A bargaining-theoretic approach to moral uncertainty

Owen Cotton-Barratt (Future of Humanity Institute, University of Oxford), Hilary Greaves (Global Priorities Institute, University of Oxford)

Global Priorities Institute | February 2023

GPI Working Paper No . 2-2023

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

A bargaining-theoretic approach to moral uncertainty

Hilary Greaves and Owen Cotton-Barratt

16 December, 2022

Abstract

Nick Bostrom and others have suggested treating decision-making under moral uncertainty as analogous to parliamentary decision-making. The core suggestion of this 'parliamentary approach' is that the competing moral theories function like (groups of) delegates to the parliament, and that these delegates then make decisions by some combination of bargaining and voting. There seems some reason to hope that such an approach might avoid standard objections to existing approaches (for example, the 'maximise expected choiceworthiness' (MEC) and 'my favourite theory' approaches). However, the parliamentary approach is so far extremely underspecified, making it largely indeterminate how such a model will in fact behave in the respects that those concerned with moral uncertainty care about.

This paper explores one way of making it precise. We treat predicaments of moral uncertainty as analogous to bargaining situations alone (setting aside voting), and apply a version of the Nash solution that is standard in bargaining theory. The resulting model does indeed perform in many of the hoped-for ways. However, so also does a version of MEC that employs a structural approach to intertheoretic comparisons. It seems to us an open question which, of this version of MEC and the bargaining-theoretic approach, is superior to the other. We identify the key points on which the two differ.

1 The problem of moral uncertainty

We often have to act under conditions of relevant uncertainty. Sometimes the uncertainty in question is purely empirical. When one decides whether or not to take the umbrella, for instance, one is uncertain whether or not it will rain. Each action one might choose is a gamble: the outcome of one's action depends,

in ways that affect how highly one values the outcome, on factors of which one is ignorant and over which one has no control.

Suppose Alice takes an umbrella but, as the day turns out, it does not rain. Does it follow that Alice made the wrong decision? In one (objective) sense of "wrong", yes: thanks to that decision, she experienced the mild but unnecessary inconvenience of carrying a bulky object around all day. But in a second (more subjective) sense, clearly it need not follow that the decision was wrong: if the probability of rain was sufficiently high and Alice sufficiently dislikes getting wet, her decision could easily be the appropriate one to make given her state of ignorance about how the weather would in fact turn out. Normative theories of decision-making under uncertainty aim to capture this second, more subjective, type of evaluation. The standard such account is expected utility theory.

We also have to act under conditions of relevant *moral* uncertainty. When one decides whether or not to eat meat, for instance, one is (or should be) uncertain whether or not eating meat is morally permissible. How should one choose, when facing relevant moral uncertainty? In one (objective) sense, of course, what one should do is simply what the true moral hypothesis says one should do. But it seems there is also a second sense of "should", analogous to the subjective "should" for empirical uncertainty, capturing the sense in which it is appropriate for the agent facing moral uncertainty to be guided by her moral credences, whatever the moral facts may be.

This way of setting out the issues hints that there is a close analogy between the cases of moral and empirical uncertainty, so that those who recognise a subjective reading of "ought" in the context of empirical uncertainty should also recognise a nontrivial question of appropriate action under moral uncertainty.[1] There is a lively debate about whether this analogy is valid.[2] There is also debate about what precisely kind of "should" is involved: rational, moral, or something else again.[3]

[1] Not everyone does recognise a subjective reading of the moral 'ought', even in the case of empirical uncertainty. One can distinguish between objectivist, (rational-)credence-relative and pluralist views on this matter. According to objectivists (Moore, 1903; Moore, 1912; Ross, 1930, p.32; Thomson, 1986, esp. pp. 177-9; Graham, 2010; Bykvist and Olson, 2011) (respectively, credence-relativists (Prichard, 1933; Ross, 1939; Howard-Snyder, 2005; Zimmermann, 2006; Zimmerman, 2009; Mason, 2013), the "ought" of morality is uniquely an objective (respectively, a credence-relative) one. According to pluralists, "ought" is ambiguous between these two readings (Russell, 1966; Gibbard, 2005; Parfit, 2011; Portmore, 2011; Dorsey, 2012; Olsen, 2017), or varies between the two readings according to context (Kolodny and Macfarlane, 2010).

[2] The view that while some form of subjectivism about empirical uncertainty is perhaps correct, objectivism the (uniquely) correct view about moral uncertainty, is defended by (Harman, 2011; Weatherson, 2014; Hedden, 2016; Weatherson, 2019). For replies, see (Sepielli, 2016; Bykvist, 2017; Sepielli, 2017; MacAskill and Ord, 2018).

[3] The view that it is a rational "should" is defended by e.g. Bykvist (2014; 2018). On the other hand, one might well worry that it is consistent with *rationality* to altogether ignore all moral considerations, but that there is nonetheless a sense in which even such an 'amoralist' ought to play it safe in suitable contexts of moral uncertainty. This consideration mitigates

For the purpose of this article, we will simply take for granted that there is a nontrivially credence-relative sense of "should" in the moral case. We will also not take a stand on what kind of "should" it is. Our question is how the "should" in question behaves in purely extensional terms. An answer to that question is a *metanormative theory*.

There are various existing proposed metanormative theories, but none commands widespread assent. The purpose of the present paper is to articulate and evaluate a new approach, based on bargaining theory.

The structure of the paper is as follows. Section 2 briefly surveys the main extant theories of moral uncertainty that we will use as standards for comparison, viz. the "maximise expected choiceworthiness" (MEC) and "my favourite theory" (MFT) approaches. Section 3 introduces the idea of treating decision-making under moral uncertainty as analogous to parliamentary decision-making. We note that many have hoped that such an approach might simultaneously avoid the standard objections to (inter alia) MEC and MFT, but that so far, the discussion of this idea has remained at a superficial level with no precise models, making it difficult to assess the hopes in question. In particular, a parliamentary approach is supposed to involve both bargaining and voting, but there has been no previous rigorous exploration of what modelling decision-making under uncertainty as formally analogous to a bargaining problem would look like.

Section 4 introduces bargaining theory, and sets out a bargaining-theoretic approach to moral uncertainty. Section 5 establishes some general results that will prove illuminating, for the purpose of understanding and evaluating the way in which the bargaining-theoretic approach treats the problem of moral uncertainty. In sections 6–10, we use these results to analyse the performance of this approach vis-a-vis (respectively) issues of sensitivity to relative stakes (section 6), compromise between moral theories (section 7), fanaticism (section 8), dependence of results on the presence 'irrelevant' alternatives (section 9) and a "problem of small worlds" (section 10). Section 11 summarises, and compares the merits of a bargaining-theoretic approach with those of MEC. We find that the bargaining-theoretic approach is similar on many of the dimensions of interest to one of the extant versions of MEC (specifically, MEC with a "structural" approach to intertheoretic comparisons). There are also, however, potentially important points of difference. The issues seem rather subtle, and we do not attempt any overall conclusion about which of the two approaches in question is in the end superior to the other. We do conclude, though, that ideas in the vicinity of 'the parliamentary approach' seem promising enough to warrant further exploration.

against the view that the "should" in question is a rational rather than a moral one. For a survey of the issues, see (Bykvist, 2017, section 2).

2 Existing theories of moral uncertainty

The most popular metanormative theory holds that one should maximise expected choiceworthiness (MEC). MEC treats moral uncertainty exactly as expected utility theory treats empirical uncertainty. That is, it holds that an agent facing moral uncertainty ought to be such that for some probability function p on (products of) moral hypotheses and states of nature, and some choiceworthiness function that assigns numerical values to pairs of moral theories and options, the agent weakly prefers A to B iff the expected choiceworthiness of A, with respect to p, is at least as high as that of B.

MEC is, however, subject to various criticisms. For our purposes, the most important criticisms are that it is well-defined only if intertheoretic unit comparisons are well-defined, and that it leads to problematic forms of "fanaticism".

Regarding intertheoretic comparisons: specifically, an MEC approach is well-defined only if there are intertheoretic *unit* comparisons. That is, for all options A, B, C, D and all theories T_i, T_j, there must be a fact of the matter as to the ratio between (i) the choiceworthiness difference between A and B according to theory T_i and (ii) the choiceworthiness difference between C and D according to theory T_j. One might worry that there are no such facts of the matter (Hudson, 1987, p.224; Gracely, 1996, p.330; Broome, 2012, p.185; Gustafsson and Torpman, 2014; Nissan-Rozen, 2015; Hedden, 2016). Among those who do not draw this negative conclusion there are, broadly speaking, three positive approaches to intertheoretic comparisons (Riedener, 2021, chapter 4;; Greaves and Ord 2017, section 4). The first is "content-based": where two theories have sufficiently similar content to one another, we might take them to agree on how much is at stake in their areas of overlap, and extrapolate from there using the resources of each theory separately, to determine intertheoretic comparisons in the areas in which the theories disagree. The second approach is "structural": we might equalise some statistical measure of the distribution of choiceworthiness (for example, the range (Lockhart, 2000, pp.84ff.; Sepielli, 2013) or the standard deviation (Cotton-Barratt et al., 2020)) across theories. Thirdly, we might take intertheoretic comparisons to be fixed in some exogenous way, whether by the theories themselves or by the agent who faces the moral uncertainty. This space of options is worth bearing in mind going forwards since, as we will see, there are in fact substantial similarities between a bargaining-theoretic approach and a version of MEC that employs a structural approach to intertheoretic comparisons.

The concern of "fanaticism" is that under the MEC approach, decisions under moral uncertainty can be dictated by a theory in which one has arbitrarily low credence, even if all other theories are unanimous in strongly condemning the decision in question. This can happen, under MEC, if the low-credence theory regards the issue in question as being higher-stakes by a sufficiently large factor

— large enough to trump the ratio of credences (Ross, 2006, pp.765-7; MacAskill and Ord, 2018, section 7(v)).

In response to the perceived drawbacks of MEC, one might consider the "my favourite theory" (MFT) approach. According to MFT, under moral uncertainty one should act in accordance with the moral theory one has highest credence in (Gracely, 1996; Gustafsson and Torpman, 2014).[4]

MFT has its own problems. First, for the purposes of MFT it makes a difference how theories are individuated, whereas this should not make a difference (Gustafsson and Torpman, 2014, section 5; MacAskill and Ord, 2018, pp.8-9). Secondly, since MFT pays no attention to any feature of the agent's decision problem other than her credences, it is insensitive to considerations of relative stakes. If one has (say) 51% credence in a theory according to which A is slightly better than B and 49% credence in a theory according to which A is enormously worse than A, MFT will simply conclude, from the slight difference in credences, that it is appropriate for the agent to choose A. As we will elaborate in section 6, in some examples this stakes-insensitivity appears troubling.[5]

3 A 'parliamentary' approach to moral uncertainty?

Dissatisfaction with extant approaches has motivated a number of commentators to suggest a 'parliamentary approach' to moral uncertainty. The basic idea is to think of the decisions facing a morally uncertain agent on the model of decisions facing a parliament, in which the delegates are representatives of the moral theories in which the agent has positive credence. The number of delegates corresponding to a given theory is proportional to one's credence in the theory in question, so that higher-credence theories have more influence in the parliament. Then "the delegates bargain with one another for support on various issues; and the Parliament reaches a decision by the delegates voting. What [the agent] should do is act according to the decisions of this imaginary Parliament" (Bostrom, 2009).

Advocates of this approach suggest that it might help with many of the problems that plague existing metanormative theories (Bostrom, 2009; Newberry and Ord, 2021). As it is hard to imagine the decisions of a parliament being dictated by a single delegate, on an issue on which that delegate's views

[4]According to the "my favourite option" (MFO) approach, under moral uncertainty one should selection an option that one has highest credence is morally optimal. For our purposes, MFO is a minor variation on MFT.

[5]There is, of course, a tension between a desire to avoid appeal to intertheoretic comparisons and a desire to be sensitive to relative stakes. We return to this tension in section 6.

were strongly opposed by all other delegates, it is natural to hope that an approach to moral uncertainty following a parliamentary model (unlike, perhaps, the MEC approach) will be immune to fanaticism. Since no parliament requires information about how the strength of one delegate's views on the issues under consideration compares, across the board, to the strength of another delegate's views, presumably a parliamentary approach would not face any problem of intertheoretic comparisons, again unlike MEC. Since a parliamentary delegate would, however, tend to bargain hard on issues that she particularly cared about, making concessions in return on issues she regards as less important, one can reasonably hope that (unlike MFT) a parliamentary approach to moral uncertainty will nonetheless exhibit an appropriate degree of sensitivity to relative stakes. And since the number of delegates corresponding to a given point of view is proportional to the agent's credence in that view, not (say) to the number of theories agreeing with that view, there is no danger here of sensitivity to theory individuation (unlike MFT). Because of these sources of hope, the idea of a parliamentary approach enjoys quite widespread support, in particular, among practitioners of effective altruism, where the spectre of fanaticism (especially) is a pressing concern.[6]

To date, however, hope is all that we can have. As the main advocates of a parliamentary approach readily acknowledge, the idea of delegates 'bargaining and then voting' could be made precise in an enormous number of ways. To date, none of these has been articulated in any detail, still less explored. As the devil has a habit of being in the details, we should suspend judgment about whether any parliamentary approach can make good on the above promissory notes until we have examined those details.

The existing literature on moral uncertainty already contains some discussion of treating moral uncertainty on an analogy with voting (MacAskill, 2016; Tarsney, 2019, 2021). The sketches that exist of a 'parliamentary approach' also, however, suggest a second component: not only voting, but also *bargaining*. In fact, bargaining (rather than voting) might be where most of the action happens in the hypothetical moral parliament. For one key idea in 'the parliamentary approach' is that the bargaining stage might often end in widespread (if not completely unanimous) agreement to vote for a particular, specified alternative, in which case the voting stage itself would merely ratify the outcome of the bargaining process.

As with voting, there is a well-developed theory of bargaining (for an overview, see, e.g., Muthoo (1999)). To date, however, there has been no exploration of what happens when one applies the tools of bargaining theory to the problem of moral uncertainty. The present paper makes a start on that exploration.

[6]For example, a survey post on Effective Altruism Forum lists 'the parliamentary model' as one of four accounts of moral uncertainty, alongside normative externalism, MEC and MFT, despite the absence of peer-reviewed academic literature on the former (https://forum.effectivealtruism.org/topics/moral-uncertainty; accessed 14 December 2022).

4 A bargaining-theoretic approach

4.1 Bargaining among persons

In a bargaining problem, two or more 'players' each stand to gain from cooperation (relative to some relevant 'status quo' state of affairs), but there is more than one alternative that is strictly Pareto superior to that status quo.[7] There is therefore an open question precisely which Pareto superior alternative (if any) the players will settle on. Bargaining theory addresses this question.

Our project in this paper is to explore applying the ideas and formal machinery of bargaining theory to the problem of moral uncertainty. For that application, as we will explain in more detail below, we will be taking the 'players' to be moral theories. For the benefit of readers who are unfamiliar with bargaining theory in general, though, we will start by illustrating the workings of bargaining theory in its normal applications, that is, to cases in which the 'players' are *persons*, with differing interests. (Readers who are familiar with bargaining theory can skip the present subsection.)

Suppose, for example, that Kate goes to a car dealership, with a view to buying a second hand car. There is one particular car of interest. Kate would be willing to buy this car at any price no greater than £5000. The dealer would be willing to sell it at any price no less than £3000. There is therefore a wide range of agreements that Kate and the dealer could make — transfer of the car in exchange for any amount between £3000 and £5000 inclusive — that would be Pareto improvements over a scenario in which no transaction occurs. Kate will of course try to push for a lower price, while the dealer will push for a higher price. The open question is which of these Pareto-improving agreements the parties will settle on.

To model this formally, bargaining theory takes an *n-person bargaining problem* to be a structure

$X = (P, U, \mathcal{A}_X, u_X, d_X)$, where

- $P = (p_1, ..., p_n)$ is a set of players.

In our example, n=2, and P = (Kate, the dealer).

- $U = (U_1, ..., U_n)$, where, for each $i = 1, ..., n$, U_i is the space of possible utility levels for player i.

[7] One option is strictly Pareto superior to another if the first is stricly preferred by at least one person, and strictly dispreferred by no-one.

One usually doesn't go far wrong if one simply identifies each U_i with the space of real numbers $\Re$, and indeed many presentations of bargaining theory do proceed in that way. The reason not to officially identify the $U_i$with $\Re$ is that $\Re$ has additional structure beyond that included in the U_i, in ways that are sometimes important. (Crucially for the discussion of moral uncertainty, to officially identify each U_i from the outset with $\Re$ would be to already settle questions of intertheoretic comparisons.)

- $\mathcal{A}_X$ is a set of available options.

In our example, this is the set [£3000, £5000] of possible prices.

- d_X is the disagreement point. This point represents the outcome that each player would get if the players failed to reach agreement.

In our example, the disagreement point corresponds to the scenario in which Kate and the dealer fail to agree on a price, so that Kate remains carless and the dealer remains cashless.

- $u_X = (u_1, ..., u_n)$, where for each i, $u_i : \mathcal{A}_X \cup \{d_X\} \to U_i$ is the von Neumann-Morgenstern (vNM) utility function of player i. We assume that $u_X(\mathcal{A}_X)$ is bounded and convex.

In our example, we can take $\mathcal{A}_X$ to be the set of possible Pareto-improving prices that Kate and the dealer could agree on, that is, the set [£3000, £5000].[8]

In our example, u_1 and u_2 pin down the shape of (respectively) Kate's and the dealer's utility functions as functions of sale price, beyond the minimal fact that Kate's (resp. the dealer's) utility is decreasing (resp. increasing) in sale price. For example, if Kate is relatively relaxed about price differences in the range [£3000, £3800] but cares a lot more about each additional dollar spent above the threshold of £3800, the shape of her utility function will capture this: its slope will then be much steeper in the range [£3800, £5000] than in the range [£3000, £3800].

[8]It is sometimes useful to make the conceptual distinction between the set $\mathcal{A}_X$ of acts and its image $u(\mathcal{A}_X)$ in utility space. However, for the purpose of bargaining theory, usually $u_X(\mathcal{A}_X)$ contains all relevant information about $\mathcal{A}_X$. Accordingly, we will sometimes conflate an act with its image under u_X.

The convexity of $u_X(\mathcal{A}_X)$ is guaranteed if $\mathcal{A}_X$ is closed under probabilistic mixture. That closure condition means that if a_1 is an a_1 happens with probability λ and a_2 happens with probability $1-\lambda$. Such "mixed acts" are arguably a little odd conceptually (Arntzenius, 2008). Assuming closure under probabilistic mixture is however standard in bargaining theory, for technical reasons.

The notion of a disagreement point is worthy of further comment, since its importance is distinctive of the bargaining scenario (no such reference point, for example, plays a role in utilitarian maximisation). A minimal role played by the disagreement point is that it limits the set of agreements that are worthy of consideration: an in-principle-possible agreement that either agent would regard as strictly worse than failure to reach any agreement can be excluded from the modelling exercise from the start, since at least one agent is bound to reject such a candidate agreement. More than this, though, the location of the disagreement point, relative to the utilities provided by the various possible agreements, plays a significant role in determining the balance of bargaining power among the parties. Suppose that the automobile market is highly regulated, so that sale of the vehicle in question outside the price window [£3000, £5000] is forbidden by law. Suppose that Kate is *desparate* to buy a car (suppose, for example, that without a car she cannot make frequent visit to a seriously ill relative who is hospitalised some distance away, and that making such visits is extremely important to her). That is, in utility terms, suppose that Kate's utility at the disagreement point is far below her utility at any of the points corresponding to a legally permitted sale. Suppose further that for the car dealer, the question of whether or not a sale goes through is much lower-stakes. Then, intuitively, the dealer possesses most of the bargaining power; one would expect this consideration to result in a price at a higher sale. Conversely, if Kate is relatively indifferent but the dealer is on the brink of being sacked for not making enough sales, so that the dealer's utility at the disagreement point is far below that at any permitted sale price, then Kate possesses most of the bargaining power. Because of this effect, bargaining theory has to include specification of the disagreement point in its modelling of a bargaining problem.

A *solution* to such a bargaining problem is a point $s \in u(\mathcal{A}_X)$: the utility n-tuple that is selected as a result of the bargaining procedure. A *solution function* is an object that specifies a solution for every possible bargaining problem (formally: a function $\mathcal{S}$ from bargaining problems X to solutions $s \in \mathcal{A}_X$).

The standard solution function is the *Nash bargaining solution*, according to which the selected agreement is always the point in utility space that maximises the product (across players) of the amount by which each player's utility exceeds the utility the player would have at the disagreement point:[9]

$$\mathrm{NBS}(X) = \mathrm{argmax}_{a \in \mathcal{A}_X} \Pi_{i=1,\ldots,n} (u_i(a) – d_i). \tag{1}$$

[9]This solution is, of course, not the only possible one; see footnote 19. However, it is by far the most widely discussed and used. For example, the graduate textbook (Muthoo, 1999) discusses exclusively the Nash solution.

Strategic justifications of the Nash bargaining solution based on a one-shot demand game are given by (Zeuthen, 1930; Nash, 1953; Harsanyi, 1956; Anbar and Kalai, 1978). Strategic justifications based on the Rubinstein alternating offer model are given in Binmore et al. (1986). Axiomatic justifications are given in (Nash, 1950; Lensberg, 1988; Chun and Thomson, 1990; Dagan et al., 2002; de Clippel, 2006).

Suppose, for example, that the utility functions for Kate and the car dealer respectively are given in the region of interest by

$$u_1(x) = \sqrt{\frac{£6000 - x}{£2000}}$$

$$u_2(x) = \ln\left(\frac{x - £1000}{£2000}\right),$$

where x is the sale price, on utility scales on which the disagreement point is given by $d_1 = d_2 = 0$. It is straightforward to verify (for instance, by drawing a graph of the Nash product $u_1(x) \cdot u_2(x)$ against sale price x) that the product of these two quantities is maximised when x is a little under £4800. Roughly £4800, then, is the sale price predicted by the Nash solution in our example.

4.2 A bargaining-theoretic approach to moral uncertainty

The 'bargaining' aspect of the idea of a parliamentary approach to moral uncertainty, recall, is to imagine different moral theories (or delegates, each representing a given moral theory) to bargain with one another. For the application to moral uncertainty, therefore, we will take the 'players' in bargaining theory to be moral theories that our agent has nonzero credence in, rather than people. We will take the options to be available acts that the agent could (unilaterally) choose, rather than available agreements that the multiple persons could agree to. Instead of the utility that a given person would obtain from any of the possible agreements (or from the lack of any agreement), we consider the choiceworthiness that each of the candidate moral theories assigns to the various possible acts that the single agent might perform.

Formally, to model the problem of moral uncertainty in terms of bargaining theory, we will take an *n-theory bargaining problem* to be a structure $X = (T, U, \mathcal{A}_X, u_X, p_X, d_X)$, where

- $T = (T_1, ..., T_n)$ is an n-tuple of moral theories.

We replace the set P of 'players' with a set T of theories, because we are envisaging a (hypothetical) process of bargaining among moral theories, rather than bargaining among persons.

- $U = (U_1, ..., U_n$, where for each $i = 1, ..., n$, U_i is the space of choiceworthiness levels recognised by theory T_i.

- $u_X = (u_1, ..., u_n)$, where for each i, $u_i : \mathcal{A}_X \to U_i$ is the von Neumann-Morgenstern (vNM) choiceworthiness function corresponding to theory T_i. That is, for each i, theory T_i's ranking of acts is ordinally represented by $E[u_i]$, where E is the expectation operator with respect to empirical uncertainty.

The analog of "how much utility would person i have if the outcome of bargaining were x?" is "how choiceworthy does theory i deem the possible action x?", so U and (relatedly) u_X are reinterpreted accordingly.[10]

- $\mathcal{A}_X$ is a set of available options, such that $u_X(\mathcal{A}_X)$ is bounded and convex.

These 'options' are now actions the agent could choose, rather than deals the players could agree to.

- $p_X = (p_1, ..., p_n)$ is the agent's credence distribution over T.
- $d_X \in U$ is the disagreement point.

The last two items on this list deserve more extended comment.

First, p_X, the agent's credence distribution over the set of moral theories under consideration, has no precise analog in the ordinary case of bargaining among persons. We have to include it for a plausible bargaining-theoretic approach to moral uncertainty, however, because any plausible such approach will assign higher bargaining power (other things being equal) to theories in which the agent has higher credence.

The closest analog to this in the case of ordinary bargaining among persons would be a multi-player bargaining problem in which some non-trivial groups of agents have identical utility functions. For example, if, instead of just Kate and the car dealer, there were a hundred prospective customers and one car dealer, constrained to all agree on a single sale price or none, and if all of the hundred prospective customers had identical utility functions as a function of the agreed car price x, then the Nash solution would require maximisation of the dealer's

[10]Thus, we assume that all moral theories obey the axioms of expected utility theory, in their treatment of empirical uncertainty. Not all moral theories have a structure that is consistent with this assumption. Our assumption excludes theories that violate transitivity even in the absence of uncertainty, as well as theories that accept transitivity but that deal with uncertainty in some other way (for example, via a maximin formula). This is a little awkward, since even if such moral theories seem implausible at the first-order level, ideally we would like our metanormative theory to apply to agents who have nonzero credence in such theories. However, we do not know how to adapt bargaining theory so that this assumption is not required. In this respect, the bargaining-theoretic approach is in the same boat as the MEC approach to moral uncertainty (MacAskill, 2013; Riedener, 2020, p.491; Tarsney, 2021).

utility multiplied by a customer's utility *to the power of 100*; customers would then have significantly higher bargaining power than in the case in which Kate is the sole customer.[11]

We could represent unequal credences in moral theories by simply including more copies of higher-credence theories and fewer copies of lower-credence theories in the bargaining problem, in proportion to the agent's credences. This, though, is a clumsy approach.[12] The modelling method that we favour obtains equivalent results, but in a more elegant way.

Second, much more so than in the context of ordinary bargaining among persons, it is unclear what constitutes the 'disagreement point' in the application to moral uncertainty.[13] 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 agreemement'.

One might be tempted to take this observation to be fatal to the whole idea of treating decision-making under moral uncertainty as a case of bargaining among theories. If the outcomes of bargaining depend essentially on what would happen in the absence of any mutually agreed deal, and if there is no such thing as 'what would happen in the absence of any deal mutually agreed between theories' in the case of moral uncertainty, then, the thought runs, so much the worse for the idea that decision-making under moral uncertainty should match the outcome of a hypothetical process of bargaining among theories.

But this would be much too quick. A bargaining-theoretic approach would be *more appealing* if the analogy between ordinary bargaining and the problem of moral uncertainty were more complete: in particular, if there were a fact of the matter about 'what would happen if bargaining between theories broke down', so that the element d of the bargaining-theoretic formalism had a direct and intuitive interpretation closely analogous to the interpretation it has in ordinary applications of bargaining theory. But this does not immediately mean that a bargaining-theoretic approach to moral uncertainty is *impossible*, or *incoherent*, in the absence of such a close similarity of interpretation. As long as the location of the disagreement point in extensional terms can be specified, the resulting theory might earn justification via the adequacy of its verdicts about cases, even if there is no more direct justification for why the disagreement point should be taken to be located here rather than there. Given the perceived inadequacy of extant solutions to the problem of moral uncertainty, we will press on, and investigate how much extensional adequacy is on offer.

[11] As is straightforward to verify, according to the Nash solution the agreed sale price in this example would be around £3600, i.e. significantly more favourable to the purchasers than the price of £4800 we saw in the original example above.

[12] Relatedly, it would do a particularly clumsy job of handling cases in which the agent's credences in moral theories were arbitrary real numbers, rather than rational numbers.

[13] For some discussion of subtleties of the notion of disagreement in the ordinary case of bargaining, see Binmore et al. (1986).

For a plausible metanormative theory, however, we do require that the specification of the disagreement point be reasonably simple and elegant; it would violate this very general principle of theory-building if one were to postulate a gerrymandered collection of disagreement points (across cases) so as to reverse engineer whatever verdicts about appropriate action under moral uncertainty one antecedently found plausible. Some reasonably elegant suggestions for location of the disagreement point include the following; cf. figure 1.

- For an arbitrary set of acts $\mathcal{A}$, let the 'anti-utopia point' of $\mathcal{A}$, $\underline{u}_{\mathcal{A}}$, be the point in utility space corresponding to the lowest available utility for each player. (More precisely: $\underline{u}_{\mathcal{A}} = (\inf_{a\in\mathcal{A}} u_1(a), ..., \inf_{a\in\mathcal{A}} u_n(a))$. Note that this anti-utopia point need not correspond to any available act (that is, we need not have $\underline{u}_{\mathcal{A}} \in u(\mathcal{A})$). Then we might take d_X to be the anti-utopia point relative either to $\mathcal{A}_X$ itself, or to the Pareto frontier of $\mathcal{A}_X$.[14]
- We might set $d_X = \mathrm{RD}_X$, the random dictator point, where, for each i, the act that is highest-ranked by T_i is selected with probability p_i.[15]
- We might take d_X to be determined in some ("reasonably simple and elegant") exogenous way — "exogenous" in the sense that its location in utility space does not supervene on $U(\mathcal{A}_X)$. For instance, perhaps d_X corresponds to 'doing nothing', or to performing whichever option is best with respect to non-moral considerations.

In what follows, as far as possible we will proceed in a way that is independent of how the disagreement point is identified, noting where and how the location of the disagreement point makes a difference to our qualitative results.

We return now to the distinction between U_i (for some given theory T_i) on the one hand, and the real number line $\Re$ on the other. These two objects have *similar* structures. In particular, both give rise to well-defined ratios of differences: for any quadruple (a, b, c, d) of real numbers, the arithmetic of real numbers gives us a well-defined ratio $\frac{a-b}{c-d}$, and the structure of von Neumann-Morgenstern (vNM) theory guarantees that sense can be made of the same ratio of differences if a, b, c, d are instead vNM choiceworthiness levels from the scale corresponding to some given moral theory. However, the real numbers also have *additional* structure that these utility scales do *not* possess, namely a privileged

[14]The Pareto frontier of a given set is defined to be the set of points from which no Pareto improvement is possible within that set:

$$\mathrm{Pareto}(\mathcal{A}) = \{a \in \mathcal{A} : (\neg\exists a' \in \mathcal{A})(\forall i(u_i(a') \geq u_i(a)) \wedge \exists i(u_i(a') > u_i(a))\}.$$

[15]As stated, the random dictator point is well-defined only when no theory is indifferent between two or more top-ranked options (unless all other theories are also indifferent between the options in question). We set aside the issue of whether and how the definition might be generalised to accommodate this potential obstacle.

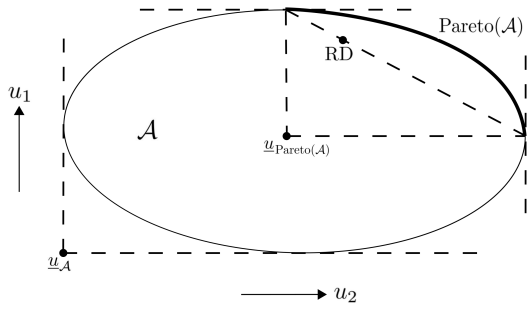

Figure 1: *An arbitrary set of available options $\mathcal{A}$, with the Pareto frontier* Pareto($\mathcal{A}$)*, anti-utopia point $\underline{u}_{\mathcal{A}}$ and random dictator point* RD *(for the case $p_{1,} = 0.75, p_2 = 0.25$) labelled.*

zero point and a privileged unit. We can represent utility levels (in the context of a given moral theory) by real numbers, and for concrete applications it is often useful to do so. But because of the surplus structure this introduces, there is a certain amount of arbitrariness, corresponding to the choice of which choiceworthiness levels to map to the real numbers 0 and 1, in any particular choice of such a real-valued representation. We will write $\mathcal{F}_n$ for the set of coordinate functions $f : U \to \Re^n$ that are admissible, in the sense that they preserve ratios of differences within each moral theory. Each such f can equally be thought of as an n-tuple $(f_1, \ldots f_n)$, where each f_i maps the choiceworthiness scale U_i for theory T_i to (one copy of) the real numbers $\Re$. If $f = (f_1, \ldots f_n)$ is one admissible coordinate system, then so is $f' = (f'_1, \ldots, f'_n)$, where for each $i, f'_i = a_i f_i + b_i$, for some positive numer a_i and some real number b_i.[16] For any $f \in \mathcal{F}_n$, we will say that the real-valued compound function $f \circ u$ *cardinally represents* the theories' choiceworthiness functions.

4.3 The asymmetric Nash Bargaining Solution

For the application to moral uncertainty, the (generally unequal) distribution of credences across theories must make a difference. This suggests considering the asymmetric version of the Nash bargaining solution:[17,18]

[16]More precisely, $\mathcal{F}_n$ is the set of all n-tuples of affine maps $(f_1, ..., f_n)$ such that $\forall i = 1, ..., n, f_i : U_i \to \Re$.

[17]Justifications for the asymmetric Nash solution are given in (Harsanyi and Selten, 1972; Kalai, 1977a; Anbar and Kalai, 1978; Anbarci and Sun, 2013).

[18]Strictly speaking, equation (2) does not make sense, since we do not have an operation of multiplication defined between elements of the utility spaces U_i. The official definition is

$$\text{NBS}(X) = \text{argmax}_{a \in \mathcal{A}} \Pi_{i=1,\ldots,n} (f \circ u_i(a) - f \circ u_i(d))^{p_i},$$

where $f \in \mathcal{F}_n$ is any admissible coordinate system. It is easy to verify that NBS(X), thus defined, is independent of the choice of f . Where the choice of $f \in \mathcal{F}_n$ thus makes no difference, we will usually omit explicit mention of f.

$$\text{NBS}(X) = \text{argmax}_{a \in \mathcal{A}_X} \Pi_{i=1,\ldots,n} (u_i(a) - d_i)^{p_i}. \quad (2)$$

Importantly, the solution (2) satisfies a condition of clone independence: since $x^\alpha x^\beta = x^{\alpha+\beta}$, it makes no difference whether we represent a given agent as having credences of p_{i_1}, p_{i_2} respectively in each of two qualitatively identical copies of a given theory T, or instead simply as having credence $p_{i_1} + p_{i_2}$ in T. This means that a bargaining-theoretic approach based on the Nash solution (2), like the "maximise expected choiceworthiness" approach, does not suffer from the problem of theory individuation that plagued the "my favourite theory" approach to moral uncertainty.

Just as the symmetric Nash solution is not the only game in town in the context of ordinary bargaining theory (cf. fn. 9), so the asymmetric Nash solution is also far from the only one we might sensibly consider in the context of moral uncertainty. Nonetheless, in this paper we will consider only the asymmetric Nash solution. This is simply to limit the size of our task; our purpose here is to make a *start* on investigating the prospects for a bargaining-theoretic approach to moral uncertainty by exploring the application of *one plausible* solution function to the problem of moral uncertainty. Comparison of the relative merits of this and other solution functions, and the prospects for an axiomatic derivation of a solution function from axioms that seem compelling in the context of moral uncertainty, lie beyond the scope of this paper.[19]

To see the asymmetric Nash bargaining solution in operation, consider an agent who has non-zero credence in only two moral theories, but has unequal credences in those theories. In the diagrams in figure 2, the disagreement point is located at the origin, the contours are lines of constant Nash product, and the heavy black line represents the Pareto frontier of a representative set of available acts. We illustrate the Nash contours for various ways of splitting credence between the two theories. Note that as credence in T_1 increases, the bargaining solution moves leftward along the Pareto frontier, favouring T_1 relative to T_2, as expected.

4.4 Reformulation in terms of maximisation of an expectation

Maximisation of the Nash product is of course equivalent to maximisation of any strictly increasing function of the Nash product. An alternative representa-

[19] Other solution functions that are relatively prominent in the literature on bargaining theory include the Kalai-Smorodinsky solution (Kalai and Smorodinsky, 1975), the Mascher-Perles solution (Perles and Maschler, 1981), and the "proportional" solutions of Kalai (1977b). (The utilitarian optimum may also be considered a solution function; in the context of moral uncertainty this of course corresponds to maximising expected moral value.)

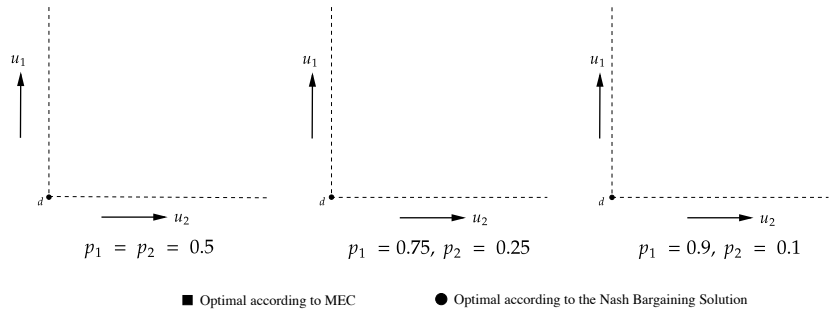

Figure 2: *Nash contours and the asymmetric Nash bargaining solution, for a generic convex Pareto frontier, illustrated for various distributions* (p_1, p_2) *of credence between two moral theories* (T_1, T_2).

tion that is of particular interest comes from taking the logarithm of the Nash product. Recall that the Nash product is

$$\mathrm{NP}(a) = \Pi_i (u_i(a) - d_i)^{p_i},$$

so that taking logarithms gives

$$\log \mathrm{NP}(a) = \sum_i p_i \log(u_i(a) - u_i(d)) \tag{3}$$

It is instructive to compare the expression (3) with expected choiceworthiness, as maximised under the MEC approach:

$$EC(a) = \sum_i p_i u_i(a). \tag{4}$$

In understanding the relationship between the MEC and Nash approaches, two differences between equations (3) and (4) are worthy of comment.

First, according to the MEC approach, overall preferences (under moral uncertainty) are risk neutral with respect to the von Neumann-Morgenstern utility functions u_i that expectationally represent *each theory* T_i*'s* choiceworthiness ordering of gambles. (This is required if preferences over gambles in the presence of moral uncertainty are to satisfy the axioms of expected utility theory, as is standardly assumed in the MEC approach.[20]) In contrast, the Nash approach,

[20]For a proof, see Riedener (2021), chapter 2.

we see from equation (3), depends nonlinearly on the utility functions u_i. Since the logarithm function is strictly increasing but strictly concave, it is tempting to say that the Nash approach is risk averse with respect to the (theory-specific) utility differences $u_i(x) - u_i(d)$, though the absence of privileged intertheoretic utility comparisons makes it a little delicate to give sense to this claim.[21]

Second, the challenge that the MEC approach faces in fixing intertheoretic comparisons is in a tolerably clear sense replaced, in the Nash approach, with the challenge of fixing the disagreement point. The quantity (4) is well-defined only insofar as there is a fact of the matter about which utility function u_i (from an infinite family of such functions, related to one another by positive affine transformations) is the appropriate representative, for MEC's purposes, of theory i's choiceworthiness verdicts. That is the issue of intertheoretic comparisons. In contrast, for the purposes of (3) it does not matter which u_i we choose to represent each theory, provided only that u_i is *a* vNM choiceworthiness function for theory i — it is a special feature of the formula (3) that substituting one such u_i for another does not alter the matter of which option(s) maximise(s) the quantity in question. The price that the Nash approach pays for this is that (3) is instead well-defined only relative to a specification of the disagreement point d.

5 General results

In this section, we establish some simple bargaining-theoretic results that will be illuminating for the purpose of seeing what bargaining theory might say about the problem of moral uncertainty.

To keep things simple, we will continue to consider cases in which the agent has non-zero credence in only two moral theories, T_1 and T_2. However, we are

[21] To say that one is risk averse with respect to a given quantity is to say that one strictly disprefers a gamble g_1 to another gamble g_2 if g_1 is a mean-preserving spread of g_2. But whether or not one option is a mean-preserving spread of another, under moral uncertainty, with respect to the theory-specific choiceworthiness functions u_i, depends in part on the choice of representative utility functions u_i. For example, with choiceworthiness levels were as in the first table below, g_1 is a mean-preserving spread of g_2. But the second table (using a different representative utility function for T_2) is an equally admissible representation of the same pair of gambles, and with those numbers, neither gamble is a mean-preserving spread of the other. What we can say is that if there exists an admissible utility representation such that g_1 is a mean-preserving spread of g_2, the Nash approach prefers g_2 to g_1.

	u_1	u_2
a	2	2
b	3	1

	u_1	u_2'
a	2	10
b	3	5

not aware of any difficulties with generalising our results to an arbitrary finite number of theories.

5.1 Two pure options

Suppose first that there are only two pure options, a_1 and a_2. Then the set of available acts, $\mathcal{A}$, contains all gambles $\lambda a_1 + (1 - \lambda a_2)$ that assign some probability $\lambda \in [0, 1]$ to a_1 and the remaining probability $(1-\lambda)$ to a_2 (formally: $\mathcal{A} = \mathrm{Conv}(a_1, a_2)$, the convex hull of a_1 and a_2).

Suppose further that T_1 strictly prefers a_1 to a_2, while T_2 has the reverse strict preference. Then all available acts lie on the Pareto frontier ($\mathrm{Pareto}(\mathcal{A}) = \mathcal{A}$): starting from any act in the available set, any change of the parameter λ that is preferred by T_1 will be dispreferred by T_2. In diagrammatic terms, the set of available options projects (via any $f \in \mathcal{F}$) to a downward-sloping straight line in $\Re^2$; see figure 3.

The Nash solution $\mathrm{NBS}(X)$ of course depends not only on the set of available options, but also on the disagreement point. Say that a two-pure-option bargaining problem X is *canonical* iff the disagreement point is the anti-utopia point of the set of available acts (if, that is, $d_X = \underline{u}_{\mathcal{A}_X}$). This is the easiest case to analyse:

Proposition 1. *If a two-theory, two-pure-option bargaining problem X is canonical, the Nash bargaining solution* $\mathrm{NBS}(X)$ *coincides with the random dictator* RD_X.

We further have

Lemma 2. *If a two-theory, two-pure-option bargaining problem X is canonical, the Nash product increases along the Pareto frontier* $\mathrm{Pareto}_{\mathcal{A}_X}$ *as one moves towards the global maximum* RD_X *from either side.*[22]

Suppose now that X is not canonical. We can nonetheless define the *canonical problem corresponding to X*, $\mathrm{Can}(X)$, by keeping the disagreement point fixed ($d_{\mathrm{Can}(X)} = d_X$) and defining the set of available acts in $\mathrm{Can}(X)$ so that it corresponds to just those points in utility space on the extension of the

[22] To prove Proposition 1, consider the first-order condition

$$\frac{d}{d\lambda}\left(\Pi_{i=1,\ldots,n}(u_i(\lambda a_1 + (1-\lambda)a_2))^{p_i}\right) = 0 \tag{5}$$

to identify the point of maximum Nash product along the straight line in utility space joining $u(a_1)$ to $u(a_2)$. For Lemma 2, consider the sign of the LHS of (5).

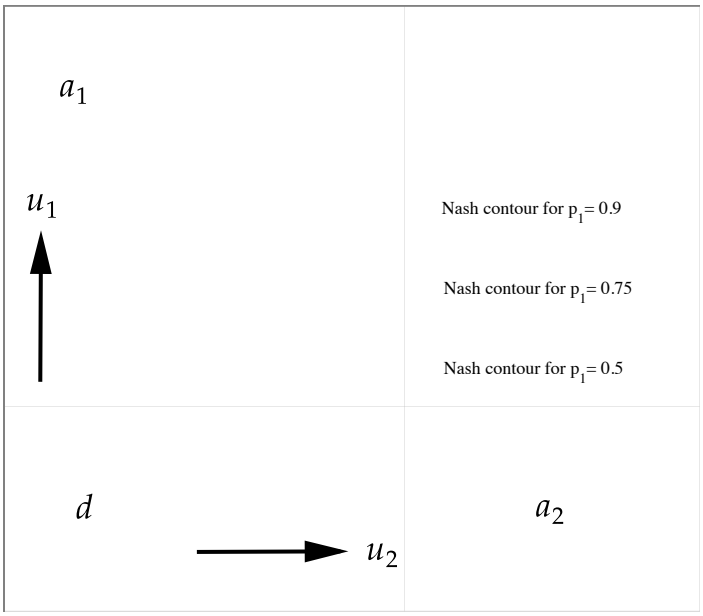

Figure 3: *The Nash solution for a canonical two-theory two-pure-option bargaining problem (Proposition 1).*

straight line $u(a_1)u(a_2)$ that are weak Pareto improvements over d_X (that is, $\{x \equiv \lambda u(a_1) + (1 - \lambda u(a_2) : \lambda \in \Re \wedge u_1(x) \geq u_1(d_X) \wedge u_2(x) \geq u_2(d_X)\}$).

The NBS for an arbitrary two-theory, two-pure-option bargaining problem X is easily stated in terms of the NBS to the corresponding canonical problem: we have

Proposition 3. *If X is a two-theory, two-pure-option bargaining problem, then* $\mathrm{NBS}(X)$ *is that point on the Pareto frontier* $\mathrm{Pareto}_{\mathcal{A}_X}$ *that is closest to the random dictator point* $\mathrm{RD}_{\mathrm{Can}_X}$ *of the corresponding canonical problem.*

The proof is immediate from Lemma 2.

As we can see with reference to figure 4, Proposition 3 covers several cases. It could be that the random dictator point corresponding to the canonical problem is itself one of the acts available in the actual decision problem ($\mathrm{RD}_{\mathrm{Can}(X)} \in \mathcal{A}_X$), so that the Nash solution simply selects this point ($\mathrm{NBS}(X) = \mathrm{RD}_{\mathrm{Can}(X)}$). This can happen whether (a) the disagreement point is a Pareto improvement over the anti-utopia point ($d_X \gg \underline{u}_X$), (b) the reverse ($\underline{u}_X \gg d_X$), or (c) the disagreement point is preferred to $\underline{u}_X$ by one of the theories (say, $d_X \succ_1 \underline{u}_X$) and dispreferred to $\underline{u}_X$ by the other theory ($\underline{u}_X \succ_2 d_X$). Alternatively, it might be that the random dictator point corresponding to the canonical problem lies

outside the set of available acts ($\mathrm{RD}_{\mathrm{Can}(X)} \notin \mathcal{A}_X$) — either off to the "top-left" or the "bottom-right" of the line segment that represents the available acts — in which case it follows from Proposition 3 that the NBS is one of the two available pure options (a_1 or a_2). Again, this can happen for various locations of the disagreement point relative to $\underline{u}_X$ (cases (d) and (e)).

5.2 Three or more pure options

If there are more than two pure options available, then the Pareto frontier can be strictly convex. This increases the tendency (already seen to some extent in section 5.1) for the Nash solution to be a point at which both theories attain a reasonably high proportion of their attainable expected utility (relative to the disagreement point), rather than a more extremal point on the Pareto frontier. This is illustrated by the following example.

> **Example 4 (Two extremal options and a unanimous nearly-as-good third option).** Suppose that (for some admissible coordinate system) the pure options A, B, C have utilities $u(a) = (3, 0), u(b) = (2.1, 0.7), u(c) = (0, 1)$, that the available acts are the probabilistic mixtures of these three pure acts (so that $\mathcal{A} = \mathrm{Conv}(a, b, c)$), and that the disagreement point is the anti-utopia point rof this set of available acts ($d = \underline{u}$). Then it is straightforward to show that for any value of p_1 between 0.3 and 0.7, the Nash bargaining solution selects b over any other pure or mixed option.[23]

In the following sections, we apply these abstract results to various issues of interest.

6 Sensitivity to relative stakes

In at least some cases, what it is appropriate to do under moral uncertainty seems to depend on issues of *relative stakes*. That is (roughly), if according to one theory the decision under consideration is a very low-stakes one, while according to a second theory the same decision is a high-stakes one, that difference should induce a shift towards following the dictates of the second theory.

[23] In contrast, the MEC approach selects a for any value of p_1 above $\frac{7}{16}$, if the intertheoretic unit comparisons are as suggested by the representative utility functions used in the above presentation of this example. (The Nash approach, of course, is insensitive to how the intertheoretic unit comparisons are settled.)

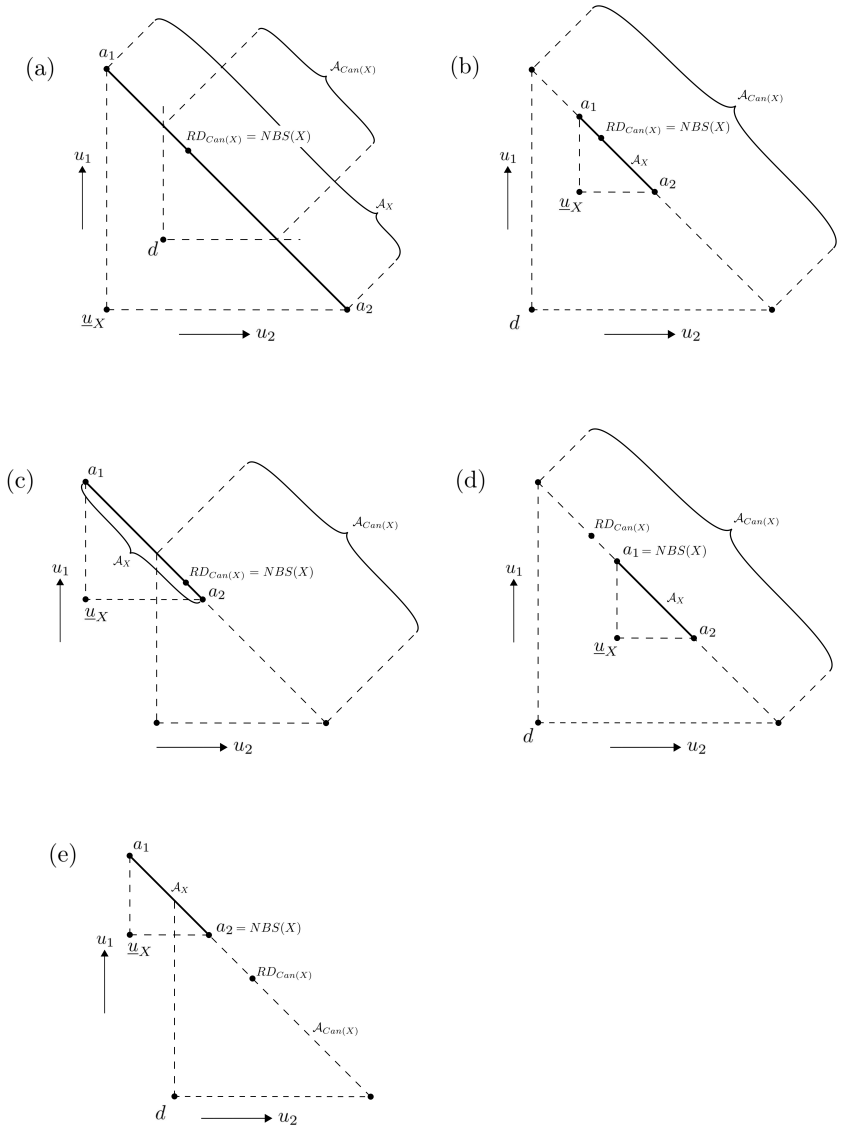

Figure 4: *The Nash bargaining solution for a variety of non-canonical two-theory, two-pure-option bargaining problems. In all cases, the Nash solution* $\mathrm{NBS}(X)$ *to the problem* X *is the available act that is closest to* $RD_{\mathrm{Can}(X)}$*, the Nash solution to the corresponding canonical problem* $\mathrm{Can}(X)$ *(Proposition 3).*

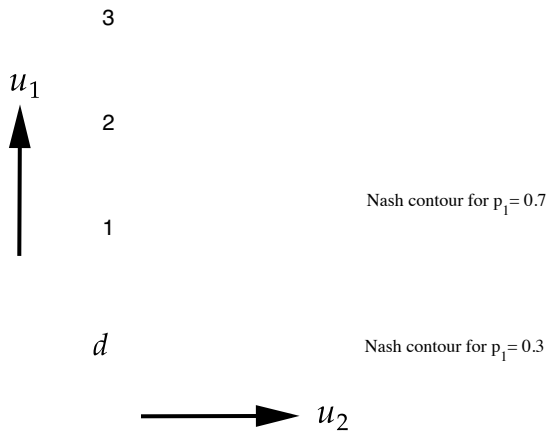

Figure 5: If there are two extremal options and a unanimous nearly-as-good third option, the Nash approach tends to select the latter (Example 4).

Suppose, for example, that one is uncertain about whether or not eating meat is morally permissible. One has 60% credence in a moral view according to which eating meat is morally unproblematic, but 40% credence that eating meat is morally on a par with cannibalism. And suppose that one has only a very slight preference, conditional on its being permissible, for eating meat over refraining (one very marginally prefers the taste of meat, but also likes vegetarian food). Then it might well be appropriate, under moral uncertainty, to refrain from eating meat (Sepielli, 2010, pp.54-6; MacAskill and Ord, 2018, pp.11-12). This is of course not a verdict with which the "my favourite theory" approach can agree. The "maximise expected choiceworthiness" approach can; this seems to be a significant advantage of the latter over the former approach (MacAskill et al., 2020, pp. 44-5).

As we noted above (fn. 5), there is a tension between sensitivity to relative stakes on the one hand, and absence of intertheoretic comparisons on the other. At first sight, it seems that there can be a difference in relative stakes only if there is a standard of intertheoretic comparisons: to say that in a given choice between options A and B, the stakes are higher according to T_1 than they are according to T_2, *just is* (it seems) to say that the magnitude of the choiceworthiness difference $u_1(A) - u_1(B)$ is greater than that of $u_2(A) - u_2(B)$. As we noted above, the Nash approach does not use intertheoretic comparisons, so one might think that it simply cannot be sensitive to any considerations of difference in relative stakes.

However, in addition to this primitively intertheoretic sense of 'difference in

relative stakes', there is also an intratheoretic sense. In the latter sense, roughly and heuristically, to say that the stakes are higher (in the choice between A and B) according to T_1 than they are according to T_2 is to say that T_1 regards the choice between A and B as *more important than other relevant choices* to a greater degree than does T_2. Since every moral theory (we are assuming) comes equipped with a standard of intratheoretic unit comparisons, such an intratheoretic notion of difference in relative stakes need not presuppose any controversial structure.

The Nash approach can recognise differences in relative stakes in this intratheoretic sense, and can respond in the intuitively appropriate way to them. Consider:

> **Example 5 (Two binary choices).** Jenny faces two independent binary choices. She can either kill one person, or let two die; and she can either donate a fixed philanthropic budget of $1m to support homeless people, or to mitigate extinction risk. Her credence is split equally between two moral theories. Jenny has 50% credence in a total utilitarian moral theory T_1, according to which it is (relatively speaking) slightly better to kill one than to let two die, but much better to direct the resources to extinction risk mitigation than to homeless support. And she has 50% credence in a commonsense moral theory T_2, according to which it is (relatively speaking) slightly better to direct resources to homeless support than to extinction risk mitigation, but much worse to kill one than to let two die.[24]

(u_1, u_2)	Kill one	Let two die
Donate to extinction risk mitigation	$(+10, -10)$	$(+9, +9)$
Support local homeless people	$(-9, -9)$	$(-10, +10)$

Similar to Example 4, in Example 5 the Pareto frontier is strictly convex, with the option of supporting extinction risk mitigation and letting two die playing the role of "good compromise" between T_1 and T_2. Similarly to Example 4, the Nash approach will tend to select the "good compromise" option.[25]

We can understand this in terms of an intratheoretic notion of "relative stakes" as follows. In a variant of Example 5 in which the *only* choice was whether to kill one or let two die, the Nash approach would in general hold, counterintuitively,

[24]The precise equality of credences is unimportant to this example; it merely simplifies the numbers.

[25]For example, this happens if the disagreement point is the anti-utopia point $\underline{u}_{\mathrm{Pareto}(\mathcal{A}_X)}$.

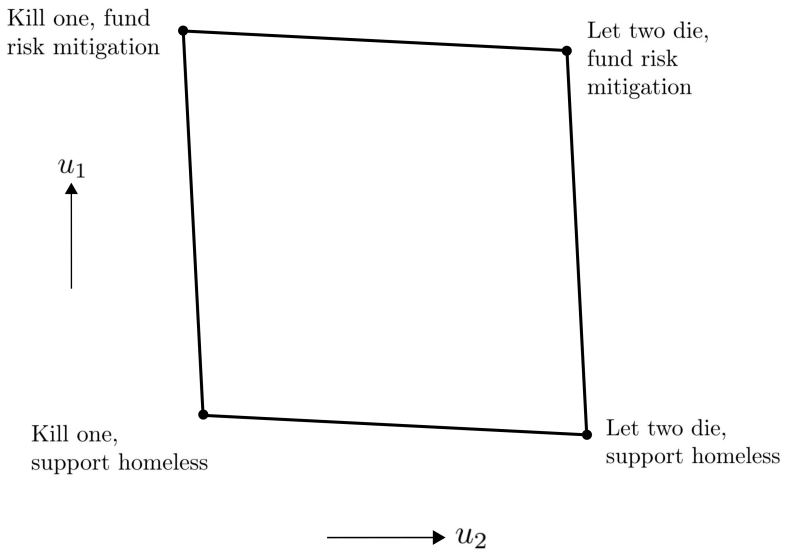

Figure 6: *In decision problems that can be decomposed into independent choices on distinct moral issues, with different moral theories regarding different issues as higher-stakes, the Nash approach tends to favour options that are good compromises between the relevant theories (Example 5).*

that the optimum is some mixed act in which one kills with positive probability. That reflects the fact that the Nash approach does not make use of any primitive intertheoretic notion of relative stakes, and so is unable to take account of any primitive sense in which avoiding killing is much more high stakes on a common-sense theory than reducing deaths is on a utilitarian theory. However, once other choices are simultaneously on the table, the Nash approach *does* recognise facts that the difference between "kill one" and "let two die" (other things equal) are much greater according to common-sense morality than the differences between "support the homeless" and "support extinction risk mitigation", and that the reverse is true according to total utilitarianism. We can mentally break down the choice between portfolios of action into a number of choices each concerning one issue (the choice between killing one and letting two die, and the choice between supporting the homeless and mitigating extinction risk). Then, even without primitive intertheoretic comparisons, a given theory can say that one issue is higher-stakes *than the other issue is according to the same theory.* When theories disagree with one another about which issues are higher- vs. lower-stakes, there is a tendency in the Nash approach to select an option, if there are any, according to which each theory gets its way on the issues that it regards as high stakes — because, in geometric terms, those options tend to 'stick out' on the Pareto frontier, in the way illustrated by Figure 6.[26]

[26]To exhibit this stakes-sensitivity, of course, it is necessary that the Nash approach models

What about the example of vegetarianism with which we opened this section? If whether or not to eat meat were the only moral choice, then the bargaining-theoretic approach would not be able to recover the (alleged) fact that in the example as specified, it is appropriate to be vegetarian on grounds of relative stakes. Some might consider this problematic. In a real-life version of the case, though, our example of vegetarianism is relevantly similar to Example 5, and so we can expect a bargaining-theoretic approach to deliver the intuitively correct verdict on that case. For whether or not to eat meat is not (in real life) the only moral issue on which a would-be diner should feel torn. And if one is torn between one view according to which it is extremely morally important not to eat meat and a matter of relative moral indifference whether or not to (say) contribute to charity, and another view according to which the reverse is the case, a bargaining-theoretic approach will plausibly recommend both vegetarianism and charity contributions, on the pattern of Example 5. In particular, it will plauisbly recommend vegetarianism, as seems appropriate.

In the examples considered in this section, the agent's credence in the various theories under consideration is moderate (in the region 40-60%). A special issue concerns the appropriate treatment of issues of relative stakes when a theory in which one has *very* low credence is also one according to which the stakes are *very* high. The latter is the issue of "fanaticism", to which we will return in section 8.

7 Splitting resources between moral theories

7.1 Splitting resources vs. flipping coins

One is often faced with a spectrum of possible pure options, spanning a range from the option that is most-preferred by one theory to the option that is most-preferred by the other theory. In some such cases, it is at the very least psychologically natural, and perhaps also appropriate, to choose some strictly intermediate option, rather than putting all one's eggs in one moral basket. The following example is of this type:

> **Example 6** (**Splitting the pot**)**.** James has a fixed philanthropic budget, and is considering two interventions that he could fund. Intervention A targets poverty, while intervention B targets animal welfare. In the absence of uncertainty, James would simply seek to maximise the amount of good done. There is (we stipulate) no

the two choices (killing vs. letting die, and where to direct resources) simultaneously, within a single four-option decision. Herein lies a problem for that approach. We will return to this in section 10.

> relevant empirical uncertainty. However, because of fundamental moral uncertainty, James is very uncertain about the appropriate "rate of exchange" between units of poverty alleviation and animal welfare promotion. Thus, he is very uncertain about which of these interventions is the more cost-effective in terms of *good done* per unit resource expended. James has some credence p in a moral theory T_1 according to which animal suffering counts for little, so that A is enormously more choiceworthy than B, but he also has some credence $1-p$ in a moral theory T_2 according to which the opposite is true. The available alternatives lie on a continuum: given a total pot of size X, for any $\lambda \in [0,1]$, James can spend λX on the first intervention and $1-\lambda X$ on the second intervention (write $\lambda(A,B) = \lambda A + (1-\lambda)B$ for this intervention).

Anecdotally, as a matter of empirical fact, many people faced with decision situations relevantly similar to this feel a powerful pull towards splitting their philanthropic pot: they are content to choose some intermediate option with $\lambda \in (0,1)$, but are not content to choose either extremal option ($\lambda = 0$ or $\lambda = 1$).

The MEC and NBS approaches are both able to rationalise this intuition, but with differences over potentially important matters of detail.

The MEC approach can rationalise it *provided* the vNM choiceworthiness function of one or both theories exhibits diminishing marginal returns to spending on the relevant intervention (that is, $u_1(\lambda(A,B))$ and/or $u_2(\lambda(A,B))$ are concave as functions of λ). If, on the other hand, the theories' vNM choiceworthiness functions are linear in λ, MEC will hold either that it is uniquely appropriate to select one or the other extremal option (whichever has higher expected utility), or that all available options are equally choiceworthy (if the two extremal options have equal expected utility to one another) (Snowden, 2019).

The Nash approach can justify splitting the pot even if the relevant choiceworthiness functions are linear in λ. For example, if the disagreement point is anti-utopia, then (by Proposition 1) given linear choiceworthiness functions the Nash approach prefers the pure option that sends a proportion p (resp. $1-p$) of the philanthropic pot to intervention A (resp. intervention B) over any other pure option.[27]

Since the question of whether or not choiceworthiness functions are linear in λ is itself a rather theoretically embedded matter, the import of this difference

[27] One natural-seeming way to conceptualise this is via thinking of the Nash approach as being risk averse, under moral uncertainty, with respect to the difference between an option's vNM choiceworthiness gain over the disagreement point. In this way of thinking, the point is that failing to split the pot is *riskier*, with respect to moral uncertainty, than splitting the pot. For a note of caution regarding this conceptualisation, see footnote 21.

is easiest to appreciate by considering which patterns of preferences between (degenerate or otherwise) gambles each of the approaches is able to rationalise. To this end, consider the following four alternatives: (a) use all the resources in the way that T_1 favours, (b) use all the resources in the way that T_2 favours, (c) split the pot, with a proportion p (resp. $1-p$) of the pot being used in the way T_1 (resp. T_2) favours, (d) flip a weighted coin to decide between (a) and (b), so that the (whole) pot is used in the way T_1 (resp. T_2) favours with probability p (resp. 1-p).

MEC can rationalise a pattern of preferences according to which $c > a \sim b \sim d$, but not one according to which $d \succ a \sim b$. The Nash approach, on the other hand, can rationalise a pattern of preferences according to which $c \sim d \succ a \sim b$ (or, via nonlinear choiceworthiness functions u_i, $c \succ d \succ a \sim b$), but (given the assumption that T_1 strictly favours one of the pure options while T_2 strictly favours the other) not one according to which $a \sim b \sim d$. Which approach gives a superior treatment of the issue of compromise then depends on whether or not coin-flipping is an adequate method of compromise in this context. This seems to us to be a question on which intuition offers no unambiguous answer.

Our conclusion from this part of the discussion is therefore not that either MEC or the Nash approach is clearly superior to the other in its treatment of Example 6. What we take away is rather that the two approaches are *interestingly different*. Both are serious contenders, and it is worth keeping both possibilities under consideration.

7.2 Violation of Independence

Consider

vNM Independence. Let L, M, N be any lotteries. If $L \preceq M$ then for any $p \in [0,1], pL + (1-p)N \preceq pM + (1-p)N$.

In its normal interpretation, vNM Independence is a constraint on preference orderings, not a constraint on choice functions. It therefore does not strictly speaking apply to the Nash approach, since (strictly speaking) the latter specifies only a choice function. However, we obtain a natural "vNM Independence" condition for choice functions by reading $x \preceq y$ as "it is permissible to choose y from the choice set $\{x, y\}$".

The Nash approach violates this Independence condition. To see this, consider again Example 6. Let $M = A$, and let $L = N = B$. Then vNM Independence would require that if $A \sim B$, then also $pA + (1-p)B \sim B$. As we saw in section 7.1, however, the mixed act $pA + (1-p)B$ can have a strictly higher Nash product

than either pure option A, B, including when $A \sim B$; thus $pA + (1-p)B \succ B$, contradicting Independence.

Meanwhile, MEC satisfies vNM Independence. This suffices to establish that, despite its various similarities to a version of MEC with structural intertheoretic normalisation, the Nash approach to moral uncertainty is extensionally distinct from any version of MEC.[28]

8 Fanaticism

Picking up the thread that we left hanging at the end of section 6: to what extent should decisions under moral uncertainty be dictated by theories in which one has extremely low credence, if according to those same theories the moral stakes are extremely high?

It does seem appropriate for such theories to have *some* influence over decision-making under moral uncertainty, in *some* such cases. It is arguably appropriate, for example, for a person who has *almost* 100% credence in consequentialism to balk at killing one for the only somewhat greater good, on the grounds that (i) she has a lingering non-zero credence in the relevant deontic constraint, and (ii) killing is *very strongly* dispreferred by theories postulating that constraint. On the other hand, intuitively there are limits to this. For example, it seems inappropriate if a theory in which one has extremely low credence (say, well under 1%) *entirely dictates* one's behaviour under moral uncertainty, simply on the grounds that that theory entails that every moral issue is (say) at least a thousand times more important than any rival moral theory says the same issue is. As is well recognised, this is potentially a problem for the MEC approach. Even if one has only credence 0.0001% in Theory 1 and has credence 99.9999% in Theory 2, for example, it nonetheless maximises expected choiceworthiness to defer entirely to Theory 1 in a given binary choice if the choiceworthiness difference between the two options in question is more than 10,000 times higher according to Theory 1 than it is according to Theory 2 (Cotton-Barratt et al., 2020, p.73, Ross, 2006, pp.765-6).

One theoretical possibility is that one should simply ignore theories in which one has extremely low credence (perhaps: less than 1%), regardless of issues of relative stakes. However, this is implausible, for two reasons. First, it is structurally incoherent: like the "my favourite theory" approach to moral uncertainty, the application of such a principle depends on how theories are individuated.

[28] We do not intend violation of Independence in itself to constitute a decisive objection to the Nash approach. Clearly, any metanormative theory that is genuinely distinct from MEC will have to violate one or more of the axioms of expected utility theory. So, if we are interested in alternatives to MEC at all, it is in the nature of the enterprise that we are open in principle to countenancing such violations.

Second, the suggestion anyway seems to give the wrong verdicts on an important class of cases — as witness our above example of the would-be utilitarian with lingering deontological doubts.

In other cases, though, it seems more plausible that it is appropriate under moral uncertainty to effectively ignore theories in which one has very low credence. An illustrative example is:

> **Example 7** (Insect suffering)**.** Heiyau is sympathetic to anti-speciesism, but is almost sure that insects are not moral patients. However, she realises that assessing moral patienthood is a tricky matter, and that she *might* be wrong. In addition, she recently read that there are about 10^{19} insects alive on Earth at any one time, 11-12 orders of magnitude more than there are humans or chickens. Heiyau thinks that if she devotes almost all her time, money and energy to the project of trying to alleviate insect suffering, she would have a non-negligible chance of succeeding in alleviating a non-negligible proportion of such suffering. Since Heiyau's credence that insects are moral patients, while very small, far exceeds 10^{-11}, she decides to devote her life to this cause, even though she is almost sure that this project is completely worthless. In the process, she foregoes opportunities she would otherwise have pursued to do all manner of things that (by her lights) are highly likely to be very valuable, including poverty relief, homeless support, and campaigning on issues of climate change and immigration reform.

At least arguably, Heiyau's behaviour is inappropriate given her very low credence in the claim that insect suffering has any moral importance, even if the numbers work out so that the enormous amount that (in her view) is at stake conditional on the proposition that insects are moral patients dominates the expected value calculation.

Here are two further examples. It might similarly seem objectionably fanatical for a society to devote almost all its resources to the project of populating the universe with computer simulations of happy experiences if the society is almost sure that that project is worthless, merely on the ground that it has a very slight credence that these projects are of truly enormous value. And it might seem objectionably fanatical to prefer even an arbitrarily populous universe in which no life is more than barely worth living over a universe of a quintillion lives of untold bliss, merely on the grounds that one has slightly positive credence in a totalist approach to population ethics according to which the former universe is astronomically better than the latter (Greaves and Ord, 2017).

It strikes us as not entirely clear what makes for the difference between cases in which intuition seems to favour following the dictates of a low-credence theory

on grounds of high relative stakes, versus cases in which the intuition seems to be that allowing a very low credence theory to have any significant effect on one's decisions would be objectionably "fanatical". Reflection on the examples above, however, suggests the following hypothesis: what is objectionably "fanatical" is a very low-credence theory dictating so much that higher-credence theories have little influence on how one's *resources as a whole* are deployed, or how one's *life as a whole* is lived, leaving little space for things that one has high credence are valuable. It might not be objectionably fanatical, for instance, for Heiyau to spend *one year* working on the issue of insect suffering as one small part of a long career mostly spent on other things, or for her to allot *some minor but significant* part of her overall philanthropic portfolio to insect suffering issues. What is objectionable about Example 7 (on this hypothesis) is specifically Heiyau's *devoting her life* to the cause of insect suffering, and the concomitant implication that, with extremely high credence, she does little that is of value *with her life as a whole*. In contrast, when our would-be utilitarian refrains from killing for the greater good on grounds of her residual slight credence in deontic constraints, there is plenty of room left in her life for efficiently promoting the good.

With this hypothesis in mind, it is worth distinguishing between three (non-exhaustive) ways one might make Heiyau's predicament more precise. For definiteness, we will assume here that the disagreement point is the anti-utopia point, though we emphasise that the results we thereby obtain are sensitive to this choice.

1. Heiyau has only two options: devote her entire life to the project of minimising insect suffering, or altogether ignore the issue of insect suffering.
2. Heiyau's only *pure* options are those two, but she can also opt for any probabilistic mixture of the two.
3. Heiyau has a continuum of pure options: she can devote any proportion of her life's effort, from none at all to all of it, to addressing issues of insect suffering. No mixed options are available.

In Case 1, the intuitively correct result is that Heiyau should ignore insect suffering. The NBS is technically silent on this case as specified, since (with anti-utopia as the disagreement point) both of the available options have a Nash product of zero. There may be some sense in which the NBS will recommend that Heiyau should ignore insect suffering (rather than devote her whole life to it) for "most" locations of the disagreement point, but it is hard to give any precise content to this claim.

In Case 2, the NBS recommends the "random dictator" mixed act, according to which Heiyau devotes her life to alleviating insect suffering with a probability

equal to her credence that insects are moral patients. It seems to us at best unclear whether or not this result is intuitively preferable to one according to which Heiyau simply ignores insect suffering; the issues here are those discussed in section 7.1.

In Case 3, if we assume (implausibly, but for simplicity) that all relevant returns to effort are linear, then the NBS recommends that Heiyau devote a proportion p of her life's effort to alleviating insect suffering, where p is her credence that insects are moral patients. This result strikes us as highly plausible. In particular, a maximally "anti-fanatical" ruling that a very low-credence theory should direct *no* part of one's resource allocation in this third case — not even a very small fraction, as small as one's credence in the theory in question — would, it seems to us, err too far in the direction of deference to the higher-credence theory.

Overall, then, the bargaining-theoretic approach seems promising in its treatment of the kind of cases that give rise to the "fanaticism" worry. It seems to strike a plausible middle ground, neither ignoring lower-credence theories altogether (as does MFT), nor allowing a very low-credence theory to dictate just about all of one's (importance-weighted) decisions on grounds of relative stakes (as do some versions of MEC). While it is not entirely clear *exactly* what are the intuitive desiderata around these issues, the bargaining-theoretic approach at least does not seem to lead to any *flagrant* violations of intuition here.

We do not claim that this discussion is conclusive. In particular, here as elsewhere, precisely which option would be selected by a bargaining-theoretic approach depends on the location of the disagreement point. In general, for any option on the Pareto frontier, there is *some* possible location of the disagreement point such that the bargaining-theoretic approach selects that option. So, in particular, there is some possible location of the disagreement point such that the NBS selects an option that our intuitions deem objectionably fanatical in cases like Example 7. But, for the same reason, a bargaining-theoretic approach to moral uncertainty must anyway (that is, issues of fanaticism aside) include some prescription as to the location of the disagreement point. Given our informal discussion above, it seems reasonable to hold out hope that any such prescription that otherwise seems plausible will also generate a strong anti-fanatical tendency.

9 Independence of irrelevant alternatives

A condition of *Independence of irrelevant alternatives* is satisfied if removing options that would not have been selected anyway does not change which option(s) is selected:[29]

[29]The terminology 'independence of irrelevant alternatives' sometimes generates confusion. In particular, in the literature on social choice theory, this term is generally reserved for a

Independence of irrelevant alternatives (IIA): Let $X = (T, U, \mathcal{A}_X, u_X, p_X, d_X)$ be an n-theory bargaining problem, for some n. Suppose $\mathcal{A}_{X'} \subset \mathcal{A}_X$, and let $X' = (T, U, \mathcal{A}_{X'}, u_X, p_X, d_X)$. If $\mathcal{S}(X) \in u_X(\mathcal{A}_{X'})$, then $\mathcal{S}(X') = \mathcal{S}(X)$.

A condition of independence of irrelevant alternatives is one of the axioms in standard axiomatic justifications of the Nash bargaining solution. The axiom is not obviously non-negotiable. It has been suggested, for instance, that if the highest utility attainable by a given player changes, that might change the bargaining solution, since it changes the corresponding player's 'levels of aspiration' (Luce and Raiffa, 1957, p.133).[30]

Whatever one thinks about this, however, the following strictly weaker condition *does* seem compelling. This condition says that removing options that were *never serious options in the first place*, in the sense that they were Pareto dominated by some other available alternative, does not change which option(s) is selected:

Independence of Pareto dominated alternatives (IPDA). Let $X = (T, U, \mathcal{A}_X, u_X, p_X, d_X)$ be an n-theory bargaining problem, for some n. Suppose $\mathcal{A}_{X'} \subset \mathcal{A}_X$ is such that $\text{Pareto}(\mathcal{A}_{X'}) = \text{Pareto}(\mathcal{A}_X)$, and let $X' = (T, U, \mathcal{A}_{X'}, u_X, p_X, d_X)$. If $\mathcal{S}(X) \in u_X(\mathcal{A}_{X'})$, then $\mathcal{S}(X') = \mathcal{S}(X)$.

To motivate the condition IPDA, consider

> **Example 8** (**Philanthropic grant-making**). Abdullah administers a philanthropic grant-making program. He receives four applications. Application a_1 proposes unconditional cash transfers to poor communities. a_2 proposes a leafletting program promoting veganism. a_3 proposes to attempt local eradication of an infectious disease that mostly affects humans, but also has some adverse effects on some non-human animals. a_4 proposes to carry out migraine research by injecting rats with one after another of an enormous number of randomly selected chemicals.

---

different condition (concerning the relationship between two choice situations in which the set of available alternatives is the same but individuals' preferences over those alternatives vary), while conditions similar to the one we state here go instead by the name of 'contraction consistency'. See e.g. (Paramesh, 1973; Sen, 2017, pp.63-4 and 317-8). Here, we follow the terminology that is standard in the literature on bargaining theory.

[30] This suggestion was of course made in the usual context of bargaining among persons. However, the analogous point for the context of bargaining among moral theories seems to have roughly the same amount of merit.

Rival bargaining solutions frequently involve retaining the other axioms that are usually involved in deriving the Nash solution, but replacing the condition of IIA with some other axiom. For example, one obtains the Kalai-Smorodinsky solution by replacing the Nash IIA axiom of with an axiom of 'monotonicity'.

The relevant credences are split between a version of utilitarianism according to which only humans have moral status, and a second version of utilitarianism according to which all sentient creatures, which the relevant parties take to include rats, have moral status. The choiceworthiness levels of the four proposals a_1—a_4, by the lights of the two moral theories in question, can be represented by the numbers in the following table:

	T_1(all-species utilitarianism)	T_2(humans-only utilitarianism)
a_1	10	10
a_2	20	0
a_3	15	9
a_4	-50	0

The Pareto frontier corresponds to a_1, a_2, a_3, together with all mixtures of a_1 and a_3 and all mixtures of a_2 and a_3. There is a nontrivial question of how to choose between a_1, a_2 and a_3, but it is clear that a_4 is not a serious contender.

Very plausibly, Abdullah should decide between a_1, a_2 and a_3 (and the relevant mixtures thereof) exactly as he would have if the application a_4 had not been submitted. Insofar as IPDA secures this result, this provides some inductive evidence that that condition is desirable.

To what extent *does* IDPA secure the result in question? Well: IPDA implies that an agent facing a bargaining problem $X' = (U, \mathcal{A}_{X'}, u_X, d_X, p_X)$ can proceed just as if they faced a different bargaining problem $X = (U, \mathcal{A}_X, u_X, d_X, p_X)$, provided that the Pareto frontiers of $\mathcal{A}_X$ and $\mathcal{A}_{X'}$ coincide. However, depending on what fixes the disagreement point, it may be that making available the Pareto-dominated option a_4 shifts the disagreement point. Here, though, is a reason for choosing some way of identifying the disagreement point that does not have that feature (it is a reason, for example, for preferring the proposal $d_X = \underline{u}_{\text{Pareto}(\mathcal{A}_X)}$ over $d_X = \underline{u}(\mathcal{A}_X)$).

It is worth comparing the bargaining-theoretic treatment of this decision context to an otherwise fairly similar version of MEC: namely, MEC with structural intertheoretic comparisons. According to the latter, recall, under moral uncertainty one should maximise expected choiceworthiness when the theories' choiceworthiness functions are normalised against one another by equalising the value of some measure of their spread across the set of alternatives. The verdicts of this type of theory do in general change when Pareto-dominated options are added to or removed from the choice set, since such addition or removal changes the structural property that is used for fixing intertheoretic comparisons. In Example 8, give this type of normalisation prescription, adding a_4 to the choice

set tends to significantly reduce the utility differences between a_1, a_2 and a_3 according to T_1, while slightly increasing them according to T_2.[31]

This naively seems to suggest a way in which the bargaining-theoretic approach is superior to a structural MEC approach. But this would be too quick. In order for IDPA to secure the desired result, it is necessary that introducing or removing a Pareto-dominated alternative does not change the disagreement point. For example, as noted above, we get the desired result if $d_X = \underline{u}_{\mathrm{Pareto}(X)}$, but not if $d_X = \underline{u}_{(X)}$. Similarly, in the case of MEC, removing a Pareto-dominated option can change MEC's recommendation if intertheoretic comparisons are fixed by equalising the chosen statistical measure as computed from the whole option set, but not if the measure is computed *along the Pareto frontier only*. Thus, the bargaining-theoretic approach and a structural version of MEC perform very similarly vis-a-vis dependence on the inclusion or exclusion of Pareto dominated alternatives.

10 Small vs grand worlds

Consider again Example 5. We noted in section 6 that the Nash approach gives the intuitive verdict in this example (viz., supporting extinction risk mitigation and letting two die) provided that it uses a "grand world" model, in which the two binary choices are combined to form a single four-pure-option decision problem.

But we might also use "small world" models. In the latter, we ask *separately* (1) whether it is appropriate for Jenny to kill one, let two die or some mixture thereof, and (2) whether it is appropriate for Jenny to direct the resources to homeless support, to mitigation of extinction risk, or some mixture thereof. According to the Nash approach, the answer is then that it is appropriate for Jenny to flip a fair coin by way of resolving each of these two binary decisions. (This follows directly from Proposition 1.) The difference in relative stakes plays no role: it *cannot* play any role in this small-world modelling, since it is a matter of the relations between the two binary choices, but to undertake small-world modelling is precisely to ignore any such relations.

This means that the prescriptions of the Nash approach depend nontrivially on the choice between small-world and grand-world modelling. This is problematic. As is widely recognised in the context of decision theory, while the maximally grand-world model is the most fundamental, it is completely impractical to

[31] 'Broad' versions of structural normalisation evaluate the structural feature in question relative to 'all possible' options, rather than only those that are available in the actual decision context. These versions avoid the potential problem as it appears in Example 8. However, it also seems desirable for the identity of the preferred alternative to be independent of whether or not this set of 'all possible' options contains some Pareto-dominated alternative.

use this maximally grand-world model in practice. For a theory to be of any practical use, there must be some small-world way of approximating the way the theory treats the grand-world problem. Further, since there is no privileged small-world description, it must be the case that structurally the *same* solution applies for (at least) a wide variety of small-world descriptions of a given decision situation. Call this the condition of *small-world consistency*. Example 5 shows that the Nash approach violates this condition.[32]

This problem has no direct analog in the context of MEC. While (as discussed in section 9) it might make a difference which options are available, on MEC it does not make the above kind of difference whether we take the options to be small-world or instead grander-world ones, provided all moral theories agree that the moral issues in question are suitably separable from one another.

11 Summary and conclusions

The most popular existing approaches to moral uncertainty (other than normative externalism) are the 'maximise expected choiceworthiness' (MEC), 'my favourite theory' (MFT) and 'my favourite option' (MFO) approaches. All of these, however, seem to have serious drawbacks. These drawbacks have motivated some commentators to propose an alternative 'parliamentary' approach to moral uncertainty, according to which moral theories function in decision-making like (groups of) delegates to a parliament. The decisions of this hypothetical parliament, it has been proposed, are determined by some combination of bargaining and voting. Ab initio, at least, it seems reasonable to hope that a suitably specified parliamentary approach might simultaneously be able to avoid most or all of the pitfalls that variously plague the extant approaches, including a need for intertheoretic comparisons, complete insensitivity to relative stakes, a tendency towards 'fanaticism', and dependence on theory individuation.

The core idea of 'bargaining and voting', however, covers an enormous number of possible models. To date, none of these has been explored with any precision or depth. This makes it impossible to know whether the apparent promise of a parliamentary approach can be made good, or is instead a mere artefact of imprecision.

It would be an enormous project to categorise and investigate all possible ways of making 'bargaining and voting' precise. In this paper, we have offered a partial treatment of one segment of the parliamentary approach. Namely, we have investigated what happens when one applies the standard machinery of

[32]This problem is somewhat similar to the "problem of small worlds" discussed by e.g. (Savage, 1972; Joyce, 1999, pp.70-77, 110-113).

bargaining theory to a decision problem featuring moral uncertainty, with moral theories playing the role of parties to the bargaining process.

For concreteness, we have further focussed on the (asymmetric) Nash bargaining solution. The resulting approach does not require that unit intertheoretic comparisons be well-defined, and is robust to the individuation of moral theories. It turns out to be similar in many ways to (although genuinely distinct from) a version of MEC that employs structural intertheoretic comparisons. We will therefore take this type of MEC approach to be our main standard for comparison, in assessing the overall merits of the bargaining-theoretic approach.

On three counts, the two approaches seem to behave in essentially the same way.

First: While neither approach recognises a *primitive* notion of intertheoretic comparisons, both nonetheless exhibit a suitable-seeming sensitivity to relative stakes, via an *intra*theoretic notion of relative stakes. Slightly more specifically, when many moral issues are simultaneously under consideration, both approaches will tend to favour options that "give each theory its way" on issues that that theory deems to be higher-stakes *compared to the stakes involved in other moral issues according to the same theory.* For example, both can capture the intuition that under moral uncertainty it is appropriate to be vegetarian, even if one has only 40% credence that there is anything morally wrong with eating meat, if that 40% credence goes to theories according to which "meat is murder".[33]

Second: In both cases, it is possible to prevent choices among Pareto optimal outcomes from being affected by the presence or absence of Pareto-dominated alternatives (that is, to ensure that a condition like the one we called "Independence of Pareto Dominated Alternatives" (IDPA) is satisfied), by a judicious choice of (respectively) the disagreement point or the set of options relative for which the chosen structural measure is evaluated. In neither case is IDPA automatically guaranteed by the basic structure of the theory, *without* such judicious choice-making.

Third, regarding "fanaticism". It is well recognised that a version of MEC employing exogenous or content-based intertheoretic comparisons can lead to problematically "fanatical"-seeming verdicts in some cases involving tiny credence in enormous stakes. In some such cases, MEC defers completely to the low-credence theory, in a way that intuitively seems problematic. We have tentatively suggested that what distinguishes such problematic cases of "fanaticism"

[33]There is a difference of detail that some readers might consider important. This is that in a structural MEC approach, one does end up recognising intertheoretic unit comparisons (albeit constructed rather than primitive ones). If even constructed intertheoretic unit comparisons were simply incoherent, this would count against the structural MEC approach, and in favour of the bargaining-theoretic approach. Our own view, however, is that this "incoherence objection" is misguided.

from cases in which it is simply *appropriate* to defer to a low-credence theory is whether or not the deference in question is essentially all-encompassing (it would "take over one's life", or "use all of society's resources", in the service of highly improbable moral ideals). If this is correct, a bargaining-theoretic approach is highly resistant to the problematically "fanatical" conclusions. Heuristically, the reason for this is that a low-credence theory can never have enough bargaining power to commandeer anything close to all resources; it is in the nature of bargaining that a given party has to *offer something substantial to the other party/ies* before the party in question can get what it wants (and, in case of very weak bargaining power, something *very* substantial). On the other hand, similar things can again be said about a structural (as opposed to a primitive or content-based) version of MEC, which similarly prevents one theory from simply "shouting louder than all other theories" across the board, and similarly requires intertheoretic "give and take" across moral issues (Cotton-Barratt et al., 2020, p.73).

On two further counts, the two approaches differ from one another in interesting and potentially important ways, though it seems to us unclear at best which approach is superior on each count.

First, regarding the notion of compromise between moral theories more generally (that is, setting aside the specific concerns that arise in cases of extreme credence). In many decision situations, if one is unsure on moral grounds which is better of two quite different alternatives, a natural inclination is to seek some option that is intermediate between those two alternatives (for example, splitting funding between two very different projects, rather than going all-in on either one to the exclusion of the other). Both the MEC approach and the Nash approach are able to rationalise this inclination, though in ways that are potentially importantly different. More specifically, while they can both rationalise a strict preference for an intermediate *pure* option over a *pure* option that more completely defers to one or the other moral theory, they differ in their treatment of mixed options according to which (for instance) it is guaranteed that one or the other theory will end up getting its most-preferred pure option, but *ex ante* there is some positive probability that it could be either theory that "wins". The Nash approach can strictly prefer such a mixed option over all of the pure options it assigns positive probability to, while the MEC approach (since it satisfies vNM Independence) cannot.

Second: both theories exhibit some undesirable dependence on arbitrary modelling choices, though the details are different the two cases. On the one hand, the Nash approach (but not the MEC approach) suffers from a "problem of small worlds": it can make a significant difference to the verdict of the Nash approach whether one chooses a smaller- or a grander-world model of one's decision problem (section 10). This is problematic, since any such choice (short of the impractical maximally grand-world model) seems arbitrary. On the other hand, the standard of intertheoretic comparisons that is used in a structural

MEC approach depends at least on what the full set of relevantly available options is – leading to discussions of whether one should take the option set to be "all actually available options" or "all conceivable options", and precisely what either of these would even mean (Cotton-Barratt et al., 2020, pp.68-70). Further, the "variance normalisation" approach, which is arguably the best structural method of fixing intertheoretic comparisons overall (Cotton-Barratt et al., 2020), requires a *measure* over the space of options, for the purpose of defining the variance of their choiceworthiness (*ibid.*, pp.69–70). Any particular choice of such a measure seems arbitrary.

The general picture from the investigation in this paper (as far as that goes) therefore seems to be that while the bargaining-theoretic approach is not obviously superior to an MEC approach — contra, perhaps, the hopes of many of the advocates of a 'parliamentary model of moral uncertainty' — neither is it clearly inferior. Nor is it equivalent; there are various potentially important differences which, on further examination, might be discovered to be advantages of one or the other approach. Our primary verdict is therefore that a bargaining-theoretic approach to moral uncertainty should be firmly on the table, whereas it has previously been (to our knowledge) entirely absent from the peer-reviewed academic literature.

If it is indeed to be on the table, the bargaining-theoretic approach is worthy of further investigation. This paper has tried to lay the groundwork for that project, and has conducted some preliminary investigation. But we have explored only one particular model of bargaining (the Nash bargaining solution), and we have not made any attempt to integrate a bargaining stage into a more complex overall model of a parliament-like procedure (as informal discussions of "the parliamentary model" sometimes suggest). We have also left largely open what might in practice settle the "disagreement point" that is essential to a bargaining-theoretic approach, and what results such an approach delivers in practice for various resolutions of that choice point. If the particular approach that we have explored here is found wanting, we hold out hope that others might be able to do better, elsewhere in the nearby space.

References

Anbar, Dan & Kalai, Ehud (1978). A one-shot bargaining problem. *International Journal of Game Theory*, 7(1):13–18.

Anbarci, Nejat & Sun, Ching-Jen (2013). Asymmetric Nash bargaining solutions: A simple Nash program. *Economics Letters*, 120(2):211–214.

Arntzenius, Frank (2008). No regrets, or: Edith piaf revamps decision theory. *Erkenntnis*, 68(2):277–297.

Binmore, Ken, Rubinstein, Ariel, & Wolinsky, Asher (1986). The Nash bargaining solution in economic modelling. *The RAND Journal of Economics*, 17(2):176–188.

Bostrom, Nick (2009). Moral uncertainty - towards a solution? Available at https://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html.

Broome, John (2012). *Climate matters*. W. W. Norton and Company.

Bykvist, Krister (2014). Evaluative uncertainty, environmental ethics, and consequentialism. In: *Consequentialism and environmental ethics*, A. Hiller, R. Ilea, & L. Kahn, ed., pages 122–135. Routledge.

Bykvist, Krister (2017). Moral uncertainty. *Philosophy Compass*, 12(3).

Bykvist, Krister (2018). Some critical comments on Zimmerman's 'Ignorance and moral obligation'. *Journal of Moral Philosophy*, 15(4):383–400.

Bykvist, Krister & Olson, Jonas (2011). Against the Being For account of normative certitude. *Journal of Ethics and Social Philosophy*, 6.

Chun, Youngsub & Thomson, William (1990). Nash solution and uncertain disagreement points. *Games and Economic Behavior*, 2(3):213–223.

Cotton-Barratt, Owen, MacAskill, William, & Ord, Toby (2020). Statistical normalization methods in interpersonal and intertheoretic comparisons. *The Journal of Philosophy*, 117(2):61–95.

Dagan, Nir, Volij, Oscar, & Winter, Eyal (2002). A characterization of the Nash bargaining solution. *Social Choice and Welfare*, 19(4):811–823.

de Clippel, Geoffrey (2006). An axiomatization of the Nash bargaining solution. *Social Choice and Welfare*, 29(2):201–210.

Dorsey, Dale (2012). Subjectivism without desire. *Philosophical Review*, 121(3):407–442.

Gibbard, Allan (2005). Truth and correct belief. *Philosophical Issues*, 15:338–350.

Gracely, Edward J. (1996). On the noncomparability of judgments made by different ethical theories. *Metaphilosophy*, 27(3):327–332.

Graham, Peter A. (2010). In defense of objectivism about moral obligation. *Ethics*, 121(1):88–115.

Greaves, Hilary & Ord, Toby (2017). Moral uncertainty about population axiology. *Journal of Ethics and Social Philosophy*, 12(2):135–167.

Gustafsson, Johan E. & Torpman, Olle (2014). In defence of my favourite theory. *Pacific Philosophical Quarterly*, 95(2):159–174.

Harman, Elizabeth (2011). Does moral ignorance exculpate? *Ratio*, 24(4):443–468.

Harsanyi, John C. (1956). Approaches to the bargaining problem before and after the theory of games: A critical discussion of Zeuthen's, Hicks', and Nash's theories. *Econometrica*, 24(2):144.

Harsanyi, John C. & Selten, Reinhard (1972). A generalized Nash solution for two-person bargaining games with incomplete information. *Management Science*, 18(5):80–106.

Hedden, Brian (2016). Does MITE make right? On decision-making under normative uncertainty. *Oxford Studies in Metaethics*, 11:102–28.

Howard-Snyder, Frances (2005). It's the thought that counts. *Utilitas*, 17(3):265–281.

Hudson, James (1987). The diminishing marginal value of happy people. *Philosophical Studies*, 51(1):123–137.

Joyce, James (1999). *The foundations of causal decision theory*. Cambridge University Press.

Kalai, Ehud (1977a). Nonsymmetric Nash solutions and replications of 2-person bargaining. *International Journal of Game Theory*, 6(3):129–133.

Kalai, Ehud (1977b). Proportional solutions to bargaining situations: Interpersonal utility comparisons. *Econometrica*, 45(7):1623.

Kalai, Ehud & Smorodinsky, Meir (1975). Other solutions to Nash's bargaining problem. *Econometrica*, 43(3):513–518.

Kolodny, Niko & Macfarlane, John (2010). Ifs and oughts. *The Journal of Philosophy*, 107(3):115–143.

Lensberg, Terje (1988). Stability and the Nash solution. *Journal of Economic Theory*, 45(2):330–341.

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

Luce, Robert Duncan & Raiffa, Howard (1957). *Games and decisions: Introduction and critical survey*. New York: Dover Publications Inc.

MacAskill, William (2013). The Infectiousness of Nihilism. *Ethics*, 123(3):508–520.

MacAskill, William (2016). Normative uncertainty as a voting problem. *Mind*, 125(500):967–1004.

MacAskill, William, Bykvist, Krister, & Ord, Toby (2020). *Moral uncertainty*. Oxford University Press.

MacAskill, William & Ord, Toby (2018). Why maximize expected choice-worthiness? *Nous*.

Mason, Elinor (2013). Objectivism and prospectivism about rightness. *Journal of Ethics and Social Philosophy*, 7(2):1–22.

Moore, George E. (1903). *Ethics: and the nature of moral philosophy*. Clarendon Press.

Moore, George E. (1912). *Principia Ethica*. Cambridge University Press.

Muthoo, Abhinay (1999). *Bargaining theory with applications*. Cambridge University Press.

Nash, John F. (1950). The bargaining problem. *Econometrica*, 18(2):155–62.

Nash, John F. (1953). Two-person cooperative games. *Econometrica*, 21(1):128.

Newberry, Toby & Ord, Toby (2021). The parliamentary approach to moral uncertainty. FHI technical report 2021-2. Available online at https://www.fhi.ox.ac.uk/wp-content/uploads/2021/06/Parliamentary-Approach-to-Moral-Uncertainty.pdf (accessed 15 December, 2022).

Nissan-Rozen, Ittay (2015). Against moral hedging. *Economics and Philosophy*, 31(03):349–369.

Olsen, Kristian (2017). A defense of the objective/subjective moral ought distinction. *The Journal of Ethics*, 21(4):351–373.

Paramesh, Ray (1973). Independence of irrelevant alternatives. *Econometrica*, 41(5):987–91.

Parfit, Derek (2011). *On what matters: Volume Two*. Oxford University Press.

Perles, Micha A & Maschler, Michael (1981). The super-additive solution for the nash bargaining game. *International Journal of Game Theory*, 10(3):163–193.

Portmore, Douglas W. (2011). The teleological conception of practical reasons. *Mind*, 120(477):117–153.

Prichard, Harold A. (1933). Duty and ignorance of fact. *Philosophy*, 8(30):226–228.

Riedener, Stefan (2020). An axiomatic approach to axiological uncertainty. *Philosophical Studies*, 177(2):483–504.

Riedener, Stefan (2021). *Uncertain Values: An Axiomatic Approach to Axiological Uncertainty*. De Gruyter.

Ross, Jacob (2006). Rejecting ethical deflationism. *Ethics*, 116(4):742–768.

Ross, William David (1930). *The right and the good*. Oxford University Press.

Ross, William David (1939). *Foundations of ethics*. Oxford: Clarendon Press.

Russell, Bertrand (1966). The elements of ethics. In: *Philosophical Essays*, B. Russell, ed., pages 13–59. New York: Simon and Schuster.

Savage, Leonard J. (1972). *The foundations of statistics*, (2nd ed.). Dover.

Sen, Amartya Kumar (2017). *Collective choice and social welfare (expanded edition)*. London: Penguin.

Sepielli, Andrew (2010). *Along an imperfectly-lighted path: Practical rationality and normative uncertainty*. PhD thesis, Rutgers University.

Sepielli, Andrew (2013). Moral uncertainty and the principle of equity among moral theories. *Philosophy and Phenomenological Research*, 86(3):580–589.

Sepielli, Andrew (2016). Moral uncertainty and fetishistic motivation. *Philosophical Studies*, 173(11):2951–2968.

Sepielli, Andrew (2017). How moral uncertaintism can be both true and interesting. *Oxford Studies in Normative Ethics*, 7.

Snowden, James (2019). Should we give to more than one charity? In: *Effective altruism: Philosophical issues*, H. Greaves & T. Pummer, ed. Oxford: Oxford University Press.

Tarsney, Christian (2019). Normative uncertainty and social choice. *Mind*, 512:1285–308.

Tarsney, Christian J (2021). Vive la difference? structural diversity as a challenge for metanormative theories. *Ethics*, 131(2):151–182.

Thomson, Judith Jarvis (1986). *Rights, restitution, and risk: Essays in moral theory*. Harvard University Press.

Weatherson, Brian (2014). Running risks morally. *Philosophical Studies*, 167(1):141–163.

Weatherson, Brian (2019). *Normative externalism*. Oxford University Press.

Zeuthen, Frederik (1930). *Problems of monopoly and economic warfare*. Routledge.

Zimmerman, Michael J. (2009). Living with uncertainty: The moral significance of ignorance. *Analysis*, 69(4):785–787.

Zimmermann, Michael J. (2006). Is moral obligation objective or subjective? *Utilitas*, 18(4):329–361.