A bargaining-theoretic approach to moral uncertainty

Owen Cotton-Barratt, Hilary Greaves

Global Priorities Institute | August 2019

GPI Working Paper No. 4-2019

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

A bargaining-theoretic approach to moral uncertainty

Hilary Greaves and Owen Cotton-Barratt*

August 9, 2019

Abstract

This paper explores a new approach to the problem of decision under relevant moral uncertainty. We treat the case of an agent making decisions in the face of moral uncertainty on the model of bargaining theory, as if the decision-making process were one of bargaining among different internal parts of the agent, with different parts committed to different moral theories. The resulting approach contrasts interestingly with the extant "maximise expected choiceworthiness" and "my favourite theory" approaches, in several key respects. In particular, it seems somewhat less prone than the MEC approach to 'fanaticism': allowing decisions to be dictated by a theory in which the agent has extremely low credence, if the relative stakes are high enough. Overall, however, we tentatively conclude that the MEC approach is superior to a bargaining-theoretic approach.

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 pack waterproofs, 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.

*The idea to investigate applying the Nash bargaining solution to the problem of moral uncertainty, and that the result might be distinctive in particular in cases like Example 7 below, originated with OCB. HG led the investigation of the remainder of the issues discussed herein, and the writing of the paper.

Suppose Alice packs the waterproofs 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 bulky raingear 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 ande mpirical 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). 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 even a person who does not care about morality in some sense ought to play it safe in suitable contexts of moral uncertainty; this consideration mitigates against the view that the ought in question is a rational rather than a moral one. For a survey of the issues, see (Bykvist, 2017, Section 2).

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. Say that 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 sets out a bargaining-theoretic approach. Section 4 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 5–8, we use these results to analyse the performance of this approach vis-a-vis (respectively) issues of dependence of results on the presence 'irrelevant' alternatives (section 5), the problem of small worlds (section 6), moral risk aversion (section 7), and sensitivity to relative stakes and (relatedly) fanaticism (section 8). Section 9 summarises, and compares the merits of a bargaining-theoretic approach with those of MEC. Our own tentative conclusion is that overall the bargaining-theoretic approach is inferior to at least one version of MEC, but this is not completely clear-cut.

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 EU 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 moral hypotheses, 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 (Gracely, 1996, p. 185; Broome, 2012, p.185; Gustafsson and Torpman, 2014; Nissan-Rozen, 2015; Hedden, 2016), and that it leads to problematic forms of "fanaticism" (Ross, 2006, pp.765-7; MacAskill and Ord, 2018, section 7(v)); see also section 8.2, below).

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).

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. In some examples this stakes-insensitivity appears troubling.[4]

A different set of angles on the problem is suggested by the thought that decision under moral uncertainty is analogous to the problem of group choice in the face of disagreement: there is conflict between internal parts of the agent, but only the agent as a whole can act. Existing work on problems of group choice includes the literatures on voting and on bargaining theory.

There has been some exploration of the application of voting theory to the problem of moral uncertainty, specifically for the case in which moral theories exhibit only ordinal structure (MacAskill, 2016; Tarsney, 2018). However, whatever its merits for the purely ordinal case, a voting-theoretic approach seems inappropriate for the treatment of moral theories that do come already equipped with cardinal choiceworthiness structure (Tarsney, nd, pp.2-3).

Unlike voting theory, bargaining theory does make use of the cardinal structure of agents' utility functions. To our knowledge, however, the application of bargaining theory to the problem of moral uncertainty has not yet been explored. The aim of the present paper is to conduct such an exploration.

3 The bargaining-theoretic approach

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. There is therefore an open question precisely which Pareto superior alternative (if any) the players will settle on. Bargaining theory addresses this question. For the

[4]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 8.

application to moral uncertainty, we will take the 'players' to be moral theories that our agent has nonzero credence in (rather than people, as in the ordinary application).

3.1 The model

An *n-theory bargaining problem* is 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.
- $U = U_1 \times \ldots \times U_n$, 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.[5]
- $\mathcal{A}_X$ is a set of available options, such that $u_X(\mathcal{A}_X)$ is bounded and convex.[6]
- $p_X = (p_1, ..., p_n)$ is the agent's credence distribution over T.
- $d_X \in U$ is the disagreement point.

The final item on this list requires some comment. In the usual application of bargaining theory, the disagreement point represents what would happen if the parties to the bargaining procedure fail to reach agreement.[7] This disagreement

[5] 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. This includes 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, 2015, pp.69-72;Riedener, 2018, p.9;Tarsney, nd).

[6] 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 $u(\mathcal{A}_X)$ contains all relevant information about $\mathcal{A}_X$. Accordingly, we will sometimes conflate an act with its image under u.

The convexity of $u(\mathcal{A}_X)$ is guaranteed if $\mathcal{A}_X$ is closed under probabilistic mixture.

[7] For some discussion of the subtleties of this notion, see Binmore et al. (1986).

point is crucial to the outcome of a bargaining procedure, since (i) a given candidate agreement is a contender only if it is at least a weak Pareto improvement over the disagreement point, and further (ii) a party that stands to gain little from agreement (relative to disagreement) has a stronger bargaining position.

In the application to moral uncertainty, it is unclear how the disagreement point should be identified. 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'. The task is simply to select some disagreement point such that bargaining theory with that choice of disagreement point supplies a satisfactory metanormative theory. Some reasonably natural suggestions 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 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 $\mathcal{A}_X$.[8]
- $d_X = \mathrm{RD}_X$, the random dictator point, where, for each i, the act that is highest-ranked by T_i is selected with probability p_i.[9]
- d_X is 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.

It is sometimes convenient to work with real-valued coordinate systems for the utility space U. To this end, let $\mathcal{F}_n$ be the set of all n-tuples of affine maps $(f_1, ..., f_n)$ such that $\forall i = 1, ..., n, f_i : U_i \to \Re$. Relative to any $f \in \mathcal{F}_n$, we can associate any point in utility space $u \in U$ (respectively, any option $a \in \mathcal{A}$) with an element $f(u) \in \Re^n$ (respectively, $f \circ u(a) \in \Re^n$). Call these functions

[8] 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:

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

[9] 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.

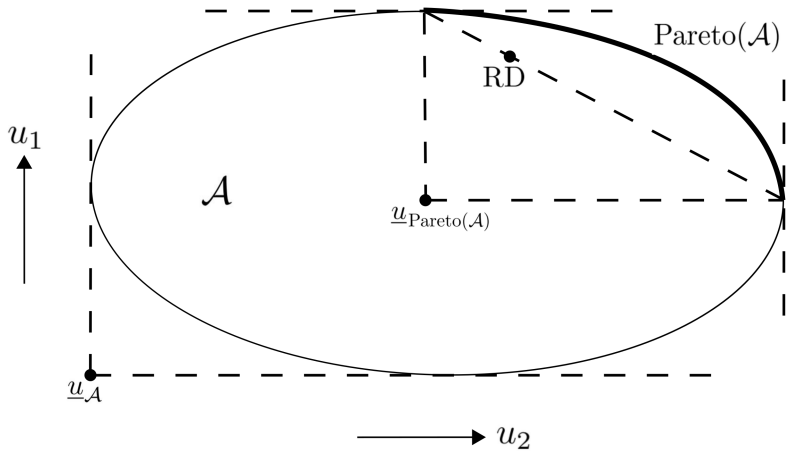


Figure 1: *An arbitrary set of available options $\mathcal{A}$, with the Pareto frontier* Pareto($\mathcal{A}$)*, anti-utopia point $\underline{u}_{\mathcal{A}}$ and random dictator point* RD *(for the case $p_{1,} = 0.25, p_2 = 0.75$) labelled.*

f admissible coordinate systems for the utility space U. For any $f \in \mathcal{F}_n$, say that the real-valued compound function $f \circ u$ *cardinally represents* the players' utility functions.

We will sometimes appeal to such numerical representations for purposes of illustration. However, such a real-valued representation involves surplus structure, since we do not assume any privileged zero point for any theory's utility function, or any privileged standard of intertheoretic level or unit comparison. The abstract representation in terms of the coordinate-free affine space U is is therefore more fundamental.

3.2 The asymmetric Nash Bargaining Solution

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 a function $\mathcal{S}$ from bargaining problems X to solutions $s \in \mathcal{A}_X$.

The standard solution function is the *Nash bargaining solution.*[10] For the application to moral uncertainty, since the (generally unequal) distribution of cre-

[10]For example, the graduate textbook by Muthoo (1999) discusses exclusively the Nash solution. Strategic justifications of the Nash bargaining solution based on a one-shot demand game are give 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).

As noted below, for the purposes of this paper we are primarily interested in a variant of

dences across theories must make a difference, we need the asymmetric version of this solution:[11]

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

We note in passing that the solution (1) 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. Recall that violation of such a clone-independence condition was one of the key problems facing the "my favourite theory" approach to moral uncertainty.

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 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.

---

the Nash solution: the *asymmetric* Nash solution. Justifications for this variant are given in (Harsanyi and Selten, 1972; Kalai, 1977a; Anbar and Kalai, 1978; Anbarci and Sun, 2013).

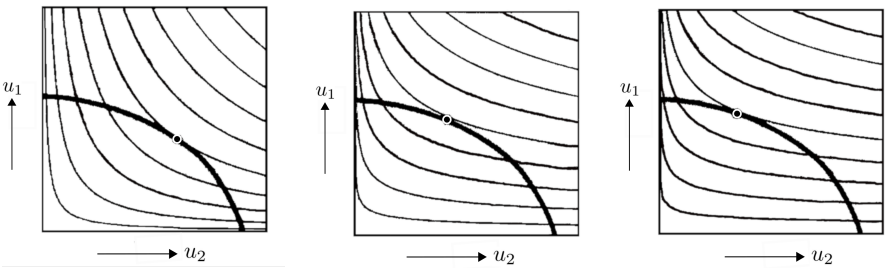
The Nash solution is of course not the only possibility. Other notable alternatives include the Kalai-Smorodinsky solution (Kalai and Smorodinsky, 1975), the Mascher-Perles solution (Perles and Maschler, 1981), and the "proportional" solutions of Kalai (1977b). Investigation of these other solutions in the context of moral uncertainty is beyond the scope of this paper. (The utilitarian optimum may also be considered a solution to a bargaining problem; in the context of moral uncertainty this of course corresponds to maximising expected moral value.)

[11] The usual (symmetric) Nash bargaining solution omits the exponents p_i.

Strictly speaking, equation (1) 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

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

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 $f \in \mathcal{F}_n$ makes no difference, we will usually omit explicit mention of f.



$p_1 = p_2 = 0.5$ $\quad$ $p_1 = 0.75, p_2 = 0.25$ $\quad$ $p_1 = 0.9, p_2 = 0.1$

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

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

4.1 Two pure options

Suppose first that there are only two pure options, a_1 and a_2. Thus $\mathcal{A} = \text{Conv}(a_1, a_2)$, 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 $\text{Pareto}(\mathcal{A}) = \mathcal{A}$. 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 solution $\text{NBS}(X)$ 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 $d_X = \underline{u}_X$. 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* $\text{NBS}(X)$ *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* $\text{Pareto}_{\mathcal{A}_X}$ *as one moves towards the global maximum* RD_X *from either side.*[12]

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

$$\frac{d}{d\lambda}\left(\Pi_{i=1,\ldots,n}(u_i(\lambda a_1 + (1-\lambda)a_2))^{p_i}\right) = 0 \qquad (2)$$

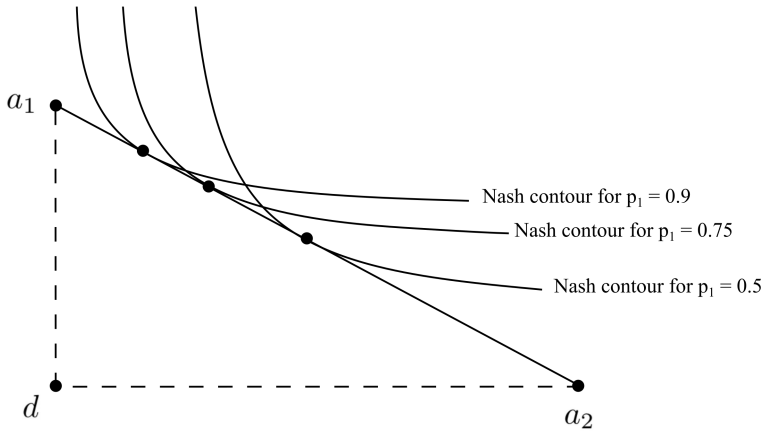


Figure 3: *Illustration of Proposition 1.*

Suppose now that X is not canonical. We can nonetheless define the *canonical problem corresponding to* X, $\text{Can}(X)$, as follows: Let $d_{\text{Can}(X)} = d_X$, and let $\mathcal{A}_{\text{Can}(X)}$ contain all and only points on the extension of the straight line a_1a_2 that are weak Pareto improvements over 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* $\text{NBS}(X)$ *is that point on the Pareto frontier* $\text{Pareto}_{\mathcal{A}_X}$ *that is closest to the random dictator point* RD_{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 $\mathcal{A}_X$ already contains $\text{RD}_{\text{Can}(X)}$, so that $\text{NBS}(X) = \text{RD}_{\text{Can}(X)}$. This can happen whether (e) the disagreement point is a Pareto improvement over the anti-utopia point ($d_X \gg \underline{u}_X$), (a) 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, we might have $\text{RD}_{\text{Can}(X)} \notin \mathcal{A}_X$,in which case it follows from Proposition 3 that the NBS to X 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 (b) and (d)).

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 (2).

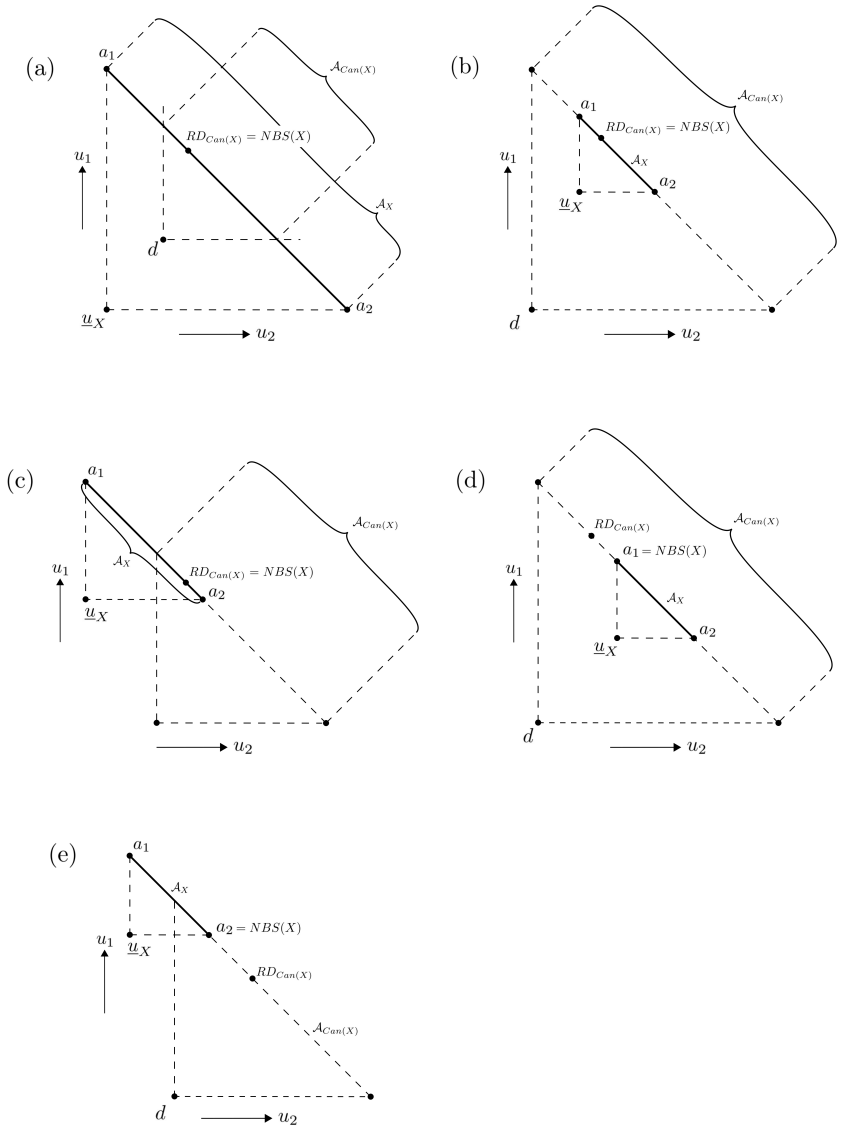


Figure 4: *The Nash bargaining solution for a variety of non-canonical two-theory, two-pure-option bargaining problems. In all cases, the Nash solution* $\mathrm{NBS}(X)$ *to the problem* X *is the available act that is closest to* $RD_{\mathrm{Can}(X)}$*, the Nash solution to the corresponding canonical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

4.2 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 4.1) 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 $u(a) = (1, 0), u(b) = (0.7, 0.7), u(c) = (0, 1)$, that $\mathcal{A} = \mathrm{Conv}(a, b, c)$, and that $d = \underline{u}$. Then it is straightforward to show that for any value of p_1 between 0.3 and 0.7, the NBS selects b over any other pure or mixed option.

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

5 Independence of irrelevant alternatives

Consider the following condition:[13]

> **Independence of irrelevant alternatives (IIA):** Suppose $X = (T, U, \mathcal{A}_X, u_X, p_X, d_X)$ is 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

[13] 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 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.

the bargaining solution, since it changes the corresponding player's 'levels of aspiration' (Luce and Raiffa, 1957, p.133).[14]

Whatever one thinks about this, however, the following strictly weaker condition *does* seem compelling:

Independence of Pareto dominated alternatives (IPDA). Suppose $X = (T, U, \mathcal{A}_X, u_X, p_X, d_X)$ is an n-theory bargaining problem, for some n. Suppose $\mathcal{A}_{X'}$ 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 this condition, consider

> **Example 5** (**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.
>
> 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

[14] 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 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.[15]

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, 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 — perhaps the range (Lockhart, 2000, pp.84ff.; Sepielli, 2013), or the standard deviation (Cotton-Barratt et al., nd) — 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 5, 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.[16]

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: we could, for example, have $d_X = \underline{u}_{\text{Pareto}(X)}$, but not $d_X = \underline{u}_{(X)}$. Similarly, in the case of MEC, introducing or removing a Pareto-dominated option will make no difference if our procedure is to equalise the measure of spread *along the Pareto frontier only*. Thus, the bargaining-theoretic approach and a structural version of MEC perform very similarly vis-a-vis IDPA.

However, we could take our lead from the above discussion of the disagreement point in the Nash approach (fn. 15), and fix intertheoretic comparisons by

[15]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 $d_X = \underline{u}(\mathcal{A}_X)$).

[16]'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 5. 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.

equalising the variance only among options on the Pareto frontier. Given the availability of this option, this type of MEC approach seems roughly on a par with the Nash approach, vis-a-vis the considerations discussed in this section.

6 Small vs grand worlds

Consider

> **Example 6** (**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 (say) $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.[17]

What, according to the Nash approach, is it appropriate for Jenny to do?

The answer to this depends on whether Jenny's predicament is modelled as two "small-world" decision problems, or instead as one "grander-world" decision problem. In the former case, 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.

For the "grand world" model, we instead take it that Jenny faces a single choice between four pure options (and mixtures among them): kill one and support the homeless, etc. In this case, the stipulated facts about (intratheoretic) relative

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

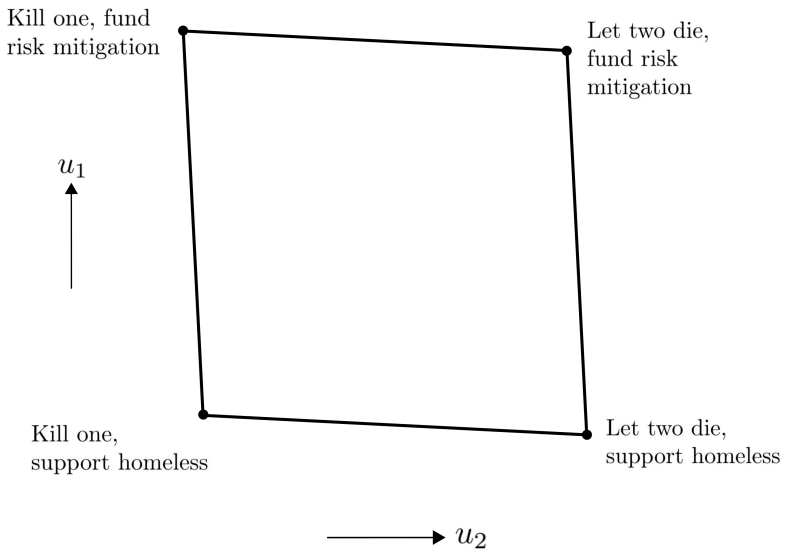


Figure 5: *Two binary choices.*

stakes *are* important. 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 this "good compromise" option. For example, this happens if the choiceworthiness functions are as in the following table and the disagreement point is the anti-utopia point $\underline{u}_{\mathrm{Pareto}(\mathcal{A}_X)}$; cf. figure 5.

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

This example brings out two important points about the behaviour of the Nash approach.

The first point is 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 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 6 shows that the Nash approach violates this condition. [18]

This problem has no direct analog in the context of MEC. While (as discussed in section 5) 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.

The second point illustrated by Example 6 concerns issues of sensitivity to relative stakes; we return to it in section 8.

7 Moral risk aversion

7.1 Moral risk aversion with respect to empirically expected utility

By definition, the MEC approach is risk neutral, in the context of moral risk, with respect to choiceworthiness – whatever cardinalisations are used for the choiceworthiness function (that is, however the intratheoretic unit comparisons are fixed, for each theory).[19] However, given only the constraint that a given theory's choiceworthiness function must correctly represent the theory's ordering of (risk-free) outcomes, these cardinalisations are underdetermined.

Arguably the most natural cardinalisation, and anyway the one we assume for the purposes of this paper, appeals to the individual theories' orderings of empirical gambles (rather than only outcomes). Specifically, we identify each theory's choiceworthiness function with its vNM representation of choiceworthiness (cf. section 3). Given this stipulation, the resulting MEC approach treats moral and empirical risk alike.

Matters are different for the Nash approach. The (asymmetric) Nash bargaining solution, recall, corresponds to maximisation of the (asymmetric) Nash product

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

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

[19] We use the terms 'uncertainty' and 'risk' interchangeably.

But this is equivalent to maximisation of *morally expected log empirically expected utility*,

$$E_m \log E_e u(a) := \sum_i p_i \log(u_i(a) - d_i).$$

Since $\log x$ is a concave function, this amounts to risk aversion, in the context of *moral* uncertainty ('risk'), with respect to empirically expected choiceworthiness (more precisely: with respect to empirically expected *gain* in choiceworthiness, relative to the disagreement point).

This difference is brought out in the following example. Suppose there are two states of nature (s_1 and s_2), and two moral theories (T_1 and T_2), that each of the resulting four theory-state combinations has probability 1/4, and that the vNM choiceworthiness levels are as in the following table:

(u_1, u_2)	s_1	s_2
a	(2,1)	(1,2)
b	(2,1)	(2,1)

Assuming further that the appropriate intertheoretic comparisons are as in this table, the expected choiceworthiness for the two acts is identical, so the MEC approach recommends indifference between a and b.

However, there is a sense in which act a is morally safe, whereas b is morally risky: a has an empirically expected utility of 1.5 conditional on either moral theory, whereas b has an empirically expected utility of 2 (resp. 1) conditional on T_1 (resp. T_2). Correspondingly, the Nash approach tends to select a over b. For instance, if the disagreement point is $(0, 0)$, then the Nash products of a and b are ~ 0.58 and 0.5 respectively.

This type of aversion to moral risk might superficially seem desirable. However, on further reflection (we will argue), it seems actively undesirable.

7.2 Preference for compromise

The phenomoneon of moral risk aversion with respect to empirically expected choiceworthiness has the following implication.

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 7 (**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 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 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 in a moral theory T_2 according to 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 $\lambda = 1$).

Clearly, the MFT approach to moral uncertainty cannot justify this pattern of preferences. The MEC approach to moral uncertainty can justify it *provided* the choiceworthiness functions 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). However, this condition may not hold. It is at least arguable that the relevant choiceworthiness functions are linear in λ even in some cases in which people feel strongly inclined to select a compromise option (Snowden, 2019). When the choiceworthiness functions have that feature, 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).

Due to its moral risk aversion, 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.

This preference for nontrivially splitting the pot might initially seem a good-making feature of the Nash approach to moral uncertainty. There is a twist, however.

As we flagged above (section 7.1), the preferences in question arise from moral-risk aversion with respect to empirically *expected* utility. Thus, for example, in Example 7, the Nash solution corresponds not only to the empirically pure option $p(A, B)$ noted above, but also to any empirically mixed option that has the same empirically expected utility according to each of the moral theories T_1, T_2. For example, the Nash approach is indifferent between (i) the empirically pure option $p(A, B)$ and (ii) the empirically mixed option according to which the whole pot goes to intervention A (resp. intervention B) with probability p (resp. $(1 - p)$).

While some sense of moral risk aversion might seem desirable, this specific phenomenon of moral risk aversion with respect to empirically expected utility seems at best dubious. We do not think, for instance, that people who share James' preference for splitting the pot will generally be just as happy to flip a (suitably weighted) coin to decide where to send the whole pot as they are to actually split the pot. This suggests that the way in which MEC treats Example 7 is superior: either the cases in question are not after all ones of linear vNM choiceworthiness functions, or (perhaps contrary to untutored intuition) splitting the pot in these cases is not after all preferable to an extremal allocation.[20]

7.3 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"

[20] One might mount a defence of randomisation on grounds of universalisability: if splitting is desirable at the population level, then it is indeed a matter of indifference whether each individual splits or instead randomises, as the law of large numbers means that at the population level the outcome will almost certainly be a form of splitting even in the latter case. However, this would be using an answer to the problem of normative uncertainty to help resolve a coordination problem. It seems to us preferable to first address the problem of normative uncertainty in the abstract, and only later to consider which algorithms have good practical consequences when implemented.

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 7. Let $L = N = A$, and let $M = B$. Then vNM Independence would require that if $A \simeq B$, then also $pB + (1-p)A \simeq A$. As we saw in section 7.2, however, the mixed act $pB+(1-p)A$ can have a strictly higher Nash product than either pure option A, B, including when $A \simeq B$; thus $pB + (1-p)A \succ A$, contradicting Independence.

Meanwhile, any version of MEC satisfies vNM Independence (on either interpretation of the condition). 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.[21]

8 Sensitivity to relative stakes and fanaticism

8.1 Inter- and intratheoretic senses of 'sensitivity to relative stakes'

In at least some cases, we might want, or expect, that under moral uncertainty what it is appropriate to do depends 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.

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 fine, 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 (see e.g.(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.

As we noted above (fn. 4), there is a tension between sensitivity to relative stakes on the one hand, and absence of intertheoretic comparisons on the other. At

[21] We do not intend violation of Independence in itself to constitute an 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.

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, 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 a typical choice*, 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. We have already seen the key point, in section 6. If there are only two pure options, then (modulo issues of how the disagreement point is fixed) there is no room for intratheoretic relative stakes, and the Nash approach simply selects the random dictator point of the corresponding canonical problem. However, when there are more than two pure options, and relatedly the Pareto frontier is strictly convex, then intratheoretic relative stakes can be highly decision-relevant. For instance, in Example 4, the Nash approach prefers the option B that is almost-optimal according to both theories, depsite not being optimal according to either theory. Relatedly, Example 6 shows how when there are several moral issues under simultaneous consideration (that is, when decisions are modelled in a "grand world" way), the Nash approach tends to choose on each moral issue in accordance with the theory according to which the particular issue under consideration is (in an intratheoretic sense) higher-stakes.

8.2 Fanaticism

It is not completely clear what counts as a good-making vs. a bad-making feature of a metanormative theory vis-a-vis issues of sensitivity to relative stakes. In some cases (such as, arguably, the eample of vegetarianism sketched above), it seems that we want sensitivity to relative stakes. This might also be true in some cases in which one has quite low credence in the theory according to which stakes are high. 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.

However, there may be limits to this. In particular, in some cases in which an option that is preferred by a moral theory in which the agent has *very* low credence is strongly dispreferred by much higher-credence theories, it can sometimes seem objectionably 'fanatical' for the low-credence theory to dictate decisions under moral uncertainty. Consider, for example:

> **Example 8** (Insect suffering)**.** Heiyau is sympathetic to anti-speciesism, but is almost sure that insects are not sentient. However, she realises that assessing sentience is a tricky matter, and that she *might* be wrong. She is also aware that there are truly enormous numbers of insects: 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. Since Heiyau's credence that insects are sentient, while very small, far exceeds 10^{-11}, she decides to devote almost all her time, money and energy to the project of trying to alleviate insect suffering, even though she is almost sure that this project is completely worthless.
>
> At least arguably, Heiyau's behaviour seems inappropriate given her low credence in insect suffering, and this despite the difference in relative stakes.[22]

Despite the apparent force of these examples, it is not completely clear that fanaticism is objectionable.[23] It is also not clear that it is reasonably avoidable.[24] In this paper, we will simply map how the Nash approach behaves vis-a-vis

[22]Here are some other examples. It might similarly seem objectionably 'fanatical' to devote one's life to the project of giving glory to God if one is almost sure that that project is worthless, merely on the ground that one has a very slight credence that these projects are of truly enormous value. It might seem objectionably fanatical to resolve never again to go on vacation, if one is almost completely convinced that the verdicts of commonsense morality on the relevant matters, merely on the ground that one has a very slight credence that the taking of vacations is morally on a par with walking past a child drowning in a shallow pond. See also the discussion of the 'effective Repugnant Conclusion' in (Greaves and Ord, 2017).

[23]Similarly, in the case of empirical uncertainty: Some people take such cases as that of the Pasadena game (Nover and Hajek, 2004) and 'Pascal's mugging' (Bostrom, 2009) to show that deviations from expected utility are required in cases that involve *extremely* small chances of *extremely* large payoffs (e.g.(Smith, 2014)). But others think that these examples merely show e.g. that utilities must be bounded (Sprenger and Heesen, 2011), or that one is rationally permitted to have zero credence in sufficiently outlandish alternatives (Hajek, 2012), so that expected utility theory does not have the suggested counterintuitive consequences in these cases.

[24]Beckstead (2013; nd, Ch 6.) argues that any decision theory that avoids fanaticism will be subject to an objectionable charge of "timidity": roughly, refusing to let even arbitrarily large increases in possible payoff outweigh arbitrarily small reductions in the probability of getting that payoff.

issues of stakes-sensitivity, and contrast it to the relevant behaviour of other theories of moral uncertainty. We leave it to the reader to assess whether this behaviour on balance favours or disfavours the Nash approach over rival theories of moral uncertainty.

8.3 The NBS and fanaticism

As above, suppose that for two options $a_1, a_2 \in \mathcal{A}$, $a_1 \succ_1 a_2$, and $a_2 \succ_2 a_1$. We assume that both a_1 and a_2 lie on the Pareto frontier. To analyse issues of fanaticism, we suppose that credence in T_2 is very low (say, 1%). (Thus, in the above examples: T_2 is the theory according to which it is massively more choiceworthy to alleviate insect suffering, and a_2 is this option; T_1 and a_1 are the more common-sensical alternatives.)

Can it ever happen that the NBS selects either a_2, or a mixed option that gives high probability to a_2, despite the agent's very low credence in the theory that favours a_2? If so, this might (modulo the caveats above) be considered problematically fanatical, just as the implications of (many versions of) MEC might be considered problematically fanatical.

In light of section 4, this case is straightforward to analyse. For simplicity, we will consider cases in which a_1 and a_2 are the *only* available pure options (that is, $\mathcal{A} = \text{Conv}(a_1, a_2)$). In this case, we know (from Proposition 3) that the NBS will select whichever available option is closest in utility space point to the solution $\text{RD}_{\text{Can}(X)}$ to the corresponding canonical problem.

In general, there is a tendency for that point to be much closer to a_1 than it is to a_2, since, when $p_2 \ll p_1$, $\text{RD}_{\text{Can}(X)}$ is located in the far top-left in the canonical problem. However, there are some disagreement points for which the NBS will select a mixed option that gives high probability to a_2, and some for which the NBS selects a_2 itself, given the choice set $\text{Conv}(a_1, a_2)$. Specifically, this can happen if the set $\text{Conv}(a_1, a_2)$ is confined to the extreme upper-left part of the diagram for the canonical problem, so that $\text{RD}_{\text{Can}(X)}$ is closer to a_2 than to a_1 (and perhaps is closer to a_2 than any other point in $\text{Conv}(a_1, a_2)$. In turn, the condition for that to happen is that the disagreement point has far lower utility according to T_1, and only slightly lower utility according to T_2, than the anti-utopia point.

More precisely, the Nash bargaining solution selects a_2 from $\text{Conv}(a_1, a_2)$ iff

$$\frac{p_2}{p_1} > \left(\frac{u_1(a_1) - u_1(a_2)}{u_1(a_2) - u_1(d)}\right) \cdot \left(\frac{u_2(a_2) - u_2(d)}{u_2(a_2) - u_2(a_1)}\right). \tag{3}$$

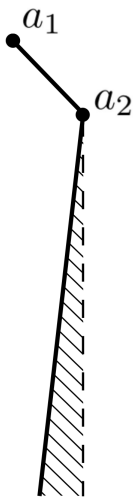


Figure 6: *The set of disagreement points (shaded area) for which the Nash solution selects the 'fanatical' act a_2. The gradient of the diagonal line bounding the shaded region is determined by the split of credences between the theories.*

The set of disagreement points d for which condition (3) is met is illustrated in figure 6.[25]

It is unclear whether the condition (3) is met in the real-life examples in which fanaticism seems a concern. This of course depends in part on the details of how the disagreement point is identified (and, as noted in section 3, there does not seem to be any uniquely compelling resolution of that choice point).

We have so far discussed the possibility that the NBS selects a_2 itself. A relevantly similar possibility is that the NBS selects a point that is *close* to a_2, to an extent that is arguably worrying for reasons of fanaticism. This could happen either because the disagreement point has $u_2(d)$ significantly higher than $u_2(a_1)$, so that the Pareto condition already rules out all options that are any closer to a_1; or because, while points closer to a_1 are among the available options, the point closer to a_2 has a higher Nash product. We will not work through the details of the conditions required for this to happen.

[25]Note that when this condition is met it is unsurprising that in a bargaining-theoretic approach a_2 or something close to a_2 is selected: this choice of disagreement point means that the high-credence theory T_1 "has a very weak bargaining position". The unfavourable-to-T_1 location of the disagreement point can outweigh the fact that in the asymmetric Nash bargaining solution, the difference in credences strongly favours T_1 over T_2. However, this (lack of surprisingness) will do nothing to mitigate the problematic nature of selecting a_2 over a_1, by the lights of anyone who is initially inclined to find this kind of fanaticism problematic.

9 Summary and conclusions

The existing literature on moral uncertainty takes its lead primarily from decision theory, with some exploration of voting-theoretic approaches. This paper has explored a new approach, based instead on bargaining theory.

For concreteness, we 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. On those particular counts it might be considered to have advantages over (respectively) the 'maximise expected choiceworthiness' (MEC) and "my favourite theory" approaches to moral uncertainty.

Despite not recognising any fundamental intertheoretic comparisons, the Nash approach is somewhat sensitive to issues of relative stakes in an intratheoretic sense. In this, it is closely akin 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 one count, the two approaches seem equally good. 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, by a judicious choice of (respectively) the disagreement point or the set of options relative for which the range or variance is evaluated (section 5).

A consideration of ambiguous sign concerns fanaticism (section 8.2): that is, the tendency of decisions to be driven primarily by extraordinarily small probabilities of extraordinarily large (positive or negative) payoffs. The Nash approach is somewhat more resistant to fanaticism than MEC, although (1) the Nash approach is not completely immune from fanaticism, and (2) it is unclear whether or not fanaticism is in the end undesirable.

On two further counts, however, an MEC approach seems superior.

Firstly, the Nash approach has a stronger tendency to prefer options that lead to reasonably high expected utility according to most or all moral theories (section 7). This superficially might seem desirable. However, the particular way in which the Nash approach implements this is via moral risk aversion with respect to empirically expected vNM choiceworthiness. This is *un*desirable, since it means (e.g.) that the Nash approach tends to favour actions that by the lights of all moral theories lead either to an extremely good or to an extremely bad outcome (with nontrivial empirical probabilities for each), no less than actions that are genuinely 'safe' according to all moral theories. In the end, MEC's treatment of the decision predicaments in question seems superior.

Secondly, 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 6). This is problematic, since any such choice (short of the impractical maximally grand-world model) seems arbitrary.

We ourselves tentatively conclude that while the bargaining-theoretic approach is interesting, in the end it seems inferior to at least one version of the standard 'maximise expected choiceworthiness' approach. The most plausible line of resistance to this conclusion concerns fanaticism.

Acknowledgements

For valuable discussions, we are grateful to William MacAskill, Andreas Mogensen, Christian Tarsney, Teru Thomas and Philip Trammell, and to audiences where this work was presented at the London School of Economics, the University of Reading and the Institute for Futures Studies.

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.

Beckstead, Nick (2013). *On the overwhelming importance of shaping the far future*. PhD thesis, Rutgers University.

Beckstead, Nick (n.d.). A paradox for tiny probabilities and enormous values. Unpublished manuscript.

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). Pascal's mugging. *Analysis*, 69(3):443–445.

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 (n.d.). Normative uncertainty, intertheoretic comparisons, and variance normalisation. Unpublished manuscript.

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.

Hajek, Alan (2012). Is strict coherence coherent? *Dialectica*, 66(3):411–424.

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.

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 & 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.

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

Nover, Harris & Hajek, Alan (2004). Vexing expectations. *Mind*, 113(450):237–249.

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(3-4):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 (2015). *Maximising expected value under axiological uncertainty*. PhD thesis, University of Oxford.

Riedener, Stefan (2018). An axiomatic approach to axiological uncertainty. *Philosophical Studies*.

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.

Smith, Nicholas J. J. (2014). Is evaluative compositionality a requirement of rationality? *Mind*, 123(490):457–502.

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.

Sprenger, Jan & Heesen, Remco (2011). The bounded strength of weak expectations. *Mind*, 120(479):819–832.

Tarsney, Christian (2018). Normative uncertainty and social choice. *Mind*.

Tarsney, Christian (n.d.). Vive la difference? structural diversity as a challenge for metarnormative theories. Unpublished manuscript.

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.

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.