Constraints on rational decision-making under moral uncertainty (draft)

Kaarel Hänni[*] Rio Popper[†]

June 2023

Abstract

This paper describes constraints on any rational approach to moral uncertainty. First, assuming a one-shot decision, we use Harsányi's Utilitarian Theorem to show that any rational decision-making procedure under moral uncertainty must have a certain form. In particular, we show that any rational decision-making procedure is extensionally equivalent to maximizing a linear combination of the underlying individual moral theories' respective choiceworthiness functions. Second, we extend the analysis to multi-shot cases — allowing the decision-maker to update based on the outcomes of prior moral observations. In particular, we use the mathematical framework provided in Desai, Critch, and Russell (2018) to justify Bayesian updating in the moral case. Throughout the paper, we discuss the work's implications for existing methods of decision-making under moral uncertainty.

1 Introduction

In this section, we clearly set out what moral uncertainty is, and we briefly survey existing methods of dealing with such uncertainty. This allows us to contextualize the work done in the remainder of the paper.

1.1 What is moral uncertainty?

Let us, for clarity, begin by very briefly considering strictly empirical uncertainty — setting aside the moral realm for now. Let's say you're deciding whether to go to a ballet or to a baseball game. Whatever you choose, you are uncertain about how much you'll enjoy the event. Perhaps you enjoy baseball only if your team wins, and otherwise don't enjoy it; and you enjoy ballet only if your favourite dancer plays a leading role and otherwise are bored. You are, then, essentially picking between two different enjoyment-gambles: utility given ballet, and utility given baseball. Generally, we account for this kind of empirical uncertainty (about which team will win, which dancer will dance, and your corresponding utility) with expected utility theory, which tells us to pick the choice with the highest expected utility. Even if it turns out you would have been happier had you chosen otherwise, so long as you choose the option with the best ex ante expected outcome, we don't say you chose 'wrongly". Moral uncertainty incorporates a new axis of uncertainty — not only might we face uncertainty about which outcomes might happen, but we also face uncertainty about which outcomes are morally 'good'. Consider the stereotypical example. You are unsure between moral theories A and B. Theory t_a tells you that animal-utility and disutility is as relevant as human-utility and disutility (so you shouldn't eat meat), and theory t_b tells you that animal-utility and disutility

[*] kaarelh@gmail.com

[†] rio.popper@philosophy.ox.ac.uk

isn't morally relevant (so it's fine to eat meat). How should you deal with such uncertainty?[1] Proposed attempts at handling such normative uncertainty — so-called meta-normative theories — attempt to answer that question.

1.2 Extant meta-normative theories

There are several existing meta-normative theories — 'My Favorite Theory' (MFT), 'Maximize Expected Choiceworthiness' (MEC), and a cluster of approaches related to voting and bargaining. We briefly survey each in turn.

1.2.1 My Favorite Theory (MFT)

MFT instructs the decision-maker to choose according to the moral theory in which the decision-maker has the most credence. (Note that this is not necessarily the same as the decision-maker's 'favorite' theory—MFT refers rather to the theory the decision-maker, say, takes most seriously or finds most probable.) To return to our example of eating meat: if you have more confidence in theory t_b, then MFT would say that you should go ahead and eat meat. MFT is attractive on the grounds that it avoids the complexity of inter-theoretic comparison, and it provides a compellingly consistent set of instructions to the decision-maker (Gustafsson and Torpman 2014). The most important criticism of MFT is that it is insensitive to relative moral stakes. To return to the meat-eating example, according to MFT, you should eat meat even if you only have slightly more confidence in t_b than in t_a and even if the decision is much higher stakes for t_a than it is for t_b. This is troubling.

1.2.2 Maximize Expected Choiceworthiness (MEC)

MEC tells the agent to choose the best option after aggregating the choiceworthiness functions of each of the underlying moral theories, accounting for both the credence the agent has in the theory and the magnitude of each theory's 'choiceworthiness' or 'preference' on the action. More precisely, it treats moral uncertainty just as expected utility theory treats empirical uncertainty. Let us return to the meat-eating example. Say you have 20% credence in t_a (which says that, since human and animal utility ought to be weighted equally, you should not eat meat) and 80% credence in theory t_b (which says it's fine to eat meat). Now, say also that we know something of the utilities in question and can confidently make inter-agent comparisons. If you eat meat, you gain 10 utils; and the animal loses 100 (from being inhumanely raised and killed). Theory t_a would output a cardinal number $-100 + 10 = -90$ (counting your utility of 10 and nothing of the animal's utility) which accounts both for the -100 of animal-disutility and your 10 of human-utility) were you to eat meat, and 0 were you to abstain from eating meat. Theory t_b would output a cardinal number 10 for eating meat (your 10 of human-utility, and the animal's are not counted) and 0 were you not to eat meat. Taking into account the respective credence-weights of 0.2 and 0.8, the expected choiceworthiness of eating meat is $0.8 \cdot 10 + 0.2 \cdot (-90) = -10$. The expected choiceworthiness of abstaining is 0. Since $0 > -10$, MEC says that you should (contra MFT) abstain. More generally, given possible actions $a_1, \ldots, a_n$, and with conflicting theories $T_1, \ldots, T_m$ in which the agent has credences $p_1, \ldots, p_m$, such that T_i assigns choiceworthiness c_{ij} to action a_j,[2] MEC says the agent should perform the action a_j

[1] Note here that 'should' does not refer to which action one 'should', morally, take. If there were one correct moral theory, the action one 'should' take would be the action indicated by that theory. The 'should' in this context refers to the way an agent should subjectively and rationally handle uncertainty. Note this reading of 'should' is slightly controversial. While, in the rest of the paper, we use 'should' in this way, see Bykvist (2017) and Section 1 of Cotton-Barratt and Greaves (2023).

[2] In Section 2 and later, our choiceworthiness functions will always map from a set of outcomes to $\mathbb{R}$. The present footnote is addressed to the shrewd reader who might then notice that this is in contrast with assigning choiceworthinesses to actions here. However, a choiceworthiness function from outcomes to $\mathbb{R}$ canonically induces a choiceworthiness function from actions to $\mathbb{R}$ — i.e., by taking the choiceworthiness of an action to be the expected

for which the expected choiceworthiness $\sum_{i=1}^{m} p_i c_{ij}$ is maximal.

While MEC is the standard position in the literature (MacAskill, Bykvist, and Ord (2020), Cotton-Barratt and Greaves (2023), it has been criticized on several grounds. First, it has the potential for Pascal's wager style fanaticism (i.e., one theory takes control of a decision by viewing it as high-stakes, even if the agent has low credence in that theory) (Ross 2006). Second, it assumes straightforward inter-theoretic comparability. Insofar as the choiceworthiness function of a theory is grounded solely in its preferences (via the von Neumann-Morgenstern utility theorem), the choiceworthiness function is only defined up to positive affine transformations; but if one were to apply such a transformation to the choiceworthiness function of just one theory, one would (in general) be changing the overall preference ordering. That said, there are ways out of this problem — i.e., ways to derive some reasonable 'exchange rate' between different theories — such as equalizing the theories' variances before maximizing the expectation.[3] Despite these critiques, MEC remains the most widely held of the meta-normative theories, and even the originators of another meta-normative theory (the bargaining approach) think it the most plausible (Cotton-Barratt and Greaves 2023).

1.2.3 Voting and bargaining approaches

The final set of approaches to moral uncertainty aggregates the preferences of the individual moral theories by using methods originally developed to facilitate and understand group decision making. In particular, there is an approach using tools from voting theory (MacAskill (2016)) for ordinal theories or theories with inter-theoretic comparability problems, and an approach using tools from bargaining theory (Cotton-Barratt and Greaves (2023)) for cardinally comparable theories. While we do not discuss these theories in depth here, we will return to them — in particular, to the bargaining-theoretic approach — later in the paper.

1.3 A brief summary of our approach

In the remainder of the paper, we set out a different approach. We begin with the intuition that a decision-making superagent is acting on behalf of moral-theory subagents.[4] To clarify the subagent framing, consider, e.g., a person deciding whether to eat meat as discussed above. In that case, the decision-maker would be the person, and the subagents (subparts of the person) would be theories t_1 and t_2. Note, of course, that in this example the subagents are not separate people but rather different parts of the same person. Our framework applies both to these cases and to broader cases of societal-level moral uncertainty, although we focus on the individual case here. From there, we use Harsányi's Utilitarian Theorem (discussed below in Section 2) to argue that the decision-maker must aggregate the choiceworthiness functions of the moral theories in a particular way. Namely, we argue that any rational decision-making procedure must maximize a linear combination of the moral-theory sub-agents' choiceworthiness functions. While this approach does not specify a unique solution point, it does significantly narrow the possible solution space.[5] This is useful both as a meta-normative theory, insofar as narrowing the solution space can work as a meta-normative theory of its own; and also on the level of evaluating other meta-normative theories (a meta-meta-normative theory). If any other meta-normative theory yields a solution outside the space picked

choiceworthiness of an outcome drawn from a distribution of outcomes conditional on a given action.

[3] A more troubling cousin of this critique is that MEC struggles to deal with theories which only have ordinal choiceworthiness orderings. There are, again, ways of adjusting MEC to handle these slightly better (see, e.g. Tarsney (2017)), but the limitations are present. Our paper faces similar limitations with ordinal ordering.

[4] One can alternatively think of the moral theories as advising the decision-maker — a framing which is more in line with Desai, Critch, and Russell (2018). While we use the subagent framing in the rest of this paper, one could equally use the advising framework.

[5] Note that the bargaining approach of Cotton-Barratt and Greaves (2023) can similarly be seen as picking out a solution space rather than a point, because the disagreement, or fallback, point is underspecified. (Although, given a specific disagreement point, it no longer shares this characteristic with our approach and simply picks out the solution point.)

out by Harsányi's condition, that decision-making procedure is irrational. In Section 3, we use the mathematical framework from Desai, Critch, and Russell (2018) to extend the analysis to sequential situations. We describe and justify a process of Bayesian updating in the moral realm. Overall, the paper describes the set of possible rational moral decision-making procedures under uncertainty. We conclude that MEC and MFT fit comfortably within the space picked out by our approach (and are clearly rational); the approach specified in Cotton-Barratt and Greaves (2023) is rational in cases where a decision-maker is only choosing a single action, but fails in cases where more than one action must be chosen.

2 Applying Harsányi's Utilitarian Theorem to moral uncertainty

Harsányi's Utilitarian Theorem relates the decision-making of a collective of rational agents to the decision-makings of its constituent individuals. Harsányi originally framed the theorem in the context of welfare economics and social choice theory, in which case the theorem says that a rational group of agents' decision-making is given by maximizing an affine function of the welfares of the individuals. In this section, we apply the theorem in the moral context, describing a set of possible rational decision-making procedures for a moral decision-maker facing moral uncertainty. We will find that these procedures are given by linearly aggregating the choiceworthinesses assigned by the moral theory subagents.[6]

2.1 Harsányi's Utilitarian Theorem in the moral context

Here, we explain the mathematical language needed to state Harsányi's Utilitarian Theorem as it applies in the moral context.[7] Our treatment applies to the case where there is a finite set of outcomes $\mathcal{A}$. We will think of both the individual moral theories and the moral decision-maker as having preferences defined on pairs of elements of the set of probability distributions on $\mathcal{A}$, which we call $\Delta(\mathcal{A})$. That is, we think of each preference-haver as having an answer to any question of the form "Would you weakly prefer that an outcome is drawn from the distribution μ, or that an outcome is drawn from the distribution ν?".

The von Neumann-Morgenstern utility theorem says that for any specification of all of a preference-haver's preferences between distributions which satisfies standard rationality constraints,[8] which we will henceforth call the preference-haver being *vNM-rational*, there is a choiceworthiness function $u\colon \mathcal{A} \to \mathbb{R}$ such that one weakly prefers μ to ν iff $\mathbb{E}_\mu[u(A)] \geq \mathbb{E}_\nu[u(A)]$.[9][10] We will from now on assume that any vNM-rational preference-haver comes with a corresponding choiceworthiness function that captures its preferences.[11] This provides all the language we need to state this section's main technical result.

Theorem 1 (A moral application of Harsányi's Utilitarian Theorem, Harsanyi (1955))**.** *Let $t_1, \ldots, t_n$ be vNM-rational moral theories, with respective choiceworthiness functions $u_1, \ldots, u_n$. Suppose the*

[6] As an example, consider again the case in which you are deciding whether to eat meat — considering two moral theories, one of which values animal utility and the other of which does not. Here, you are the moral decision-maker, and each of the moral theories (animal valuing and not animal valuing) is a sub-agent.

[7] The theorem we state will at first glance look different than the usual version (because we set it out specifically in the moral context), but the mathematical content is the same.

[8] See von Neumann and Morgenstern (1947) for a list of these rationality constraints.

[9] $\mathbb{E}_\mu[u(A)]$ denotes the expected value of the random variable $u(A)$ when A is drawn from the distribution μ.

[10] And conversely, having a preference-haver's preferences match the maximization of the expectation of a certain choiceworthiness function implies that the preference-haver is vNM-rational. So being vNM-rational is extensionally equivalent to maximizing the expectation of a certain choiceworthiness function.

[11] This choiceworthiness function is not unique — positive affine transformations of a choiceworthiness function capture the same preferences. We pick one arbitrarily.

overarching decision-maker is also a vNM-rational agent, so it has a choiceworthiness function U. Suppose further that if each t_i weakly prefers μ over ν, then the decision-maker also weakly prefers μ over ν. That is, if $\mathbb{E}_\mu[u_i] \geq \mathbb{E}_\nu[u_i]$ for all i, then $\mathbb{E}_\mu[U] \geq \mathbb{E}_\nu[U]$. Then, there are $c_0 \in \mathbb{R}$ and $c_1, \ldots, c_n \in \mathbb{R}^{\geq 0}$ such that

$$U = c_0 + \sum_{i=1}^{n} c_i u_i.$$

In other words, assuming the moral theory subagents and the overarching decision-maker are rational and that the decision-maker satisfies a Pareto condition with respect to the moral theory subagents, the choiceworthiness function of the decision-maker must be a non-negative linear combination of the choiceworthiness functions of the individual moral theories. The converse of the theorem is also true, in the sense that any choice of $c_0 \in \mathbb{R}$ and $c_1, \ldots, c_n \in \mathbb{R}^{\geq 0}$ leads to a U which satisfies the Pareto condition (and, of course, also vNM-rationality of the decision-maker). In this sense, the theorem says that any rational decision-making procedure under moral uncertainty is extensionally equivalent[12] to maximizing a non-negative affine combination of the choiceworthiness functions of the individuals.

Note that Theorem 1 does not force the decision-maker to make a particular choice when facing any one individual choice between a single pair of lotteries μ, ν — as long as there is at least one theory that strictly prefers μ to ν and another theory that strictly prefers ν to μ, the weights $c_0, c_1, \ldots, c_n$ can be set such that the resulting decision-maker strictly prefers μ to ν, or such that it strictly prefers ν to μ. However, across multiple such decisions — while it may be the case that each is indeed compatible with maximizing some linear combination — it is in general not the case that all the decisions maximize the same linear combination.[13] That is, according to Theorem 1, the multiple decisions which the decision-maker makes must 'hang together' — must maximize the same linear combination across all the decisions (as opposed to a different linear combination for each decision). This distinction will be discussed in more detail in Section 2.2.3.

2.2 Specific meta-normative theories in the context of Theorem 1

In this section, we study how MFT, MEC, and Nash bargaining relate to Theorem 1, in particular examining whether each is deemed irrational by the theorem.

2.2.1 MFT

MFT is not deemed an irrational strategy by Theorem 1. That is, following MFT is indeed maximizing a linear combination of the underlying moral theories' choiceworthiness functions. To wit, MFT puts a weight of 1 on the choiceworthiness function of one, 'my favourite", theory and 0 on all the other theories. It then, of course, simply maximizes that one theory's choiceworthiness function.

2.2.2 MEC

Similarly, MEC is also not deemed an irrational strategy — it does not contradict Section 2.1 above. Here, the coefficients $c_1, \ldots, c_n$ are the credences in each of the underlying theories $t_1, \ldots, t_n$. The agent maximizes the expectation of the linear combination of the choiceworthiness functions of the theories with those credence-coefficients.

[12] We use the word 'extensionally' here to make it clear that the theorem does not say that such a decision-making procedure has to internally resemble computing a linear combination of the choiceworthiness functions for a range of options, and then choosing an option for which this is maximal. The theorem speaks about the verdicts of the decision-making procedure, not about its internals.

[13] The reader may note this section only refers to one-shot cases. By multiple, in this case, we mean multiple decisions while the agent remains in the same state of mind, keeps the same credences on the different moral theories, and does not update based on new moral information. The agent may make multiple decisions, but one can equivalently model those decisions as happening all at once.

Note that, if we take credences or inter-theoretic comparisons (or both) as a free parameter, then any Harsányi-approved strategy can be modeled as a MEC, simply by arbitrarily setting the credences or inter-theoretic comparisons to specific values.[14] However, assuming that we do not take either of credences or inter-theoretic comparisons as a free variable, this is no longer the case.

2.2.3 Nash bargaining

Cotton-Barratt and Greaves (2023) discuss multiple bargaining approaches. We limit the below analysis to the Nash solution they provide, and — in particular — omit discussion of the commonly-considered Kalai–Smorodinsky solution.[15] If the competing moral theories $t_1, \ldots, t_n$ have choiceworthiness functions, respectively, $u_1, \ldots, u_n$, and have the decision-maker's credences in them being $p_1, \ldots, p_n$, then the Nash solution picks the option for which $\prod_{i=1}^{n}(u_i - d_i)^{p_i}$ is maximal, where $(d_1, \ldots, d_n) \in \mathbb{R}^n$ represents the disagreement, or fallback, point. To put $(d_1, \ldots, d_n) \in \mathbb{R}^n$ in other words, it is the point at which utilities would end up were bargaining to break down.[16] Note that for any one-off decision (with finitely many possible sure outcomes), there is a Harsányi-style linear combination is maximized by the Nash solution. This is because the Nash solution by definition yields a point on the Pareto frontier. Any point on the Pareto frontier is on the boundary of the set of options, which for a convex set implies that there is some direction along which it is maximal in the set — i.e., some linear combination of utilities which is maximized at that point.[17]

However, as noted earlier, an essential piece of Harsanyi's rationality constraint is that multiple decisions must hang together — that is, they must maximize the same linear combination. Across multiple Nash decisions — while each is indeed the maximum of some linear combination — it is not the case that all the decisions necessarily maximize the same linear combination. The choices endorsed by Nash bargaining fail to hang together in this sense. For instance, consider a case where there are two theories t_1, t_2 and the disagreement point is $(0, 0)$. In the first decision, suppose that there are two sure outcomes A_1, A_2, giving choiceworthiness tuples $A_1 \to (1, 0)$ and $A_2 = (0, 1)$ respectively, and the full set of outcomes is the set of all probability distributions on $\mathcal{A} = \{A_1, A_2\}$. The Nash bargaining solution is then $\left(\frac{1}{2}, \frac{1}{2}\right)$, and the only linear combinations for which this point is a maximizer are $c_A(u_1 + u_2)$ with $c_A \geq 0$. Whereas with an analogous setup in which the two sure outcomes have choiceworthiness tuples $(2, 0)$ and $(0, 1)$, the only linear combinations which are maximized at the Nash solution are $c_B(u_1 + 2u_2)$. The only way for these to be equal is if $c_A = c_B = 0$. There is a sense in which the choices made by the Nash bargaining solution are indeed compatible with the $c_1, \ldots, c_n = 0$ Harsányi case (that is, where the decision-maker does not have any preferences at all), but any decision-making process is compatible with Harsányi when seen as having no preferences and merely picking some arbitrary option each time. If we are to see the Nash bargaining solution as actually preferring the options it picks over other options, then the Nash solution is not rational in the sense of Section 2.1.[18] Thus, the Nash solution does not provide a rational meta-normative decision procedure.

[14] Remember that insofar as capturing the preferences of an individual theory is concerned, one can apply positive affine transformations to choiceworthiness functions. Supposing that for each theory t_i, the credence p_i assigned to it is positive, we can rescale the choiceworthiness function of the theory to be $u_i' = \frac{c_i}{p_i} u_i$. Note that maximizing $U = c_0 + \sum_{i=1}^{n} c_i u_i = c_0 + \sum_{i=1}^{n} p_i u_i'$ is then equivalent to MEC for the same credences p_i and choiceworthiness functions u_i'.

[15] See generally Kalai and Smorodinsky (1975).

[16] As Cotton-Barratt and Greaves (2023) discuss, there are several options for what the disagreement point might be in the moral uncertainty case. Candidates they propose include the worst possible outcome for both theories, a random-dictator-decision-procedure, and the more qualitative metric of the default behavior — doing nothing (or the societally approved choice, or some other abdication of moral decision-making). The following analysis does not depend on how one conceptualizes the disagreement point.

[17] A longer explanation: by a direction along which a point $\boldsymbol{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n$ is maximal in a set $S \ni \boldsymbol{u}$, we mean a unit vector $\boldsymbol{c} = (c_1, \ldots, c_n)$ such that $\boldsymbol{u} \in \arg\max_{\boldsymbol{y} \in C} \boldsymbol{c} \cdot \boldsymbol{y}$; this is equivalent to saying that $(u_1, \ldots, u_n)$ maximizes the linear combination $c_1 y_1 + \cdots + c_n y_n$.

[18] The (only) assumption of Harsányi's Utilitarian Theorem which the Nash bargaining solution fails is the vNM-rationality of the decision process, and the only reason it fails that is that it fails independence.

3 The sequential case — updating on moral intuitions

In this section, we explore how moral decision making changes when one learns over time from one's moral intuitions. We consider what that learning looks like for a rational decision-maker making a sequence of decisions under moral uncertainty. In particular, we formally justify the claim that any rational sequential moral decision-making process is close to MEC with Bayesian updating on intuitions (in a sense which we make precise later). In this way, our formal work builds on the intuitive, philosophical work in Beckstead (2013) who first suggested Bayesian updating on intuitions as a method of moral learning.

The formal justification is comprised of three sections. In Section 3.1, we set out a framework for various decision-making procedures under uncertainty, which we take from Desai, Critch, and Russell (2018). In 3.2, we apply that framework to the moral case. And, in 3.3, we explain how the work in the previous two sections applies to various meta-normative theories.

3.1 The partially observable Markov decision process (POMDP) setup

Instead of strictly moral theories, let us briefly allow $t_1, \ldots, t_n$ to be 'holistic systems' that hold both empirical and moral views. Formally, following Desai, Critch, and Russell 2018, we think of each theory t_i's view of the decision-making situation as a partially observable Markov decision process (POMDP) $D_i = (\mathcal{S}_i, \mathcal{A}, T_i, \mathcal{O}, \Omega_i, N, u_i)$, where:

S_i is the set of possible states of the environment[19] according to theory t_i

$\mathcal{A}$ is the set of actions, a, available to the decision-maker.

T_i contains the data of all the transition probabilities $\mathbb{P}_i\left(s_{j+1}|s_j a_j\right)$ according to theory t_i, as well as the probability distribution for the initial state the decision-maker finds itself in, $\mathbb{P}_i\left(s_1\right)$.[20]

$\mathcal{O}$ is the set of possible observations.[21] More precisely, it is the set of all possible complete specifications of information the decision-maker might get from the environment at a particular time.[22]

Ω_i contains data about all the observation probabilities $\mathbb{P}_i\left(o_j|s_j\right)$ according to theory t_i.[23]

N is the horizon, i.e. the total number of steps in the decision process.

$u_i\colon$ $(S_i)^N \to \mathbb{R}$ is theory t_i's utility function that maps sequences of states $(s_1, \ldots, s_N)$ to $\mathbb{R}$.

A POMDP is solved by a *policy* π. A policy is a function that takes in a *history* of observations and actions of arbitrary length $1 \le j \le N$, denoted $h_j := (o_{\le j} a_{<j}) := (o_1, a_1, o_2, a_2, \ldots, o_{j-1}, a_{j-1}, o_j)$, and that outputs a probability distribution $\pi\left(o_{\le j} a_{<j}\right)$ over the next action. In other words, a policy is something that looks at past data to which the decision-maker has access and tells it what to do next.

[19] We will later take the environment to contain both moral and non-moral information — there will be a discussion of this below.

[20] In other words, a 'transition probability' refers to the probability, given the environment is in a certain state and the decision-maker takes a specific action, that the environment will end up in a given state in the next time-step.

[21] Observations give potential information about the state of the environment. Recall, however, that this is only a partially observable decision process; meaning that the decision-maker's observations might differ from the full reality of the underlying state, and only provide partial information about it. This reflects our actual world, in which we operate under imperfect information.

[22] For, e.g., a human decision-maker, an element of $\mathcal{O}$ might be a complete specification of all the sensory experiences the human has in one instant. (Note that, by 'in one instant', we simply mean that the human has all of the experiences specified in an observation 'simultaneously' — that is, we do not mean that they all occur at one specific time step j_1 or j_2, but only that they all occur at the same time.)

[23] In other words, Ω_i is an individual theory's function that maps pairs of a given state s_j and a given observation o_j to the probability that, when in state s_j, the decision-maker observes observation o_j.

Desai, Critch, and Russell (2018) provide the following description of Pareto optimal policies in this setting:

Theorem 2 (Desai, Critch, and Russell (2018))**.** *Given a tuple of theories $t_1, \ldots, t_n$ with associated POMDPs $D_1, \ldots, D_n$ as above, a policy π is Pareto optimal for these theories if and only if there are $w_1, \ldots, w_n \geq 0$ with $\sum w_i = 1$ such that for all $1 \leq j \leq n$ and for any history $h_j = (o_{\leq j} a_{<j})$, the policy components satisfy the following:*

$$\pi(h_j) \in \underset{\alpha \in \Delta(A)}{\arg\max} \sum_{i=1}^{n} w_i \mathbb{P}_i \left(o_{\leq j} | a_{<j} \right) \mathbb{E}_{i,\pi}[u_i | h_j, a_j \sim \alpha],$$

where $\mathbb{E}_{i,\pi}$ denotes the expectation according to the POMDP D_i, conditional on policy π being followed (except when choosing the action a_j).

That is, at each step, the decision-maker must choose an action which maximizes a particular linear combination of the underlying theories' expected utilities. The particular linear combination is dictated by the accuracy of the theories' past predictions. For example, if a theory gives a low probability to some possible observation which ends up happening, the decision-maker must then weight that theory less in future steps.

3.2 The moral case

At this point, one might say that this is all well and good, but what can this say about the moral uncertainty case? In this subsection, we further refine the model to account for the specifics of moral uncertainty.

3.2.1 Philosophical intuitions

We take the position that moral theories do in fact differ in their 'empirical' predictions: we assume (following Beckstead (2013)) that intuition can provide us moral information; and this assumption naturally leads to the claim that moral theories differ in their predictions about what will happen when certain moral intuitions are queried. To return to the meat-eating example, a theory that doesn't value animal-utility might predict a lack of guilt if one, say, killed an animal; and a theory that valued animal-utility might predict some guilt.[24] We think of this as using one's moral intuition to look at the part of the moral domain which specifies whether it is bad to kill an animal.

A moral realist could think of this as our moral intuitions accessing a real moral domain — specifically, accessing data about whether a particular moral proposition is objectively true. A moral antirealist might still buy this picture — as capturing how one's intuitions access some underlying reflective preferences (or underlying character).

It is worth noting that our moral intuitions—whether viewed from a realist or antirealist point of view—clearly do not get straightforward access to the underlying moral data. When one considers how one ought to update one's credences in different theories based on the results of one's moral intuitions, one should also consider how one's moral intuitions might systematically be wrong. These systematic errors might, for example, reflect epistemic incompetence such as underlying biases. Overall, a moral theory should expect the results of moral queries to match its beliefs about the moral domain, but to also be prepared for the existence of systematic ways in which those queries might fail to track moral truth.[25]

[24] This assumes guilt, qua ethical response, is affected by the moral realm. If you reject this, you can construct an analogous example with a different moral-response emotion, or you might think of all moral-response emotions — guilt, pride, etc. — as proxies for some purer moral intuition.

[25] We consider the moral measurement-device in more depth in section 3.2.4

3.2.2 Formalizing the moral case

We formalize the intuitive picture described in Section 3.2.1 by considering a special case of the setup from Desai, Critch, and Russell (2018) with significantly more structure than in the original setup. We will think of our theories t_i as agreeing on all purely empirical matters — allowing a focus on the case of pure moral disagreement.[26] We make several assumptions on top of the POMDP setup, each of which helps refine the setup for the moral context.[27]

Observation Storage: Each state contains the decision-maker's memory. And, the states are such that any new observation o_j is added to this memory (which, recall, is part of the state itself) at that same time-step.[28][29]

Unchanging Moral Data: The environment contains unchanging moral data. That is, while empirical facts change from state to state within the environment, in all states of the environment the moral facts remain the same. One's actions can update one's beliefs about morality, or can move one from believing in one moral theory to another, but one's actions cannot change the underlying morality itself — and nor can anything else.

Accessing Moral Data: There are also unchanging facts in each state which determine exactly what happens if one's moral intuition accesses the moral data — in particular, about ways (if any) in which one's moral intuitions systematically fail to accurately track the underlying moral data.

Moral Contemplation: The underlying moral data can be accessed by being at a state in which one queries one's moral intuition about the truth of some particular moral proposition (a *moral contemplation state*); from this state, the transition is to a state in which one receives a moral intuition saying whether the proposition is true or false.[30] The result of one's query — one's moral intuition — is deterministically predicted by the underlying moral data and the facts about how the decision-maker's moral intuition is systematically wrong (if it is), both of which are contained in the initial, querying state.

Empirical Agreement: All moral theories agree on the state space: $S_i = S$. They also agree on all transition probabilities $\mathbb{P}_i\left(s_{j+1}|s_j a_j\right) = \mathbb{P}\left(s_{j+1}|s_j a_j\right)$, but they disagree on the initial probability distribution $\mathbb{P}_i\left(s_1\right)$. That is, since the underlying moral data is constant across all time-steps, the moral theories differ about what moral data the environment starts off with, but — since the disagreement is contained within that initial disagreement — they do not disagree about the likelihood of ending up in a given future state given

[26] One can think of each theory as in fact a combination of one of the moral theories with the singular correct or best scientific theory — which all moral theories share.

[27] There is some technical freedom in how to fit the moral uncertainty case to the POMDP format and apply Theorem 2. We make certain assumptions — specified below — but there are other assumptions that one could equally make and end up at the same result.

[28] Note that there are two meanings of 'observation', which diverge slightly. In one sense, an 'observation' refers to an element of $\mathcal{O}$ which the decision-maker directly receives in the POMDP sense. In another sense, 'observation' refers to a copy of this formal observation which, we stipulate, is stored in each state in the decision-maker's memory. The keen reader may notice that our description does not seem to respect the causal structure of the POMDP, since the observation cannot feed back into the state which produced it. What we technically really mean here is that the observation appears (say, as a pre-observation) at the top of the memory stack in the state first, then gets copied formally as a POMDP-observation to the decision-maker (that is, the observation distribution at that state is simply a point mass on that one observation), and then has become a solidified true observation in the memory stack by the next state.

[29] See Section 3 of Sutton and Barto (2018) for a longer discussion of the agent-environment boundary.

[30] More generally, the intuition could be probabilistic instead of a true-false binary. We consider the binary case for simplicity.

that one is already in some present state and that one acts in a certain way.[31] The theories may also differ in their hypotheses about ways in which the decision-maker's moral intuition might systematically fail to track the underlying moral data.[32] But the theories agree on the observation probabilities $\mathbb{P}_i(o_j|s_j) = \mathbb{P}(o_j|s_j)$.[33]

Choiceworthinesses: Of course, the moral theories may differ in their choiceworthiness functions u_i, which we take to be functions only of the empirical part of a state, and not of the moral data or the facts about how the decision-maker's moral intuition tracks that moral data.

So, to summarize the assumptions, the moral theories t_i differ only in their choiceworthinesses and their initial probability distributions on the moral data and the facts about how moral intuitions could be (systematically) off. Together, the Moral Contemplation and Empirical Agreement assumptions imply that the only times different theories can give different probabilities for observing some observation in the next state is when the current state is a moral contemplation state.[34]

In this setup, Theorem 2 simplifies to only including the moral parts of the updating. For the intuition, recall that we have framed this such that the only times different theories give different probabilities for observing some observation in the following state is when the current state is one of moral contemplation, and the only observations involved here are observations of the decision-maker's moral intuition. For the formal argument, see Appendix A.

We can now state the central theorem of this section.

Theorem 3. *We have n moral theories $t_1, \dots, t_n$, whose views of the moral decision-maker's situation are given by the associated POMDPs $D_1, \dots, D_n$ which have the structure specified above. A policy π is Pareto for these moral theories if and only if*[35] *there are $w_1, \dots, w_n \geq 0$ with $\sum w_i = 1$ such that for all $1 \leq j \leq n$ and any history h_j, the policy components satisfy the following:*

$$\pi(h_j) \in \operatorname*{arg\,max}_{\alpha \in \Delta(A)} \sum_{i=1}^{n} w_i \mathbb{P}_i \left((P_1 \to b_1, \dots, P_r \to b_r) \right) \mathbb{E}_{i,\pi}[u_i | h_j, a_j \sim \alpha],$$

where $P_1, \dots, P_r$ are the moral propositions queried in history h_j, $b_1, \dots, b_r \in \{0, 1\}$ are the outcomes of the queries,[36] *and $\mathbb{E}_{i,\pi}$ denotes the expectation according to the POMDP D_i, conditional on policy π being followed (except when choosing the action a_j).*

That is, in the moral-uncertainty case with updating, a rational decision-maker must choose an action which maximizes a particular linear combination of the underlying theories' expected

[31] In fact, we assume they even all agree on what the initial probability distribution says about all purely empirical matters. We also assume they all agree that the transition probabilities $\mathbb{P}_i(s_{j+1}|s_j a_j)$ are all 0 for s_j, s_{j+1} with different moral data, and that, except for transitions out of moral contemplation states, the transition probabilities between states with the same moral data only depend on the empirical parts of the states. The theories only differ on the moral data in the state, each moral theory t_i being certain that the moral data agrees with what it thinks is correct about every ethical proposition.

[32] The keen reader may notice that, since a specification of the ways in which the decision-maker's moral intuition might be wrong is taken to be contained in the state, this may seem to be in contradiction with all theories having the same S. However, what we mean by this is just that the initial probability assignment to hypotheses about how the decision-maker's moral intuition can systematically fail to track the underlying moral data may be different for different t_i.

[33] In fact, we assume that these only depend on the empirical part of the state. Recall in particular that moral observations are formally simply copies from the memory in the state, which we consider to be within the empirical part of the state.

[34] That is, when the next state is one in which the decision-maker will observe some moral intuition.

[35] To be precise, the 'only if' direction is true up to identifying policies that only differ on probability 0 histories. A more precise if and only if statement would say that the equation only needs to hold for histories with nonzero *empirical part of the likelihood of the history* — see Appendix A for a definition — and for which all observations immediately following the observation of starting to contemplate a particular proposition P actually give a moral intuition on P (as opposed to giving some random empirical observation).

[36] See Appendix A for the definition of $\mathbb{P}_i((P_1 \to b_1, \dots, P_r \to b_r))$ in terms of previously defined probabilities — we have omitted it here, hoping that it is intuitive.

choiceworthiness functions. The weight given to a theory is determined by how accurately it has predicted moral intuitions in the past. That is, if a theory assigns a low probability to some outcome of querying a moral intuition, if that outcome happens, the decision-maker must weight that theory less in future steps.

3.2.3 A special case: additively decomposing utility

To illustrate the above theorem, let us consider a special, simple case where utilities decompose additively. Though the theorem applies to the moral case as stated in general, the simple case considered here helps to elucidate it. Suppose that the horizon $N = 4m$, and the environment is set up such that it keeps cycling through the following four kinds of states for $1 \leq j \leq m$:

1. First, a state where the observation gives one all the possibly morally relevant empirical information about what would happen conditional on one taking various actions at that time. To return to our example from previous sections, let's say you're state is 'in a restaurant' and you observe that you have a decision about whether or not to eat meat. If you decide to eat meat, then (for the sake of a clear illustration, stipulate) one additional animal will be raised and killed. One might think of the decision as the decision-maker deciding whether or not to 'put in an order' for an animal, which will only be raised, killed, and eaten (by the decision-maker) if that decision-maker puts in the order.

2. Second, a state where the value/disvalue from your action is realized.[37] To return to our example,[38] the state would either be s_1, where the animal is raised, killed, and eaten ($u_{t_a}(s_1) = -90$, $u_{t_b}(s_1) = 10$) or the state s_2, where the decision-maker eats a vegetarian meal instead ($u_{t_a}(s_2) = u_{t_b}(s_2) = 0$).

3. Third, a state where one starts to contemplate a particular moral proposition P (as explained earlier). For example, the decision-maker might contemplate whether animal utility matters.[39]

4. Fourth, a state where one receives the moral intuition query result b_j about P_j (e.g., receives the intuition that animal utility matters or the intuition that it does not matter).

After going through a cycle of 4 such states, the process goes back through another cycle of 4 such states, generally with a different decision problem and a different moral proposition. For simplicity, we assume that the moral decisions faced and the moral propositions queried are deterministic and independent of previous actions. Additionally, we assume that utility decomposes linearly — that is, the utility a theory assigns to a full sequence of states decomposes as a sum of utilities, with one summand for each state in which we say value gets realized above. In the first step of the jth iteration of the above loop, we take the actions $\mathcal{A} = \{a_1, \ldots, a_\ell\}$ to (deterministically) lead to second states with theory-i-utilities (respectively) $u_{i,j,1}, u_{i,j,2}, \ldots, u_{i,j,\ell}$.[40] Then, Theorem 3 says that the Pareto optimal policies are precisely those for which there are coefficients $w_1, \ldots, w_n$, such that the probability distribution $\alpha_j = (\alpha_{j,1}, \ldots, \alpha_{j,\ell}) \in \Delta(A)$ chosen at the decision step on the jth iteration[41] is a maximizer of

[37] More precisely, instead of a single state, it might be a deterministic sequence of states (all conditional on one's previous action). That is, in the above, a single state s can span multiple time-steps $j_1, \ldots, j_k$.

[38] Just as in the earlier set-up of this example, we let theory t_a value animal-utility and disutility as just as relevant as human-utility and disutility, and we let theory t_b value animal-utility and disutility not at all. We further assume, for ease of calculation, that the animal being raised and killed causes -100 utils of animal-disutility; that eating animal causes 10 utils of human-utility; and that eating the vegetarian meal causes 0 in human and animal utility.

[39] Note, however, that the intuition which one considers need not be related to the decision that one has made. Although in many practical cases the intuition may indeed relate to the decision.

[40] If our decision problem has fewer options than the full set of actions, we can model this by just saying that the other actions lead to very low utility according to all theories.

[41] This distribution chosen is allowed to, and generally will, depend on the previous history

$$\sum_{k=1}^{\ell} \alpha_{j,k} \sum_{i=1}^{n} w_i \mathbb{P}_i \left((P_1 \to b_1, \ldots, P_r \to b_r) \right) u_{i,j,k}.$$

That is, the Pareto optimal strategies are precisely those for which on step j, one picks for oneself a distribution over actions which maximizes the expectation of a linear combination of the utilities the moral theories assign to the possible options, with the coefficients being given by the initial weights w_i weighted by the degree to which a moral theory has predicted moral intuitions accurately thus far, $\mathbb{P}_i \left((P_1 \to b_1, \ldots, P_r \to b_r) \right)$. In fact, one can think of these weights as implementing Bayesian updating, as explained in the general case in Section 3.3 below.

Remark. *One can model 'moral empiricism' (i.e. the process of updating what one thinks of as moral based on one's actions and the results of those actions) in this toy setting by requiring the moral intuition queried to always be directly related to the action one ought to have taken in the decision situation in the same cycle.* [42]

3.2.4 Notes on the moral-intuition measurement-device

Let us (following Beckstead (2013)) consider the decision-maker's moral intuition as a measurement device — a device that attempts to measure the underlying moral data in the environment. It is the sole information channel from the moral domain to the agent. Recall also that different theories can differ on their predictions about the flaws in this measurement-device. One theory might, for example, think that ex ante each measurement has a 10% chance of being mistaken, and another theory might think the chance only 1%.[43] These differing predictions about the moral measurement devices are relevant because a specific prediction counts in favor of, or against, a theory more or less depending on what that theory predicts about the likelihood of a mistaken measurement.

Moreover, it seems unlikely that conceptualizing a moral-intuition measuring device's mistakes as a flat, equal chance across all measurements is particularly accurate. An example may clarify. Let us say that a decision-maker's intuition says that it's wrong to eat meat for dinner tonight. One theory — which gives no weight to animal-utility and disutility — would be surprised by this intuition measurement (because it thinks it does not match the true moral data). That theory's opinion should, therefore, be weighted less in future decisions. But say now that the decision-maker is later querying the intuition about whether it would be wrong to eat meat as a late-night snack. Were the meat-eating theory to still rule an anti-meat intuition mistaken, it seems that it should not be a new update down on the theory. Intuitively, this new intuition about eating meat as a late-night snack should not contain much new information beyond what was already contained in the intuition about eating meat for dinner — a reasonable theory might say that both incorrect intuitions are explained by one's "measurement device" being broken in the same way (in this case, perhaps that one's intuitions which intuitively factor through the moral worth of animals are generally mistaken). Intuitively, the most extreme failure here is to keep querying the same intuition over and over again, which will, in most cases, give the same result — and using this to arbitrarily repeatedly update down on a moral theory. Allowing one's moral theories to have more sophisticated predictions about one's moral intuition-errors — namely, theories that consider systematic errors — allows one to avoid these failures.[44]

While these brief notes on moral-intuition measurement-devices do not influence the general-case mathematical constraints we discuss elsewhere in the paper, they do have implications for

[42] Neither the cyclic structure nor utility decomposing linearly is crucial for this though — one can analogously model moral empiricism in the non-toy setting.

[43] We can intuitively tell that the error rate must be above 0% because we have contradictory moral intuitions — moral intuitions which cannot both be true, at least under the assumption that moral truth itself cannot be contradictory. (Though the claim that base moral truth — or the underlying moral data in a state, even if one wants to view that data from an antirealist perspective — must be consistent is itself a non-trivial assumption.)

[44] For more on this, see Beckstead (2013).

how this updating must work in practice. In particular, a decision-maker engaged in a process of moral updating must understand how much, or little, different 'measurements' are correlated, and weight specific updates accordingly. Is there effectively a finite amount of total information to be learned about the moral domain via one's intuition, because the moral domain only has so many "independent axes"?[45] How would one go about determining what kinds of new information about the moral domain, if any, can be inferred from a particular intuition query result? Can one seek out ever wackier thought experiments to query the moral domain in uncorrelated ways? Ought one to do so? We see these as questions for further study.

3.3 Specific meta-normative theories in the context of Theorem 3

In this section, we study how MFT, MEC, and Nash bargaining relate to Theorem 2, in particular examining whether each is deemed irrational by the theorem.

3.3.1 MFT

A version of MFT which involves committing to theory t_i for all time is Pareto optimal in the sense of Theorem 3. To wit, with $w_i = 1$ and every other weight set to 0, a policy π which maximizes $\mathbb{E}_{i,\pi}[u_i]$ satisfies the condition from the theorem statement.

However, a different — and potentially more compelling — version of MFT fails the condition. In particular, it is plausible that one should endorse a version of MFT that allows one to (while acting on one's 'favorite' theory) be open to changing one's mind about which of the theories is the 'favorite' (based on the results of querying one's moral intuitions). For clarity, consider an example. Supposing for simplicity that there are just two theories t_1 and t_2, if at first t_1 is the 'favorite', but t_2 eventually overtakes it because t_2 better agrees with later-considered intuitions, then this version of MFT would tell one to initially completely ignore the opinion of t_2 and subsequently to completely ignore the opinion of t_1. This, however, is generally not Pareto optimal because there could be decisions that only barely matter to t_1 but really matter to t_2 which are made while the decision-maker's credence in t_1 is higher, and then decisions that really matter to t_1 but only barely matter to t_2 while the decision-maker's credence in t_2 is higher. Since plausibly both t_1 and t_2 would prefer a trade in this situation — i.e., there is another policy which is a Pareto improvement — this version of MFT generally fails Pareto optimality.[46]

To put the previous paragraph's conclusion in different terms, the second version of MFT (which allows the decision-maker to change which theory is the 'favourite') is incompatible with the above condition, since the decision-maker's credence weights in this example case are not 0 and 1 but rather change and take on intermediate values — and yet the linear combination that the decision-maker maximizes does not reflect those weights.

[45] The picture we have in mind here is that a reasonable moral 'error theory' (i.e. a theory about how one's intuitions might be in error, not the meta-ethical view) might consider it possible for the moral intuition measurement process to have a malfunctioning sensor for reading the coordinate along each individual axis of morality-space. For instance, the correct morality might fall somewhere on the axis between averaging vs summing, and on the axis between prioritizing love vs decency, and on the axis between freedom and order, and on the axis between hedonism and preference satisfaction, and plausibly on 20 other such conceivably independent axes, but plausibly not on infinitely many such independent axes. And once the moral 'error theory' has settled on a specification of which of the finitely many sensors are broken in the mapping from the moral domain to intuition-query answers (assuming one can then accurately predict future intuitions, which seems intuitively plausible), the decision-maker will stop updating down on that moral theory.

[46] And in fact, given that one assumes that, e.g., this version of MFT also updates its credences in a Bayesian way, one can set up the mathematical details such that this turns out to be the case, although we omit a formal description. Note here that we assumed Bayesianism in the previous sentence — but Bayesianism is only one example of an updating procedure that would yield this non-Pareto result. Plausibly any updating procedure paired with MFT would also yield a lack of Pareto optimality for the same general reason.

3.3.2 MEC

In analogy with Section 2.2.2, any Pareto optimal policy in the sequential setting can be seen as closely related to MEC with Bayesian updating on intuitions. The situation here, though, is a tad subtler.

Begin by considering MEC: in the standard case, the decision-maker simply chooses the action that maximizes some linear combination of the theories' choiceworthiness functions. The weights — even in the standard case — are the decision-maker's credences in each theory: $w_i = p_i$. The natural rational extension of this to the case with updating on one's moral intuitions says that one should update one's credences in the theories in a Bayesian manner. That is, after seeing the sequence of intuition query results $(P_1 \to b_1, \ldots, P_r \to b_r)$, the probabilities one assigns to theories should be proportional to $p_i \mathbb{P}_i \left((P_1 \to b_1, \ldots, P_r \to b_r)\right) = w_i \mathbb{P}_i \left((P_1 \to b_1, \ldots, P_r \to b_r)\right)$. Since these are exactly the coefficients that show up in Theorem 3, the policy of maximizing expected choiceworthiness with each of one's decisions, updating on one's intuition queries in a Bayesian way, is clearly compatible with Theorem 3.

More complex is the related claim that 'all rational sequential strategies are a form of MEC with Bayesian updating'. More precisely, the claim is this: the requirement placed on a policy by Theorem 3 is equivalent to saying that there ought to be some MEC with Bayesian updating on intuitions — i.e. an MEC with certain priors; or alternatively, given a nonzero prior probability in each theory, with certain inter-theoretic comparisons (all this is in complete analogy with Section 2.2.2) — such that the policy is optimal according to that MEC. The subtlety is that while the corresponding MEC says one should always be indifferent between different options with the same expected choiceworthiness, the theorem does not say one needs to be indifferent between e.g. the different options in the $\arg\max$. The theorem, even after choosing w_i, does not really specify the preference structure the decision-maker needs to have at all at any one time beyond specifying a condition on what the decision-maker must choose in the decision problems it could actually encounter.[47]

3.3.3 Nash Bargaining

Much like MFT, the Nash solution also fails Pareto optimality, although it's failure isn't related to any updating procedure.[48] It fails Pareto optimality on grounds discussed by Cotton-Barratt and Greaves (2023) — in particular, on small vs large world concerns.

Consider two decisions, A (a choice between a_1 and a_2) and B (a choice between b_1 and b_2. Consider also two theories, t_1 and t_2, in which the decision-maker has equal credence. In decision A, let t_1 prefer a_1 very much, and let t_2 prefer a_2 only a little. In decision B, let the positions be reversed — that is, let t_1 prefer b_1 only a little, and let t_2 prefer b_2 very much. Since we assumed that the decision-maker has equal credences in both theories, the Nash solution would tell us to proverbially 'flip a fair coin' to make each, independent decision.[49] This fails Pareto optimality, since both theories would prefer to trade from the equal mixture to outcome a_1, b_2.

The crucial point here is that making each decision in isolation (as a 'small world') yields a different result than were the multiple decisions taken together (as a 'grand world'). Taking the decisions together, Nash yields a Pareto optimal solution (a_1, b_2) in our example. Taking the decisions separately, with an individual bargaining solution for each, yields a different — non-Pareto — result.

[47] In case the set of possible amounts of choiceworthiness realized is such that the policy π is the unique $\arg\max$ of the expectation of $\sum_{i=1}^{n} w_i u_i$, it seems more fair to say that the theorem does just say one needs to act as the corresponding Bayesian MEC, as in that case, the theorem fully specifies the unique action one is allowed to take.

[48] We mention it here only because the failure is often considered in the sequential case.

[49] Much of this example is taken from Cotton-Barratt and Greaves (2023). As they also emphasize, the $50-50$ credences aren't needed to make the example work. They simply simplify the example for clarity.

4 Conclusions

Most existing work in moral uncertainty does one of two things. Some work attempts to propose particular methods of dealing with normative uncertainty (see, e.g., MacAskill, Bykvist, and Ord (2020) and Cotton-Barratt and Greaves (2023)). These meta-normative theories may approach the problem from different disciplines—from decision theory, voting theory, or bargaining theory—but they share an end of proposing a specific method for handling these cases. Other work has made progress on specifying and parsing out difficulties in the problem of moral uncertainty—without necessarily connecting those observations about the structure of the problem to any proposed meta-normative theories. Some example work in this category includes work done on (the difficulty of) inter-theoretic comparison—although that work often translates to work on proposing specific meta-normative theories (see, e.g., Tarsney (2021)). The approach we take in this paper does not try, from the ground up, to construct a meta-normative theory. Instead, we begin with what the constraints on any rational meta-normative theory must be. From these constraints, we argue that any rational meta-normative theory must fall inside a particular space, and we describe this space both in the one-shot and in the sequential case. With that space in mind, we further justify MEC with Bayesian updating as a rational theory, and make comments on other theories in either the one-shot or the sequential case. In that way, our paper does not attempt to set out or apply a meta-normative one—instead, it sets out and applies a meta-meta-normative theory.[50]

A Reducing Theorem 3 to Theorem 2

We will be assuming the POMDP has the form specified in Section 3.2.2, and providing a reduction of Theorem 3 to Theorem 2. Recall that the different theories only differ in their predictions for observations which are just after a previous observation which says that the decision-maker has started contemplating a particular moral proposition P, which we will call the moral query observations. For a particular history h_j, we let the set of all of its moral query observations be $Q(h) = \{o_{j_1}, \ldots, o_{j_m}\}$, corresponding respectively to propositions $P_1, \ldots, P_m$. Note that because the theories t_q and t_r agree on all empirical matters, the products $\mathbb{P}_q\left(o_{\leq j} | a_{<j}\right) = \prod_{\ell=1}^{j} \mathbb{P}_q\left(o_\ell | h_{<\ell}\right)$ and $\mathbb{P}_r\left(o_{\leq j} | a_{<j}\right) = \prod_{\ell=1}^{j} \mathbb{P}_r\left(o_\ell | h_{<\ell}\right)$ differ only at indices in $Q(h)$. Let us establish some notation for the part of the products which is the same for different hypotheses:

$$p_e(h) := \prod_{\ell \in [j] \setminus Q(h)} \mathbb{P}\left(o_\ell | h_{<\ell}\right),$$

which we call *the empirical part of the likelihood of history h.* The rest is

$$p_m(h; i) = \prod_{\ell \in Q(h)} \mathbb{P}_i\left(o_\ell | h_{<\ell}\right),$$

which we call *the moral part of the likelihood of history h, according to theory i.* For our moral POMDPs, conditional on having entered the moral contemplation state for a particular proposition P, the moral intuition querying mechanism has the property that the next moral measurement is independent of all past empirical observations.[51]

[50] More precisely, our paper does narrow the set of possible actions that a rational decision-maker could take. In that way, it is a meta-normative theory — to be a decision-making procedure, something need not necessarily pick out one single right choice rather than a set of possible right choices. The theory here, then, can be conceived of both as a meta-normative theory that simply picks out a space rather than a point, or as a meta-meta normative theory that bounds the space of other meta-normative theories. The difference is semantic.

[51] Let us explain this by appealing to an implicit causal graph modeling the state transitions. Let o'_j denote the node in the causal graph for the last observation that happens in the jth state. We think of o'_j as having a moral node m_j feeding into it, which is set up so as to have a special null value when the query is non-moral. TODO: finish stating this as a causal graph Markov boundary thing

In this case, the moral part $p_m(h;i) = \prod_{\ell \in Q(h)} \mathbb{P}_i \left(o_\ell | h_{<\ell}\right)$ simplifies further because, conditional on having entered the moral query state for proposition P as evidenced by the previous observation, the probability of one's moral intuition telling one that P or that $\neg P$ is independent of everything in the history $h_{<\ell}$ except for the sequence of previous moral intuition query results. To make this independence of the rest of the history (and of the observation indices) explicit, if the history h has propositions $P_1, \ldots, P_r$ being claimed to have truth values $b_1, \ldots, b_r \in \{0,1\}$ by the moral queries,[52], we have $p_m(h;i) = \prod_{r=1}^{q} \mathbb{P}_i \left(P_r \to b_r | (P_1 \to B_1, \ldots, P_{r-1} \to b_{r-1})\right) = \mathbb{P}_i \left((P_1 \to b_1, \ldots, P_r \to b_r)\right)$. Writing the probability term in Theorem 2 as a product of these $p_e(h)$ and $p_m(h;i)$ and collecting the factor of $p_e(h)$ yields Theorem 3.

[52] Recall, as discussed above, that we might also consider other options for the outputs of the moral intuition measurement device — for instance, probabilities in $(0,1)$ — but this does not change much.

References

Beckstead, Nicholas. 2013. "On the overwhelming importance of shaping the far future," https://doi.org/doi:10.7282/T35M649T.

Bykvist, Krister. 2017. "Moral Uncertainty." *Philosophy Compass* 12 (3): e12408. https://doi.org/10.1111/phc3.12408.

Cotton-Barratt, Owen, and Hilary Greaves. 2023. "A Bargaining-Theoretic Approach to Moral Uncertainty." *Journal of Moral Philosophy.*

Desai, Nishant, Andrew Critch, and Stuart J Russell. 2018. "Negotiable Reinforcement Learning for Pareto Optimal Sequential Decision-Making." In *Advances in Neural Information Processing Systems,* edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/5b8e4fd39d9786228649a8a8bec4e008-Paper.pdf.

Gustafsson, Johan E., and Olle Torpman. 2014. "In Defence of My Favourite Theory." *Pacific Philosophical Quarterly* 95 (2): 159–174. https://doi.org/https://doi.org/10.1111/papq.12022. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/papq.12022. https://onlinelibrary.wiley.com/doi/abs/10.1111/papq.12022.

Harsanyi, John C. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." Remark: I actually used https://hceconomics.uchicago.edu/sites/default/files/pdf/events/Harsanyi_1955_JPE_v63_n4.pdf, *Journal of Political Economy* 63 (4): 309–321. ISSN: 00223808, 1537534X, accessed March 22, 2023. http://www.jstor.org/stable/1827128.

Kalai, Ehud, and Meir Smorodinsky. 1975. "Other Solutions to Nash's Bargaining Problem." *Econometrica* 43 (3): 513–518. ISSN: 00129682, 14680262, accessed May 16, 2023. http://www.jstor.org/stable/1914280.

MacAskill, William. 2016. "Normative Uncertainty as a Voting Problem." *Mind* 125 (500): 967–1004. ISSN: 00264423, 14602113, accessed May 14, 2023. http://www.jstor.org/stable/26361891.

MacAskill, William, Krister Bykvist, and Toby Ord. 2020. "39Maximizing Expected Choiceworthiness." In *Moral Uncertainty.* Oxford University Press, September. ISBN: 9780198722274. https://doi.org/10.1093/oso/9780198722274.003.0003. eprint: https://academic.oup.com/book/0/chapter/267645210/chapter-pdf/49946271/oso-9780198722274-chapter-3.pdf. https://doi.org/10.1093/oso/9780198722274.003.0003.

Ross, Jacob. 2006. "Rejecting Ethical Deflationism." *Ethics* 116 (4): 742–768. ISSN: 00141704, 1539297X, accessed May 14, 2023. http://www.jstor.org/stable/10.1086/505234.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction.* Cambridge, MA, USA: A Bradford Book. ISBN: 0262039249.

Tarsney, Christian J. 2017. "Rationality and Moral Risk: A Moderate Defense of Hedging."

———. 2021. "Vive la Différence? Structural Diversity as a Challenge for Metanormative Theories." *Ethics* 131 (2): 151–182. https://doi.org/10.1086/711204.

von Neumann, John, and Oskar Morgenstern. 1947. *Theory of games and economic behavior.* Second edition. Princeton University Press.