to reach consensus.

information about one another and can't vote strategically.[12] It is less clear how much group-level transparency and goodwill we should assume in the case of bargaining. For example, should we permit worldviews to extort others? If not, how do we formally prohibit it?

Types of Voting Processes

Voting methods matter. Holding a group of people and their preferences fixed, changing the method of voting can yield very different results. There are various desirable formal features that we would want from a voting method, along with practical considerations like ease of implementation.[13] Unfortunately, no voting system satisfies all proposed desiderata in all circumstances. Voting theory is a centuries-old discipline, and we cannot do justice to the enormous number of proposed voting methods and arguments for and against them. Here, we will briefly present popular voting methods, including those that we have used in the Moral Parliament Tool, giving some sense of their representational commitments and functional properties.

In the tool, worldviews are represented by a number of delegates proportional to the credences the user has in those worldviews. These delegates vote on options for allocating funds across various projects. The option set contains all possible allocations across selected projects, up to a certain level of grain. For example, the option set includes: {100% funding to Tuberculosis Initiative, 0% to all others}; {90% funding to Tuberculosis, 10% to Direct Transfers, 0% to all others}; {equal funding across all projects}, etc. Worldviews' preferences over allocations are a function of what they care about and the assumption that funding to any project will have diminishing marginal returns.[14]

---

[12] "While tactical voting is a real problem when it comes to aggregating the stated preferences of people, it is no problem at all in the context of decision-making under normative uncertainty. Theories are not agents, and so there is no way that they can conceal their choiceworthiness ordering. If a decision-maker were to pretend that one theory 's choice-worthiness ordering were different from how it is, she would only be deceiving herself" (MacAskill 2016, 998)

[13] Example desiderata include: Pareto efficiency, if every voter prefers candidate A to candidate B, then B should not defeat A; Majority rule, if most voters prefer A to B, then A should defeat B; Monotonicity, A's receiving more support from the voters should not adversely affect A's chances of winning. And so on.

[14] More specifically, our default is to model preferences as the sum of the square roots of demands. The square root function, being concave, aptly represents risk-averse behavior, typical in economic

Plurality Voting

In plurality voting, every participant gets a single vote, and the option with the most votes wins. If every worldview has a different favorite option in the option set (which we expect to be common in real-world cases such as dividing up a budget), then plurality voting will yield the same results as My Favorite Theory. If some theories have the same favorites, then they can outvote the top theory, and the vote will represent more of a consensus view.[15] In our working example, Species-Neutral Justice casts 40% of the votes in favor of saving the forest, while Species-Neutral Utilitarianism and Humans-Only Prioritarianism cast the remaining votes for razing it.

Though plurality voting is extremely common and quite straightforward, it has serious flaws.[16] For example, it can pick a "Condorcet loser, an option that would lose to every other option in a head-to-head match-up. Consider the following example from Pacuit (2019):

# Voters	Ranking
1	$A \ B \ C$
7	$A \ C \ B$
7	$B \ C \ A$
6	$C \ B \ A$

Option A receives the most first-place votes so is declared the winner. However, most voters prefer each of B and C to A. Plurality voting will miss out on consensus options that might be ranked highly on many theories, even if they are the top

decision-making, where the value of each additional unit of investment decreases as more is invested. This aligns with the realistic scenarios of funding allocation, where initial investments yield significant returns but lead to diminishing benefits as investment increases. Moreover, this modeling choice strikes a balance between capturing complex economic behaviors and maintaining mathematical tractability, making it a pragmatic choice for analyzing and simulating budget distribution decisions. The Utility Discount setting can be changed by users.

[15] This is more likely to happen when there are fewer options in the option set. In the Moral Parliament Tool, the option set is comprised of distributions over projects with a fairly fine grain. Therefore, we would expect plurality voting to usually match My Favorite Theory.

[16] Indeed, when voting experts took a vote on preferred voting methods, plurality voting received no support (Laslier 2012). Approval voting, which we will discuss below, was the favorite.

choice of few. One problem is that plurality voting is insensitive to the way voters order non-first-place options.

Approval Voting

Unlike plurality voting, approval voting allows delegates to express support for more than one option. The option that has the highest total number of approval votes is the winner. As a result, approval voting tends to recommend options that have broad appeal, even if they are no one's top choice. The threshold for approval can be modeled in different ways. Voters may have some absolute threshold that an option must cross to be judged acceptable or they might instead use some relative threshold (e.g., approving their top 25% of options).

In the approval voting method of our Moral Parliament Tool, the option set is the set of all possible allocations of funds across projects (specified to a certain grain).[17] Each delegate has an ideal allocation, and the utility that they would get from that ideal allocation is their maximum utility, u*. For each potential allocation, the utility for each delegate is calculated. Delegates can approve or disapprove of a proposed distribution. Their decision is based on whether a distribution would yield them a sufficiently high amount of utility (by their own lights). An allocation is approved by an agent if the utility exceeds a certain percentage (determined by a 'strictness' parameter) of their maximum possible utility, u*. We use a default strictness of 0.8, meaning that delegates will approve of all distributions that get them 80% of u* and disapprove of all those below it.[18] A lower strictness means that more distributions will be approved, whereas a strictness of 1 means that delegates will only vote for their ideal distribution. Approval voting tends to favor allocations that diversify across projects, giving many worldviews some of what they want.

Approval voting reduces voters' judgments to a binary yes-I-approve or no-I-disapprove. This may be an apt model of judgments of moral permissibility

[17] We only consider those allocations that fully utilize the budget. It's optimal to exhaust the budget in this context (where, for at least one project – and typically all of them– more funds allocated to it is always better).
[18] This can be adjusted via the "Voting Threshold" parameter.

or impermissibility, and unlike some other methods, it can be used with worldviews that give merely ordinal rankings.[19] However, by leaving out information about how strongly voters approve or disapprove of various options, it is relatively insensitive to stakes. On the one hand, this means that it is resistant to fanaticism.[20] On the other, it can favor watered-down consensus options. For example, if 50% of voters approve of option A, all judging it to be merely okay, and 45% of voters approve of option B, deeming it to be by far the best option, approval voting will select A.

Ranked Choice Voting

In Ranked Choice Voting (also known as Instant Runoff Voting), voters rank options according to their preferences. If no candidate receives a majority of first-place votes, the candidate with the fewest first-place votes is eliminated, and their votes are redistributed to the candidate that is the highest-ranked remaining option on each ballot. This process continues until a single candidate has an absolute majority of first-preference votes.[21]

RCV aims to ensure that the winner enjoys broad support by allowing voters to express their preferences more fully than in a simple majority or approval voting system. Like approval voting, RCV heeds more than just the top choice of each delegate. However, like plurality voting, it is highly sensitive to first-place rankings. For example, suppose that voters are split between options A-G as their first preferences but all rank option H second. H will be eliminated off the bat, even though it intuitively has more support among the body of voters than any of A-G. Unlike approval voting, the way that voters rank non-optimal choices is also taken into account. Therefore, it tends to favor projects that some worldviews like the most and that lots of worldviews rank fairly highly.

In our Moral Parliament, RCV differs from the other aggregation methods in that delegates vote on single projects rather than allocations across projects. The

---

[19] Approval voting resembles My Favorite Option, according to which you ought to choose that option that is most likely to be permissible across all of your worldviews. See Tarsney (2021) for a discussion.
[20] Indeed, fanatical theories tend to be punished since they will only approve of their favorite options.
[21] This is one way of doing Ranked Choice Voting, but there are others (see here for a list and examples of uses of each).

effects of RCV are somewhat unpredictable depending on the worldviews and projects involved.

Other voting methods

Condorcet voting methods compare candidates in a round-robin tournament, and a Condorcet winner is a candidate who would win a head-to-head contest against every other candidate.[22] If there is no such winner, then there are numerous methods for selecting the best option based on head-to-head records. For example, the Simpson-Kramer method selects the candidate who has the smallest worst loss in all head-to-head comparisons, whose largest margin of defeat is smaller than that of any other candidate.

MacAskill (2016) defends the Credence-Weighted Borda Score as the best voting method for choosing the most choiceworthy option under moral uncertainty. Borda voting is an alternative round-robin procedure that takes into account a candidate's performance across all interactions, not just their worst loss. A project's Borda Score according to a worldview is calculated by counting the number of options that are less preferred than it, and subtracting the number of options that are more preferred.[23] A project's Credence-Weighted Borda Score is the sum of its Borda Scores across all worldviews, with each score weighted by the proportion of delegates from that worldview in the parliament. In the Moral Parliament Tool, we evaluate Borda scores for individual projects (i.e. allocations that give the entire budget to a single project)[24], and the method selects the project with the highest such score.

[22] Recall the point above that a Condorcet winner can nevertheless lose in a plurality voting contest.
[23] Unlike traditional Borda Scores, which often omit the subtraction step, our method includes it for tie-breaking purposes, like in Saari (1990).
[24] This was done for computational reasons since a round-robin tournament among every possible allocation of funds would be extremely complex.

Bargaining

In a bargaining process, delegates are self-interested[25] and control shares of the group's resources. A delegate attempts to maximize the amount of utility that she will receive from the aggregate's allocation of resources and can use her share of the pot as a way to influence the actions of others.

To simplify matters, we will consider pairwise interactions between delegates in a parliament who can either agree or disagree to a proposed allocation of resources. A delegate will only agree if they would get more utility from the new allocation than if they decline; it is a Pareto improvement. Among the set of Pareto improvements, the Nash bargaining solution is the one that improves on the base allocation the most, where this is interpreted in terms of the product of differences in payoff for each worldview if they agree compared to if they disagree.[26] It thus "favors equal divisions of the choice-worthiness gains to be had from trade between theory representatives" (Kaczmarek, *et al.* ms, 13).

There are two central questions when modeling and predicting the effects of a bargaining process:

- What is the disagreement point? That is, what happens if the parties fail to reach an agreement?
- Given the choice of a disagreement point, under what conditions will participants get more utility by agreeing to a new allocation than disagreeing?

Consider a morally uncertain individual: what happens if your worldviews cannot agree on a course of action? Perhaps you would be paralyzed by indecision, in which case nobody gets any utility. Alternatively, you could default to a non-democratic process, such as My Favorite Theory, so a worldview that would fare very poorly under an alternative arrangement (unlike a worldview that would fare well) would have strong incentives to cooperate.

---

[25] They are trying to maximize the utility that they get, relative to their own preference ordering. Of course, in this context, the "utility" in question is moral value, so "self-interested" is a bit of a misnomer.
[26] For this reason, the problem of intertheoretic comparisons arises here too.

We will assume that if no bargain is reached, each worldview will default to spending its share of the budget as it sees fit. This disagreement point may be representationally apt when modeling group agents in which different departments have autonomy over shares of the budget. It is less apt for modeling individuals. Nevertheless, we agree with Greaves and Cotton-Barratt (2019, 6) that "the talk of different theories 'bargaining' with one another is only metaphorical, and there is not obviously any empirical fact of the matter regarding 'what would happen in the absence of agreement'. The task is simply to select some disagreement point such that bargaining theory with that choice of disagreement point supplies a satisfactory metanormative theory."

In the Moral Parliament Tool, we by default assume that if no bargain is reached, each worldview will allocate its share of the budget (proportional to its credence) in the way that will maximize its own pre-bargaining preferences (u*). An agreement will only be arrived at only if a different proposed allocation will exceed u*. One reason is that this is a plausible model of group agents and the EA community as a whole, who need not come to an agreement before investing. The second is that it provides a greater contrast between bargaining and other approaches. In the absence of a bargain, agents will split resources in accordance with their normative uncertainty. Proportionality is the default assumption, and deviations from proportionality are the thing to be explained.

In the tool, the effects of bargaining can be seen by comparing the results of Nash Bargaining and Moral Marketplace. Moral Marketplace gives each parliamentarian a slice of the budget to allocate as they see fit and then combines their chosen allocations into one budget. The Nash Bargaining solution will differ from Moral Marketplace if and only if there were bargains to be found among the delegates. In our tool, bargains happen somewhat rarely. Worldviews can find a compromise if there are two members of the parliament who have different first choices but agree on their second choices and the second choice is more than half as good as the first. Still, in Nash Bargaining all parties need to benefit from any bargain in order for it to be accepted.

There are more types of moral bargains than our tool can capture (for example, it cannot model diachronic bargains). What follows is a broader (yet still partial) taxonomy of kinds of moral bargains we might expect to see. Each taxon describes one reason agents with different preferences might agree to a collective choice over each pursuing their own favorite option.

Compromises

In a compromise, two agents devote some of their collective resources to a project that is neither of their favorites, but which they both agree would be better for them collectively to support than for each to just devote themselves to the preferred options. For instance, if one agent slightly prefers project A to project B and cares naught for C and a second slightly prefers project C to project B and cares naught for A, they might both prefer donating $200 to B than $100 to A and $100 to C. If they each have $100 to give, they might compromise on B.

Compromises depend on specific circumstances. Compromises can exist when agents have slightly different preferences or when some possible projects do well when evaluated under fairly different preferences. If each agent strongly prefers a distinct range of projects, no compromise will be possible because there will be no acceptable middle ground for them to coordinate on. Instead, each agent will see the cost of giving up their support of their favored project as too great to be met by an increase in the support of any compromise project. Whether a compromise is in the best interests of both parties will also depend on the projects' cost curves. For example, if money given to a cause has increasing marginal utility—such that the pooled efforts of both parties would make B much more effective than it would be if only one party gave to B—then compromises can yield more utility.

We might expect to see few compromises possible between agents representing standard EA worldview divisions. Each worldview has strong preferences for helping some group (animals, current humans, posterity) that has a comparatively weak claim in other worldviews. There are projects that multiple different worldviews will see as promising (e.g. neartermist and longtermist worldviews might both see value in pandemic prevention), but they must meet a

fairly high bar in order for everyone to be happy paying the opportunity costs. As a rule of thumb, we shouldn't expect to see compromise on a project unless it is regarded as at least half as promising as the favored projects of each worldview. Compromises might be more common when standard worldviews are subdivided into families of similar worldviews, because they would have more shared values to coordinate on, but we should expect to see fewer compromises across major divisions.

Trades

In a trade, two agents agree to shift support from a project they personally prefer to a project they find less appealing due to a conflict with another worldview. For instance, a trade might involve each party avoiding projects they value that the other agent finds noxious. An agent favoring aid to current human populations might make a deal with an animal welfare sympathist to avoid funding projects particularly likely to exacerbate factory farming in exchange for the latter not engaging in projects that are more likely to place disproportionate burdens on some of the world's poorest people.

If both parties were instead to pursue their favorite options, their efforts would cancel each other out. The very low utility at the disagreement point opens the opportunity for compromise. For example, suppose Worldview 1 favors giving to gun rights organizations and Worldview 2 favors gun control projects. Both agree that Oxfam is a worthy, though non-optimal, charity. Recognizing that their donations to gun causes will cancel each out, both agree to redirect their resources to Oxfam instead (Ord 2015; see Kaczmarek, *et al.* ms for similar cases). Even if Oxfam is not rated very highly at all by either party, giving to Oxfam can still be a Pareto improvement over the disagreement point (whether neither party gets any utility).

How often will EA worldviews recommend conflicting projects, and are there available trades that would leave everyone better off? Some major EA priorities rarely directly counteract one another: money given to preventing malaria does not, for instance, make AI misalignment any more probable. On the other hand, there might be significant conflicts among EA worldviews. Efforts to

end factory farming may be at cross-purposes with efforts to promote economic development in the developing world. Pursuing aligned AI might have bad consequences for animals. Developing AI may come with serious negative effects on the climate. And so on.

Like compromises, trades require special circumstances. Unlike compromises, they don't require much middle ground. If worldviews' preferences counteract, an alternative only needs to be weakly preferred by each to be an improvement over disagreement. In other cases, there don't have to be any projects that both agents like. Instead, agents need to have stronger preferences about each other's preferred projects than the proponents do.

Wagers

In a wager, two agents make a deal to agree to some distribution of support and to shift that support based on some additional unknown information about the world. The shift would be to the advantage of one agent and to the detriment of the other, but each agent may agree that the initial distribution plus the conditional shift is worthwhile in expectation.

One form a wager might take is a bet. If different worldviews have different expectations about the probability of the condition, each may think that the expectation of the bet is to their advantage. A longtermist worldview might expect that the probability of rapid advances in AI is much higher than a common sense worldview, and so might be willing to shift funding to GHD causes in case such advances don't materialize on the condition that more money is allocated to their favored causes at first. These sorts of bets depend on agents having epistemic disagreements, which may not occur between idealized worldviews. Bets might also appeal to worldviews that differ in levels of risk aversion. If one worldview discounts small probabilities, they may be willing to offer mutually beneficial bets at low probabilities to risk-neutral worldviews.

Another form a wager might take is insurance. If an agent is concerned about avoiding worst-case outcomes, they might be willing to pay a fee in order to secure additional resources conditional on the worst-case options looking

plausible. This may leave them generally less able to pursue their vision of the good but raises the floor on how bad things might be.

A third form of wager is a bet for leverage. The effectiveness of possible interventions may depend on how the world turns out. Agents are incentivized to place bets on their own effectiveness, because the value of any resources they secure on a bet on their own effectiveness will exceed the cost they pay if it turns out they are not very effective. Unlike regular bets, this does not require any epistemic disagreements between agents.

Wagers are less contextually dependent than trades or compromises. They don't require the existence of strong preferences between other agents' preferred projects. For this reason, we should expect that in a true moral bargaining scenario, wagers may be among the most common form of bargain and could conceivably have a significant effect on how resources are allocated.

Wagers have the downside that they incur risk in real-world settings beyond the explicit terms of the wager. If two agents make an agreement that has upfront costs in exchange for future benefits, circumstances may change by the time the wager is to be honored.

Loans

In a loan, two agents make a deal to exchange resources at different times so that those resources can be allocated to each's preferred options at those times. The idea here is that agents may see different value in spending resources now versus later, so it may make sense for them to coordinate when their resources are used.

Loans can make sense when different projects are time-sensitive. If one agent's favored project can only be carried out at a given time, but another agent's favored projects will be nearly as effective later, then a loan might allow both to achieve their goals more effectively. Loans also make sense between agent's whose projects tend to ramify, having greater impact the earlier people pursue them.

Like wagers, loans are less contextual than trades or compromises and require some trust that the terms will be agreed to.

Conclusion

As we can see, there are many proposals for how we can navigate moral uncertainty that differ significantly in their representational aptness, formal properties, and normative commitments. By playing with our Moral Parliament Tool, you can also see that different methods often recommend very different proposals about what we ought to do, all things considered. This result may feel unsatisfying, as we might have hoped that metanormative uncertainty would be less severe than first-order moral uncertainty. However, that does not appear to be the case. The Moral Parliament Tool allows us to explore the ramifications of these various methods and assist us in coming to some reflective equilibrium about how to deal with normative uncertainties at numerous levels.

Acknowledgments

The Moral Parliament Tool is a project of the Worldview Investigation Team at Rethink Priorities. Arvo Muñoz Morán and Derek Shiller developed the tool; Hayley Clatterbuck, Derek Shiller, and Arvo Muñoz Morán wrote this post. Thanks to Bob Fischer and David Moss for helpful feedback. If you like our work, please consider subscribing to our newsletter. You can explore our completed public work here.