FHI TECHNICAL REPORT 2021-2

The Parliamentary Approach to Moral Uncertainty

Toby Newberry & Toby Ord

Research Scholars Programme & Macrostrategy Research Group
Future of Humanity Institute, University of Oxford

2021

Abstract

We introduce a novel approach to the problem of decision-making under moral uncertainty, based on an analogy to a parliament. 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 your credence in that theory. We present what we see as the best specific approach of this kind (based on proportional chances voting), and also show how the parliamentary approach can be used as a general framework for thinking about moral uncertainty, where extant approaches to addressing moral uncertainty correspond to parliaments with different rules and procedures.

Cite as: Newbery, T. & T. Ord (2021): "The Parliamentary Approach to Moral Uncertainty", Technical Report #2021-2, Future of Humanity Institute, University of Oxford

This report is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 669751).

1 Introduction

Where Plato famously defended a model of the state as the 'soul writ large', this paper sets out an approach to moral reasoning that might be characterised as the 'state writ small'.

The basic idea behind this approach is straightforward: in conditions of moral uncertainty (see section 2), you should act as if the moral theories you find plausible are represented in an internal 'Moral Parliament' whose decisions determine your action. Just as real-world parliaments work to resolve uncertainties and disagreements in fundamental societal values, building consensus around shared goals, the Moral Parliament works towards similar ends in the context of individual decision-making. In this sense, it is an optimistic proposal: one that assumes the possibility of 'intertheoretic dialogue', and that aims at genuine compromise.

This concept of a Moral Parliament was first suggested by Nick Bostrom in 2006, and developed with Toby Ord. While an early version of the idea was presented briefly and informally by Bostrom (2009), it hasn't been set down in a formal context until now.

Section 2 presents the problem of moral uncertainty and surveys the solutions that have been proposed along with the challenges they face. Section 3 presents the parliamentary approach, drawing out some of its advantages, and assessing how it compares to the existing solutions. Section 4 considers further directions in which this approach could be developed.

2 Moral Uncertainty

2.1 The Problem

Just as one can be uncertain about empirical questions, so one can be uncertain about basic moral questions. To take an example of the former, one might be uncertain about whether it will rain today, and thus about what you should wear. To take an example of the latter, one might be uncertain about whether it is wrong to eat meat, and thus about what you should have for dinner. Philosophical approaches to *moral uncertainty* attempt to negotiate the second kind of situation: they provide accounts of what decision a person ought[1] to take, given various details of their decision-situation, including the options they face and their credences in different moral theories. More precisely, the central question addressed by approaches to moral uncertainty is as follows:

For any given set of credences in moral theories and set of options that a decision-maker can have, what is the appropriateness ordering of the options within that option set?[2]

Here, 'option sets', 'decision-makers', and 'credences' should all be understood in their standard decision-theoretic senses. The 'appropriateness' of an option refers to its intuitive superiority (or not) given the other facts of the case — it plays a role analogous to 'goodness' or 'rightness' when considering options from the perspective of a single moral theory, but applies at the meta-normative (*inter*theoretic) level, rather than the normative (*intra*theoretic) level. See MacAskill et al (2020) for a full explanation of this terminology.

[1] There is active debate about how to interpret this kind of 'ought' because it operates at a higher level than the 'ought' of each first-order moral theory under consideration and it is not clear whether it really is best thought of as an 'ought' at all. To sidestep this issue, we will thus often rephrase things in terms of 'appropriate' choice under moral uncertainty, leaving open what this amounts to, as discussed later.

[2] MacAskill et al (2020).

Perhaps surprisingly, the history of moral uncertainty — as a philosophical problem, rather than an empirical phenomenon — is just a few decades long. Aside from a localised flowering among Catholic theologians (e.g. Medina (1577), Pascal (1657)), the problem had not received serious philosophical consideration until the end of the last century. For examples of early analytic treatments, see Lockhart (1977) and especially Gracely (1996). For a succinct and relatively recent introduction to the debate, see Bykvist (2017). Harman (2015) provides a sceptical challenge to the problem *qua* problem. And for an up-to-date book-length discussion, see MacAskill et al (2020).

The following case serves to animate the central question given above, while providing an example to call on in the discussion that follows:

Kira's Dinner

Kira is deciding which of three options to order for dinner: pork pie, linguine with clams, or lentil curry. Her credence is split between four different moral views: (1) a human-centred form of utilitarianism, according to which only human welfare has moral significance (*Human Welfare*), (2) an animal-welfarist form of utilitarianism, according to which the welfare of many non-human animals (including pigs) has moral significance, but the welfare of molluscs (including clams) is negligible (*Vertebrate Welfare*), (3) a broader animal-welfarist form of utilitarianism, according to which the welfare of all animals (including humans, pigs, and clams) has moral significance (*Animal Welfare*), and (4) a deontological view, according to which it is impermissible ever to eat pork (*No Pork*). On each of the three utilitarian views only the option that maximises (morally significant) welfare is permissible.

The pork pie is the option that would maximise Kira's own welfare, followed by linguine with clams, followed by lentil curry. The pork pie and linguine with clams come with significant costs in terms of (respectively) pig and clam welfare.

Supposing Kira has credences in these theories as specified in the following table. What is her most appropriate course of action?

	40%	30%	10%	20%
	Human Welfare	*Vertebrate Welfare*	*Animal Welfare*	*No Pork*
Pork pie	Permissible	Impermissible	Impermissible	Impermissible
Linguine with clams	Impermissible	Permissible	Impermissible	Permissible
Lentil curry	Impermissible	Impermissible	Permissible	Permissible

Table 1: Overview of permissibility of different dinner options under moral theories Kira has credence in.

2.2 Proposed Solutions

2.2.1 My Favourite Theory

A simple and natural approach to moral uncertainty is to simply follow the moral theory in which you have highest credence:

My Favourite Theory (MFT)

An option is appropriate for some agent *iff* it is a permissible option according to the moral theory this agent has highest credence in.

In *Kira's Dinner*, MFT recommends choosing the pork pie; as the theory in which she has highest credence is human-centred utilitarianism, which considers this the only permissible option.

As an approach to moral uncertainty, MFT has a number of practical and theoretical virtues. First, it is admirably easy to apply. Second, it avoids making any 'intertheoretic comparisons': one need only inspect the recommendations of a single theory, without weighing these against those of other theories.[3] Variants of this approach have been endorsed by Gracely (1996) and more recently by Gustafsson and Torpman (2014).

However, MFT also faces strong objections. In particular, it is vulnerable to challenges concerning how moral theories are to be individuated.[4] Suppose you are almost certain that some form of utilitarianism is true — you hold 90% credence in this claim. At the same time, you hold 10% credence in ethical egoism. However, utilitarianism comes in many subtly distinct varieties, so your credence in it is split between ten (at least superficially) distinct variants, which you find equally plausible. MFT thus recommends you follow egoism as this has a 10% credence compared with 9% in any particular type of utilitarianism. But this means you may be required to perform an act you are 90% sure is impermissible, even when there is an alternative you are 90% sure is permissible.

In fact, the situation is worse than this single example implies: in the absence of an account of theory-individuation, MFT simply fails to function. Without such an account, one's 'favourite theory' might always turn out to be a set of different theories, meaning the determination of favourite can never be made with confidence.

2.2.2 My Favourite Option

A closely related approach, which preserves some measure of MFT's intuitive appeal while resisting the objection noted above, is My Favourite Option (MFO). Here the recommendation is to choose whichever *option* one thinks is most likely to be *permissible*, rather than following whichever *theory* one thinks is most likely to be *true*.

My Favourite Option (MFO)

An option is appropriate for some agent *iff* it is one of the options this agent thinks is most likely to be permissible.

[3] We will discuss this property in more detail when we discuss the theory it most famously affects: Maximise Expected Choiceworthiness.

[4] In responding to this objection, Gustafsson and Torpman (2014) propose an explicit account of moral theory individuation, though it suffers problems, as discussed by MacAskill and Ord (2020).

In *Kira's Dinner*, this approach recommends that Kira should choose linguine with clams, as she has a combined credence of 50% that this is permissible (30% from her credence in *Vertebrate Welfare* and 20% from her credence in *No Pork*), which is higher than for any other option.

Much like MFT, this approach has a strong intuitive appeal. And MFO does not require a robust account of theory-individuation, resisting the objection raised above. In the previous example, if the utilitarian variants recommended the same option, then it wouldn't matter to MFO whether they comprise a single theory with 90% credence or ten theories adding up to 90% credence.

But MFO is vulnerable to a different objection — that of being insensitive to stakes. This objection hinges on an intuition that the relative stakes of options, according to different moral theories, should play a role in determining their appropriateness ordering. Returning to *Kira's Dinner*, we can elicit the intuition by considering the problem from the perspective of each theory in turn. From the perspective of the human-centred utilitarian view, the differences between options are plausibly mild: reflecting things like the difference in pleasure Kira will have while eating each meal, or the (presumably modest) longer-run effects of this single meal on her welfare. But from the perspectives of each of the other theories, the stakes appear considerably higher. For the two animal-welfarist views, certain of Kira's options implicate the mistreatment and slaughter of moral patients. For the deontological view, one option involves violating an absolute prohibition. The intuition here is that these three views should have 'more say' in this particular decision, because this is a decision they care deeply about. MFO offers no means of capturing the intuition, which seems to be a failing.[5]

2.2.3 Maximise Expected Choice-worthiness

Taking its lead from Expected Utility Theory, the approach known as Maximise Expected Choice-worthiness (MEC) explicitly incorporates a sensitivity to stakes, as well as likelihoods. MEC's basic recommendation is as follows: insofar as this is possible, you should impute a choice-worthiness function to each moral theory in which you have any credence. Given the details of your decision situation, this function assigns to each option a number representing its degree of choice-worthiness. You then take an expectation over the choice-worthiness values of the different options (weighting each theory's choice-worthiness assignments by your credence in that theory), and select an option with the highest expected choice-worthiness.

Maximise Expected Choice-worthiness (MEC)

An option is appropriate for some agent *iff* it is one of the options that has maximal expected choice-worthiness (where this can be determined).

As written, *Kira's Dinner* does not offer enough structure for MEC to reach a verdict. But it seems plausible that the approach would recommend she chooses the lentil curry — against the recommendations of both MFT and MFO. Recall that Kira's decision intuitively has far higher stakes from the perspective of some of the moral theories. In particular, it is plausible that the difference in choice-worthiness between options containing pork and options not containing pork would be vast for the deontological theory. Similarly, it is plausible that the difference between options which implicate the slaughter of moral patients and options that do not would be high for the animal-welfarist theories. Lentil curry is the only option that falls on the right side of these divisions.

[5] It's noteworthy that MFT is also stakes-insensitive in this way.

This approach sacrifices some of MFT and MFO's ease of application, but also resists each of the objections raised so far: it does not depend on how theories are individuated and is explicitly sensitive to stakes. The approach also draws strength from its analogy with Expected Utility Theory, which reigns largely unchallenged as the correct approach to take in cases of empirical uncertainty. In fact, of all extant approaches to moral uncertainty, MEC has received the most sustained and recent support (including a book-length defence in MacAskill et al. (2020)). It may therefore be regarded as the central alternative to which the Moral Parliament should be compared — although, as we shall see, the question of which approaches count as genuine alternatives is itself non-trivial.

MEC faces at least two important objections: fanaticism and the problem of intertheoretic comparisons. The first of these can be viewed as a charge of over-sensitivity, mirroring the charge of insensitivity lodged against MFO and MFT. The issue is that MEC appears overly sensitive to theories that posit extremely high stakes, even when one's credence in these theories is very small.[6] For example, consider a theory which held that killing any animal, including the smallest of insects, was as bad as killing a human (even if the killing is unintended). This could come to dominate your moral decision-making despite you having extremely little credence in it. Here, the problem is that MEC appears to give too much weight to the stakes of a theory, relative to one's credence in that theory, with the result that views with 'high-stakes and low-credences' appear advantaged in a way that violates intuitions.

The problem of intertheoretic comparisons is that there is no widely accepted method for comparing choice-worthiness across moral theories. For example, it is *prima facie* unclear how to compare how wrong it is to eat pork on a deontological theory, where this is absolutely forbidden, compared to how wrong it is to eat the alternatives under a utilitarian theory, where this is grounded in weighing up the welfare gains and losses. Moreover, the question of whether intertheoretic comparisons are even possible in principle remains contested.

One attempted solution to this problem involves normalising theories against each other. For example, you might suggest that the most choice-worthy options according to each theory should count as having equal choice-worthiness, and similarly for the least choice-worthy. The aim would be to capture a notion of 'equal say': the idea that the choice-worthiness assessments of different moral theories should be placed on equal footing. Cotton-Barratt et al (2020) investigate the approach and conclude that normalising by *variance* represents the most promising path forward. However, normalisation methods invite problems of their own. In particular, they can conflict with decision-theoretic principles[7], or else render MEC even more difficult to apply.

3 Moral Parliament

Imagine that each moral theory in which you have credence got to send delegates to an internal parliament, where the number of delegates representing each theory was proportional to your credence in that theory.[8] Now imagine that these delegates negotiate with each other, advocating

[6] There is also an acute form of this objection: the problem of infectious incomparability. In place of 'extremely high stakes', some theories posit widespread *incomparability of options*. On one understanding, such theories break MEC by leading to the result that all options are incomparable in all decision-situations (even when one has virtually no credence in one of these infectious theories). See MacAskill (2013) and Ross (2006).

[7] For example: Independence of Irrelevant Alternatives; Contraction Consistency.

[8] In the simplest case, this would mean one percentage point of credence corresponds to one delegate, for a 'total parliament size' of 100. This strains the analogy when one's credences involve fractions of a percent, since it's not clear what a fraction of a delegate would look like. If all of one's credences were rational numbers, it would always be

on behalf of their respective moral theories, until eventually the parliament reaches a decision by the delegates voting on the available options. This would provide a novel approach to decision-making under moral uncertainty that may avoid some of the problems that beset the others, and it may even provide a new framework for thinking about moral uncertainty more broadly.

Moral Parliament

An option is appropriate for some agent iff it is one of the options endorsed by that agent's Moral Parliament.

This central analogy is more than skin-deep: there is a genuine sense in which the problems that real-world parliaments exist to address resemble the problem of moral uncertainty. In the same way that existing parliaments serve to balance competing interests and to reconcile differences in deeply-held values, the Moral Parliament is intended to function as a value-neutral crucible in which compromise, and consensus, might be forged. Similarly, as we shall see, many of the problems that arise in political theory give rise to analogues in the moral case.

At first blush, the proposal above leaves much unspecified: on what basis are motions to be set, voting methods determined, or term-limits decided? In practice, however, many of these questions can be set to one side. For all of the approaches to moral uncertainty raised so far — MFT, MFO, and MEC — draw on the framework of decision theory and thus make a number of simplifying assumptions. Specifically, these approaches take decision-situations — replete with option-sets, credence distributions, and so on — as given.[9] By making the same assumption here, we can sidestep questions around how the Moral Parliament is established and, to some extent, when 'moral elections are held'. It is as if the Parliament manifests mid-session when a decision-situation involving moral uncertainty arises.

Even with these simplifications in hand, however, the Moral Parliament remains underspecified: one can imagine variants using different voting methods or with different rules for how motions are to be set. Below, we outline several potential specifications of the approach and suggest a tentatively preferred option. More specifically, we suggest using 'proportional chances voting' (described below) and that motions should be set in accordance with common-sense (in combination with decision-theoretic norms), rather than aspiring to include every possible motion. This falls short of a comprehensive formal specification, but offers enough substance to consider the advantages and disadvantages of the approach.

Real-world parliaments and electoral systems use a variety of voting methods. Among the most common is plurality voting, the method according to which each voter may vote only once on a given issue and the option with the most overall votes wins. Other methods include instant run-off voting, where voters provide ordinal rankings of the options, and approval voting, where each voter may vote to 'approve' any number of options. The literature on voting theory is extensive and covers the varied strengths and weaknesses of these methods, as well as many others. There are famously a number of impossibility theorems, showing that all voting methods suffer from at least one important theoretical defect, and there is no consensus on which method is best in practice. There are thus a variety of plausible ways the Moral Parliament's voting might be specified. In this sense, the Moral Parliament can be viewed as a family of possible solutions to the problem of moral uncertainty, rather than a single well-defined proposal.

possible to 'multiply through', potentially making for a much larger parliament but avoiding the problem of 'partial people'.

[9] In fact, this is assumed in how the problem of moral uncertainty is defined.

That said, for concreteness it is useful to examine what would happen under a more concrete proposal. We suggest that 'proportional chances voting'[10] is a strong contender for how its voting method might best be specified. Under proportional chances voting, each delegate receives a single vote on each motion. Before they vote, there is a period during 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. The delegates then cast their ballots for one particular option in each motion, just as they might in a plurality voting system. But rather than determining the winning option to be the one with the most votes, each option is given a chance of winning proportional to its share of the votes.

One advantage of this voting procedure is that it resists the 'tyranny of the majority'. This is a classic problem in political theory in which the interests of minorities are systemically underrepresented in the decisions of the overall polity as a result of electoral systems that privilege the majority. Consider, for example, plurality voting: in cases where one group achieves more than 50% of the voting power, this will amount to a dictatorship by majority. Proportional chances voting avoids this problem because it allows for *any* option, no matter how small a minority supports it, to have some chance of being selected. Moreover, it makes this allowance in an intuitive way by having the chance of each option be directly proportional to that option's degree of support.

A second advantage of this system is that it creates a strong incentive for delegates to find good compromise options, where a 'good compromise' is one in which all parties are almost as happy as if they got their own way entirely. To see this, note that one implication of proportional chances voting is that every option which receives *any votes whatsoever* can have a significant impact on the expected value of a given vote. This means that all parties have an incentive to bargain with even small minorities to see if there is a more mutually satisfactory option that they can agree upon. Thus all parties — even those with strong majorities — are incentivised to offer meaningful concessions, such that the final lottery contains no ballots they would strongly want to avoid being drawn.

It might seem that this voting method creates a new problem: in some cases, the lottery will play out in such a way that you end up selecting an option with very little support. This would certainly be true if it were implemented as a practical voting system, and its theoretical virtues when considered *ex ante* may seem less compelling *ex post* if and when an almost universally shunned option were to end up winning.

However, this is where we can make use of features of the Moral Parliament not present in real-world parliaments to finesse the method. Specifically, we can stipulate that the delegate negotiation, and voting, take place *as if* the final decision will be made by a proportional lottery, but then actually decide the outcome via plurality voting. This would preserve the incentive for compromise and continue to resist tyranny of the majority, while avoiding the possibility suggested above. In real-

[10] There is a related voting procedure in the literature known as 'Random Dictator', where a lottery is held among the voters for which one gets to determine the outcome according to their own preferences. This is mathematically similar to determining the outcome with proportional chances based on the number of votes for each option (as one way to implement it is to mix all the votes and pick one randomly), but the name 'Random Dictator' strongly implies that one voter gets their way entirely. In contrast, we are imagining that there is a round of bargaining first, such that the vote may have been cast for a compromise option — chosen based on an appreciation of the preferences and power of the other delegates. This is not something that is well described by the word 'dictator'.

world parliaments, of course, such a deception would be obvious to the delegates (at least over time). But in the case of Moral Parliament, we are free to stipulate their ignorance in this way.[11]

Along with voting methods, real-world parliaments differ in the rules by which motions are set. In the case of the Moral Parliament, many of the associated practicalities can be set to one side — we need hardly specify which dates the Moral Parliament will sit. However, there is an important question that does merit attention: at what scale of decision should the Moral Parliament be applied? For example, one might apply it to the decision about what to have for a single meal, as in *Kira's Dinner*. Then again, one might apply it to the decision about what to have for dinners in general, or even to an omnibus decision about all possible options one might ever face (culinary or otherwise). Broader decisions will offer more room for compromise to be struck, and risk less inconsistency across the scope of all one's actions, but they will heighten the difficulty in practically determining what outcome the moral parliament would have selected.

As noted above, one approach here would be to sidestep the question. In the same way that existing approaches to moral uncertainty help themselves to the conventions of decision theory and do not concern themselves with the question of 'how to identify appropriate decision-situations', we might simply consider the issue of scale to be out of scope. Alternatively, a principled basis on which the scale of decisions might be set is the ultimate version of the omnibus proposal sketched above: the thought that Moral Parliament need only be convened once, where it aims to consider and navigate all possible options over one's life-course. This suggestion has at least one advantage, which is that it avoids a charge of arbitrariness that might be levelled against any approach where Moral Parliament convenes more regularly. At the same time, it has notable disadvantages — among them the severe practical difficulties involved with thinking through all one's future decisions simultaneously.

Here, we endorse a common-sense approach to the question of scale which has much in common with standard decision-theoretic conventions. The suggestion is that one should convene Moral Parliament for those decision-situations to which it is intuitively appropriate, such as those involving non-trivial moral stakes, where the possible options are relatively well-defined, and so on. Normatively speaking, if Moral Parliament is the right approach to take to moral uncertainty, then it may also be right to apply it to all decision-situations (however this is defined). But practically speaking, this would be very difficult to achieve. This move has essentially the same implications as the approach of sidestepping the question but comes with a positive endorsement of Moral Parliament's application to 'the kinds of decision-situations typically described in papers on moral uncertainty'. This is the sense in which the common-sense approach resembles standard decision-theoretic conventions.

How, then, would the Moral Parliament, as specified above, navigate *Kira's Dinner*? As with MEC, the letter of the example does not offer enough structure to reach a formal verdict, but we can make a reasonable assessment nonetheless.

Given that all of Kira's credences are round numbers, we are free to imagine a hundred-member parliament comprising 40, 30, 10, and 20 delegates for each of *Human Welfare*, *Vertebrate Welfare*, *Animal Welfare*, and *No Pork*, respectively. In addition, there are several plausible assumptions we can make about the positions these delegates would take. For example, delegates from the last three views would likely be united in their stern opposition to pork pie, while only the delegates

[11] We could similarly sidestep some other problems of voting systems, such as the fact that most voting systems provide incentives for voters to misrepresent their preferences, as described by the Gibbard-Satterthwaite Theorem due to Gibbard (1973) and Satterthwaite (1975).

from *Animal Welfare* would have a similarly strong objection to linguine with clams. Meanwhile, no delegates at all would strongly object to lentil curry — though it would not be the first choice for many of them.

Given the proportional lottery system, the delegates from *Animal Welfare* would have a clear incentive to negotiate others away from voting for linguine with clams, while the delegates from all views except *Human Welfare* would have a similar incentive to negotiate against votes for pork pie. Overall, then, it seems most plausible Kira's Moral Parliament would recommend lentil curry on this occasion — in agreement with MEC.

3.1 In Favour of Moral Parliament

One reason for considering Moral Parliament a promising response to moral uncertainty is simply that it resists each of the objections facing other approaches.

Unlike MFT, Moral Parliament does not depend crucially on an account of theory-individuation; delegates can vote as 'individuals'. Thus we don't need to worry about whether there are 90 delegates representing utilitarianism or 9 for each of ten different varieties of utilitarianism that all have the same views on this choice.

Unlike MFO, it is somewhat sensitive to the stakes that moral theories accord different options. For example, it doesn't treat all impermissible options equally and may be prepared to compromise to secure an option all theories believe is the lesser evil. Moreover, if multiple decisions are allowed to be simultaneously brought to the parliament, then delegates are free to spend as much or as little of their negotiating capital on any given decision depending on how they see its stakes.

In addition, the Moral Parliament is arguably less vulnerable than MEC to both the problems of intertheoretic comparisons and fanaticism. The Moral Parliament does not make comparisons of choice-worthiness across moral theories. Instead, each theory's power is set solely by its credence, which determines its number of delegates. And it is up to the theory how it uses that to push towards preferred options or away from strongly dispreferred ones. But it cannot get more sway overall simply by treating more situations as having high stakes, thus avoiding the problem of fanaticism. Moreover, the Moral Parliament creates incentives for compromise that discourage voting for extreme options.

We can also expand the manner in which Moral Parliament resists fanaticism into a positive argument in its favour, which is that it appears to capture our intuitions about compromise more closely than does MEC.

Consider an agent with non-negligible credence in only two moral theories, T_1 and T_2, facing a decision-situation with only two options, A and B. T_1 considers option A to be highly choice-worthy and option B to be barely conscionable. By contrast, T_2 holds option A to be barely conscionable and option B to be highly choice-worthy. This situation is represented graphically below.

Here, the vertical and horizontal axes show the degrees of choice-worthiness according to theories T_1 and T_2, respectively, and options are shown as labelled points.

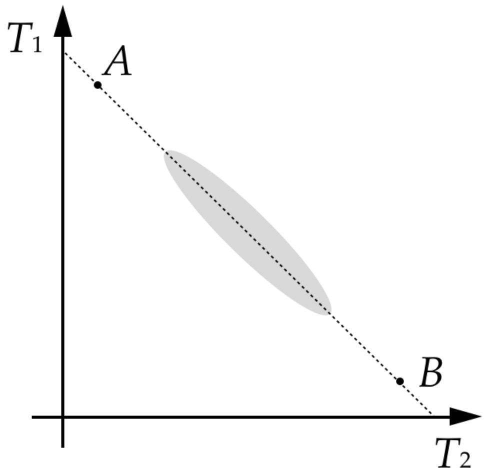

Figure 1: Illustration of choice-worthiness of options A and B according to theories T_1 and T_2.

In this case, there is a strong intuition that neither option will do: they are each considered barely conscionable by one of the theories in which the agent holds credence. Indeed, we might think that if there were any options in the shaded region, these would be more appropriate than either *A* or *B*, since these represent compromises that are considered at least somewhat choice-worthy by both theories. However, MEC has trouble recommending options in this region.

We can see this by the use of indifference curves, joining up locations where options would be considered equally appropriate. These are like contour lines of appropriateness. For MEC these indifference curves always take the form of straight lines, such as the dotted line shown. The slope of the lines is determined by both the relative credence in each theory and the way the intertheoretic comparisons of choice-worthiness are made.

Note that there is no possible combination of credence distribution and intertheoretic comparison between T_1 and T_2 for which MEC recommends points inside the shaded region and below the straight line shown. Further, there is only a narrow range of credence distributions and intertheoretic comparisons for which MEC can recommend points inside the shaded region and on or above the line. We can test this by considering straight lines of different slopes that lie outside the set of options and imagining slowly moving them towards the bottom left the diagram until they touch an option. Whichever option a line contacts first is its recommendation for an agent with that credence distribution and intertheoretic comparison. This makes it clear that only a narrow range of slopes for these indifference curves recommend options in the shaded region. In other words, MEC tends to recommend extreme options like *A* and *B* and struggles to converge on compromise options.[12] Moral Parliament, on the other hand, has the tools to identify and hone in on options of this sort.

In fact, this example demonstrates more than just that MEC struggles to find compromise options — it demonstrates that at least some versions of Moral Parliament are formally distinct from any version of MEC. Namely, any specifications of a Moral Parliament which permit the recommendation of options in the shaded region below the line or which facilitate the recommendation of options in the shaded region above the line.

One final consideration in favour of Moral Parliament is that it offers relatively clear conceptual handles and may consequently be easier to put into practice than MEC. This is a different kind of

[12] We thank Owen Cotton-Barratt for bringing this point to our attention.

argument to those raised previously: it is an argument in favour of Moral Parliament as a decision procedure rather than a criterion of rightness. Section 2.2 made the point that each of MFT and MFO are admirably easy to apply, whereas MEC sacrifices some of this practicality in navigating the various objections. In fact, these solutions represent successive increases in complexity. MFT is especially straightforward: you need only know which theory you have most credence in and what that theory recommends. MFO introduces greater complexity, in that you need now consider the recommendations of multiple theories and perform an aggregation. Finally, MEC adds the complications that you must assess the strength of recommendations (as well as the direction), compare the strengths of recommendations across theories, and then take an expectation. These increases in complexity come with costs. In particular, from a practical point of view, increasing a solution's complexity in this way diminishes its usefulness as a decision procedure. While Moral Parliament is certainly trickier to apply than MFT or MFO, it is perhaps more practically workable than MEC. It seems plausible that, in practice, trying to abide by the decisions of one's personal Moral Parliament would lead to decisions that more closely approximate 'ideal appropriateness' than would attempting to make calculations of expected choice-worthiness, even should the latter prove stronger as a criterion of rightness.

3.2 Against Moral Parliament

A major theoretical problem with varieties of Moral Parliament is that they can issue intransitive judgments of appropriateness across choice situations. This is something they inherit from voting theory due to an analogue of the famous Condorcet Paradox:

	1/3	1/3	1/3
	T_1	T_2	T_3
A	Best	Worst	Second
B	Second	Best	Worst
C	Worst	Second	Best

Table 2: Illustration of intransitivity

If a single decision is being made between options *A* and *B*, *A* is preferred by two thirds of delegates and would be deemed appropriate. If it is between *B* and *C*, *B* is preferred by two thirds and would be deemed appropriate, and similarly in a choice between *C* and *A*, *C* would be deemed appropriate. This kind of intransitivity across choice situations is a major theoretical defect (that is also suffered by MFO).

It also creates a kind of dilemma for how one sets up the Moral Parliament, where limiting the damage of this intransitivity raises other problems. As we saw earlier, there are variants of Moral Parliament that consider only one individual narrowly-construed choice at a time, all the way up to those that consider all possible choices over one's life simultaneously.

Versions that break things up into narrower decisions and treat them separately can lead you to make a sequence of decisions that is dominated by a different sequence of decisions you could have made instead. This comes from the fact that each of these decisions are being considered separately, so the apparatus of the Parliament can't take account of whether it is a higher-stakes decision for one theory than for another — their number of delegates are determined by their credence alone. This property is what allows one to avoid fanaticism, but it doesn't give you the tools to see whether the delegates of a moral theory should concede in an earlier decision (that they see as very low stakes) in exchange for getting their way in a later decision (that they see as much higher stakes), even if every theory would agree that this combination of decisions was best.

This problem is reduced as more and more decisions are bundled together, eventually disappearing entirely if everything is dealt with in one omnibus decision. But then Moral Parliament faces the other horn of the dilemma: an intractably complex decision that may be theoretically unimpeachable but offers little or no practical guidance.[13]

It is issues like these, of transitivity and avoiding dominated sequences of choices, that push one towards maximising expected utility (under empirical uncertainty) and maximising expected choiceworthiness (under moral uncertainty).

This intransitivity (and resulting dilemma) that afflicts Moral Parliament is closely linked to its avoidance of fanaticism. Indeed, it is almost a necessary price for any approach to moral uncertainty that avoids fanaticism. As Beckstead (2013) has shown (in the context of empirical uncertainty), all approaches to uncertainty are either problematically reckless (corresponding to fanaticism), problematically timid, or violate transitivity across choice situations.[14]

There is another, quite different, objection to Moral Parliament arising from its analogy to political theory. In the case of MEC, the analogy with Expected Utility Theory functions as a strength in part because Expected Utility Theory is widely accepted as the canonical response to empirical uncertainty. But in the case of politics, things are far less settled. Real-world parliaments vary in almost all respects[15] and there is hardly an 'accepted solution' that Moral Parliament can import into the case of moral uncertainty.

This suggests that the approach faces a considerable challenge: arriving at an acceptable and maximally specified account may be roughly equivalent to solving the problem of ideal governance. As noted, however, this is a challenge to be overcome and not a knock-out blow. Plato, at least, would hardly be surprised if the problems of achieving justice in the soul and the state prove approximately equivalent.

4 Further Discussion

The discussion above leaves a number of loose threads. This section tugs briefly on three of these, as a gesture towards those areas of the literature on moral uncertainty in general, and Moral Parliament in particular, that might provide fertile ground for further exploration. These threads have to do with (1) systematising the solutions to moral uncertainty and demonstrating equivalence results, (2) relaxing the assumptions imported from decision theory, and (3) identifying common desiderata for acceptable solutions.

Section 3 noted that Moral Parliament can be viewed as a family of possible solutions, rather than a single well-specified proposal. Section 3.1 then pointed out that the proposed solutions to moral uncertainty fall on a 'spectrum of complexity'. Building on these two observations, it seems

[13] This dilemma is closely related to that between narrow and broad normalisation techniques for MEC, as explained in Cotton-Barratt et al (2020).

[14] If fanaticism is to be avoided, it may be useful to explore moral uncertainty analogues of timid approaches.

[15] To say nothing of political systems that are not parliamentary in the first place.

plausible that all of the proposed solutions might be brought under a common theoretical framework and that Moral Parliament represents one promising way of doing so.

Each of the proposed solutions in Section 2 can be recast in parliamentary terms.[16] MFT can be seen as a system where the party with the most delegates is granted absolute power over decisions. This would be a kind of 'tyranny of the plurality'. Note that MFT's requirement that we can individuate theories has generated a requirement that we can group delegates into parties. MFO would correspond to a form of approval voting, with the stipulation that delegates are not allowed to vote strategically nor to make deals. Finally, MEC would correspond to a system of range voting — the method in which voters give each option a score and the option with the highest total score wins (again with no strategic voting or negotiation allowed). Different ways of handling intertheoretic comparisons on MEC would correspond to different constraints on the ranges, such as honest reporting of absolute choice-worthiness ratings on some common scale or always setting the worst option to 0 and the best to 1.

This exercise could be read as problematic, insofar as a demonstration of general equivalence might imply that Moral Parliament adds little of substance to the debate. At the same time, systematising proposed solutions in this way facilitates comparison, offers new angles for evaluation (drawing on results from political theory)[17], and may afford other advantages too.[18] It remains to be shown which precise specifications of Moral Parliament are formally equivalent to which proposed solutions, and whether parliamentary approaches are constrained in important ways or else can accommodate all plausible solutions.

Section 3 also noted that all the proposed solutions, including Moral Parliament, help themselves to simplifying assumptions from decision theory. One avenue for further exploration of the problem of moral uncertainty would be to consider in greater detail how to relax these assumptions. This is especially relevant when considering the proposed solutions as decision procedures, since the real world hardly comes divided into neat decision-situations. Instead, a solution to moral uncertainty that aims to be practically serviceable would need to answer questions about the circumstances in which this kind of decision-making should be entered into, the ways in which options are to be defined, and the conditions in which decisions may be overturned. In the parliamentary case, these are questions about when to convene Moral Parliament, how to set motions, and when to repeal policies. There is considerable practical philosophy to be accomplished in this space.

5 Conclusion

This paper introduced the problem of moral uncertainty, presented a novel way of approaching the problem, and situated this approach within the existing literature.

Section 2 provided a statement of the problem and introduced three proposed solutions. We found that each of these approaches — *My Favourite Theory*, *My Favourite Option*, and *Maximise Expected*

[16] This is closely related to the analogy between moral uncertainty and social choice theory, introduced by MacAskill (2010) and further developed in MacAskill et al (2020).

[17] For example, 'dictatorship by plurality' hardly seems like a desirable solution.

[18] For example, MEC might benefit from the 'clearer conceptual handles' of a parliamentary approach, when recast in these terms.

Choice-worthiness — face notable objections. In the case of MFT and MFO, the objections appear fatal. In the case of MEC, there is an ongoing discussion of how they might be overcome.

Section 3 then presented the idea of Moral Parliament, offered several arguments in its favour, and raised a new objection. Here, we saw that Moral Parliament resists each of the objections raised in Section 2, that it appears to capture intuitions about compromise better than MEC, and that it potentially shows greater promise as a decision procedure as well. We also noted that its recommendations can be intransitive across choice situations, and that it appears vulnerable to a charge of inherited insolubility: political theory has yet to reach consensus on what constitutes the ideal parliament, and Moral Parliament may face an equivalent struggle.

Finally, in Section 4, we sketched three directions in which this thinking could be further developed, having to do with systematising approaches to moral uncertainty, relaxing assumptions imported from decision theory, and working towards a complete set of desiderata for proposed solutions.

References

Beckstead, N., *On the Overwhelming Importance of Shaping the Far Future."* PhD Thesis. Department of Philosophy, Rutgers University (2013).

Bostrom, N., "Moral Uncertainty: Towards a Solution?", https://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html (2009).

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

Bykvist, K., "Moral Uncertainty," *Philosophy Compass* Vol. 12, Issue 3: (2017).

Cotton-Barratt, O., MacAskill, W., Ord, T., "Statistical Normalization Methods in Interpersonal and Intertheoretic Comparisons," *The Journal of Philosophy* Vol. 117, Issue 2: (2020).

Gibbard, A., "Manipulation of voting schemes: a general result." *Econometrica* Vol. 41: (1973).

Gracely, E., "On the Noncomparability of Judgments Made by Different Ethical Theories," *Metaphilosophy* Vol. 27, No. 3: (1996).

Greaves, H., Cotton-Barratt, O., "A bargaining-theoretic approach to moral uncertainty," GPI working paper No. 4: (2019).

Gustafsson, J., Torpman, O., "In Defence of My Favourite Theory," *Pacific Philosophical Quarterly* Vol. 95, No. 2: (2014).

Harman, E., "The Irrelevance of Moral Uncertainty," *Oxford Studies in Metaethics* Vol. 10: (2015).

Lockhart, T., "Another Moral Standard," *Mind* Vol. 86, No. 344: (1977).

MacAskill, W., "How to Act Appropriately in the Face of Moral Uncertainty". BPhil thesis, University of Oxford: (2010).

MacAskill, W., "The Infectiousness of Nihilism," *Ethics* Vol. 123, No. 3: (2013).

MacAskill, W. and Ord, T., "Why Maximize Expected Choice-worthiness," *Noûs* Vol. 54, No. 2: (2020).

MacAskill, W., Bykvist, K., and Ord, T., *Moral Uncertainty*: Oxford University Press: (2020).

Medina, B., *Expositio in primam secundae angelici doctoris D. Thomae Aquinatis*: (1577).

Pascal, B., *Lettres Provinciales*: (1657).

Plato, Adam, J. trans., *The Republic of Plato*: Cambridge University Press, (2009).

Ross, J., "Rejecting Ethical Deflationism," *Ethics* Vol. 116, No. 4: (2006).

Satterthwaite, M., "Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions." *Journal of Economic Theory* Vol. 10: (1975).