SUPPOSITIONAL DESIRES AND RATIONAL CHOICE UNDER MORAL UNCERTAINTY

Nicholas Makins

Forthcoming in *Ergo*. Please cite published version.

Abstract

This paper presents a unifying diagnosis of a number of important problems facing existing models of rational choice under moral uncertainty and proposes a remedy. I argue that the problems of (i) severely limited scope, (ii) intertheoretic comparisons, and (iii) 'swamping' all stem from the way in which values are assigned to options in decision rules such as Maximisation of Expected Choiceworthiness. By assigning values to options under a given moral theory by asking something like '*how much do I desire this option, supposing this theory is true?*' rather than '*how much value does this theory assign to this option?*' these problems can be avoided, while the appealing features of these accounts can be preserved. This amendment provides a role for the preferences, desires, or goals of rational agents that is curiously absent from the existing discussion of what individuals rationally ought to do when they are uncertain about what they morally ought to do.

1 Introduction

Recent years have seen something of a boom in philosophical work concerning the question of how agents should act under moral uncertainty. That is, what they should do when they do not know what they should do. A number of proposals for answering this question understand it as a challenge for rational choice (Bykvist, 2014; Lockhart, 2000; Ross, 2006; Sepielli, 2014) and adopt the tools of decision theory to answer it. It is intriguing, therefore, that these accounts make little or no mention of agents' preferences or desires–notions that are central to traditional decision-theoretic accounts of practical rationality–but instead work entirely with the values provided by the moral theories about which the agent is uncertain. In so doing, these accounts fail to recognise the ways in which rational agents must strike a balance between their moral commitments and other, non-moral considerations, such as their own self-interest. In this paper, I will argue that this feature is the root of several of the most stubborn problems facing what I call the Analogical View: the view that we should treat moral uncertainty as analogous to empirical uncertainty in theories of rational choice. By finding a role for some conventional notion of preference or desirability, arrived at through a compromise between self-interest and moral commitment, I aim to modify this view in such a way as to avoid these problems.

This is a somewhat modest task: I do not aim to provide a full defence of the Analogical View, nor to compare it to all alternatives. Some readers will take issue with the view for reasons beyond those considered here. However, I at least aim to show that this modified approach fares better than existing instantiations of the Analogical View, by avoiding their most substantial flaws, while maintaining their major advantages. §2 will characterise the Analogical View in more detail. §3 will explore three well-known objections to this approach: (i) limited scope to numerically representable theories; (ii) the problem of intertheoretic comparisons; and (iii) 'swamping'. §4 will suggest a unifying diagnosis of these problems, §5 will propose a remedy, and §6 concludes the paper.

2 The Analogical View

A popular approach to rational choice under moral uncertainty is to suggest that we should treat it as roughly analogous to rational choice under empirical uncertainty about non-moral matters and use something like expected utility theory to guide our decisions. Call this the Analogical View (Aboodi, 2022; Bykvist, 2014; Carr, 2020; Dietrich & Jabarian, 2022; Hicks, 2018; Lockhart, 2000; MacAskill, 2014; MacAskill et al., 2020; MacAskill & Ord, 2020; Riedener, 2020, 2021; Ross, 2006; Sepielli, 2009, 2010).

One prominent formulation of the Analogical View is a procedure called Maximisation of Expected Choiceworthiness (MEC) (MacAskill et al., 2020; MacAskill & Ord, 2020). On this account, a choice under moral uncertainty is represented by the following components: options $A_1,\ldots,A_m$ between which the agent must choose; moral theories $T_1,\ldots,T_n$ about which the agent is uncertain; a probability function P over the moral theories, which represents the agent's moral credences; and a set of choiceworthiness functions $cT_1,\ldots,cT_n$ which assign a number to each option, representing the choiceworthiness of that option according to the theory under consideration. The expected choiceworthiness of each option is given by:

$$EC(A_i) = \sum_{j=1}^{n} cT_j(A_i) \cdot P(T_j)$$

The claim made by proponents of MEC is that, insofar as choices that maximise expected utility are rational under empirical uncertainty, choices that maximise expected choiceworthiness are rational under moral uncertainty. For example, imagine a person who wins the lottery and wants to 'do the right thing' by giving much of their winnings away, but is uncertain about who would be the most morally worthy recipient. On the one hand they think that it might be best to give the money to whichever charity could be shown most effectively to promote wellbeing and alleviate suffering. On the other hand, they give some credence to the view that they have special obligations

to help their close family. Suppose that according to the first view, call it T_1, it would be much better to donate the money to Malaria Consortium, since it is regarded as the most effective charity around (GiveWell, 2023). This view could be represented with the choiceworthiness function cT_1 such that cT_1(Malaria Consortium) = 1000 and cT_1(Family) = 10. However, according to the second view, call it T_2, the preferential weighting of benefits given to one's nearest and dearest means that it would be slightly better to use the money to provide financial security for a few close relatives. This view could be represented with the choiceworthiness function cT_2 such that cT_2(Malaria Consortium) = 40 and cT_2(Family) = 60. Suppose that our lottery winner thinks that the family-oriented moral view T_2 is more likely to be correct than the impartial, utilitarian alternative T_1, such that $P(T_1) = 0.4$ and $P(T_2) = 0.6$. The expected choiceworthiness of these options would then be given by:

$$\text{EC(Malaria Consortium)} = (1000 \times 0.4) + (40 \cdot 0.6) = 424$$

$$\text{EC(family)} = (10 \cdot 0.4) + (60 \cdot 0.6) = 40$$

According to MEC, therefore, they should donate the money to Malaria Consortium, as that option has the higher expected choiceworthiness.

This approach has some appealing features. For one thing, its structural similarity to standard expected utility theory gives a parsimonious and consistent account of rational choice under uncertainty. There are many types of proposition about which we may be uncertain, and we do not adopt different decision procedures for each. We may be uncertain about the weather, about upcoming elections, or about financial markets, but we do not adopt different decision theories for meteorological, political, or economic uncertainty. Absent further argument, we should not treat moral uncertainty as substantially different from uncertainty about any other kind of proposition (MacAskill & Ord, 2020).

Another advantage of MEC is that it is sensitive to how much is stake according to different moral theories. If you are fairly confident that *A* is slightly better than *B*, but give some credence to a view according to which *A* is much worse, it seems as though you might want to give this low credence, high-stakes view some sway. For example, if I believe that eating meat is probably morally acceptable, but there is a small chance that it is severely morally wrong, then it would seem sensible to avoid taking the considerable moral risk of eating meat for the comparatively small benefit of a slightly tastier meal. Analogously, even if you think it is much more likely that you will not be involved in a car crash than that you will, you still wear a seatbelt, because if that did happen, the stakes would be far higher. Note that this feature is not captured by other prominent views on moral uncertainty, such as the view that what you ought to do is simply whatever the true moral theory says you ought to do (Harman, 2015; Weatherson, 2014, 2019), or the view that you ought to do whatever is recommended by the moral theory that you deem most likely to be true (Gracely, 1996; Gustafsson & Torpman, 2014).

Despite these appealing features of the Analogical View, there are number of well-known and substantial objections to this approach. It is these to which I turn next.

3 Three Problems for the Analogical View

3.1 Limited Scope

The first problem is that the scope of the Analogical View is severely limited to only those cases in which all moral theories under consideration can be numerically represented in a particular way. Calculating expected choiceworthiness is only possible if the theories under consideration assign choiceworthiness values that are measurable on an interval scale, or provide the sort of ordering that can be used to construct an interval scale representation (for example, an ordering that satisfied the von Neumann and Morgenstern axioms) (Riedener, 2021). An interval scale allows

ratios of value differences to be expressed. For example, the difference in choiceworthiness between *A* and *B* is double that between *C* and *D*. However, many moral views are simply not in the business of assigning values to options, or generating orderings. For example, the Ten Commandments are a set of prescriptions and proscriptions, with no in-built measure of choiceworthiness or even ordering of options. These moral laws do not say that remembering the Sabbath has a value of 10 and coveting thy neighbour's wife has a value of −50. Nor do they imply any ordering of options in terms of the degree to which the Commandments are satisfied. Therefore, any agent who holds some degree of belief in a view like the Ten Commandments will be unable to calculate expected choiceworthiness for the options from which they must choose. The requirement that all theories under consideration can be represented by an interval scale of choiceworthiness severely limits the scope of the Analogical View.

It is worth noting that (MacAskill et al., 2020) propose alternative tools from decision and social choice theory that may be employed when considering theories that do not have the structure required by MEC. However, these back-ups are unnecessary if this problem can be avoided from the start, as I will argue it can.

3.2 The Problem(s) of Intertheoretic Comparisons

The second problem is that MEC requires intertheoretic comparisons of choiceworthiness and, even when all moral theories under consideration are representable on an interval scale, these comparisons can still not meaningfully be made. There are two reasons for thinking that such intertheoretic comparisons are not possible, which give rise to two different versions of the problem. The first is that some different moral theories refer to fundamentally different conceptions of moral choiceworthiness and it does not seem that comparisons between measures of these different conceptions are meaningful. None of these moral theories contains any information regarding its conception of moral choiceworthiness in the terms of the others. Nor is

there some more general third theory that can be used to convert the units of one theory into the units of another. Call this the Reference Problem (Broome, 2012; Gracely, 1996; Hudson, 1989; Riedener, 2019; Tarsney, 2018a).

The Reference Problem applies when attempting to make comparisons across moral theories that differ in their explanation of the nature of moral choiceworthiness. However, sometimes one's credence may be divided only between theories that agree on this matter. For example, one may be certain of prioritarianism for the distribution of scarce healthcare resources, but remain uncertain about the precise weighting of benefits given to people at different levels of welfare. However, there is a second problem, which undermines intertheoretic comparisons even in this sort of case. This stems from the fact that interval scales are uniquely determined only up to positive affine transformation.[1] This means that there are multiple equivalent representations of any one theory, and using different representations will provide different answers to the question of which option maximises expected choiceworthiness. There is nothing within the theories themselves that can tell us how to calibrate their scales, so MEC can recommend one option or another, depending on which scale we choose. And that, ultimately, is no recommendation at all. Call this the Scale Problem (Hedden, 2016; Hicks, 2018; Lockhart, 2000; MacAskill et al., 2020; Nissan-Rozen, 2015; Sepielli, 2010).

3.3 Swamping

The third problem for the Analogical View is that it is subject to 'swamping' effects. Theories that posit larger differences in choiceworthiness have a greater effect on the expected choiceworthiness of options, ceteris paribus. This means that a choice can be dictated by some moral theory that seems highly implausible but assigns wildly large differences in choiceworthiness between the

[1] If T is a function producing an interval scale, then T^* is a positive affine transformation of it if and only if it takes the form $T^* = aT + b$, where a is a positive constant and b is any constant.

options. This problem and sensitivity to stakes are in fact two sides of the same coin: because expected choiceworthiness is sensitive to stakes, a highly dubious theory can swamp all other (more likely) theories under consideration, so long as it posits large enough differences in choiceworthiness (Bykvist, 2017; Hedden, 2016; MacAskill, 2014; MacAskill et al., 2020; MacAskill & Ord, 2020; Ross, 2006).

It is important to distinguish this from a closely related problem concerning infinite choiceworthiness. It is well-known in decision theory that the possibility of infinite utilities make trouble for expected utility theory (Arrow, 1971; Nover & Hájek, 2004; Samuelson, 1977) and these problems have analogues for the use of decision theoretic approaches to choice under moral uncertainty. For example, if an option has infinite choiceworthiness on one moral view and negative infinite choiceworthiness on another moral view, both with some positive probability, then the expected choiceworthiness of that option will be undefined (Tarsney, 2018b). There is a large literature on how best to handle the problems produced by infinite payoffs in decision theory, but this issue is distinct from swamping itself. [2] The problem with infinite values is that, without some further fine-tuning, the tools of decision theory break down altogether (MacAskill et al., 2020). The problem with swamping is that sufficiently large (but not necessarily infinite) choiceworthiness differences cause MEC to produce counter-intuitive recommendations.[3]

[2] These two issues are termed "swamping" and "fanaticism" by (MacAskill et al., 2020), while (Ross, 2006) distinguishes between "fanaticism" (finite) and "ultrafanaticism" (infinite). Elsewhere, the term "fanaticism" is sometimes used for problems with infinite value or utility (Bostrom, 2011) and sometimes for problems that can arise with finite value or utility (Wilkinson, 2022).

[3] One common strategy for dealing with these sorts of problems is to use bounded utilities. Something like this strategy might be effective against swamping in the context of moral uncertainty. However, bounded value functions introduce a host of further contested issues that needn't trouble us if the problem can be avoided from the start (Beckstead & Thomas, 2023).

4 A Unifying Diagnosis

There is a curious feature of MEC, as an instantiation of the Analogical View, that I believe is responsible for all three of these problems: it is an attempt to bring decision theory to bear on choices under moral uncertainty, but it makes little or no mention of agents' ends, desires, or preferences, notions central to conventional decision-theoretic accounts of rational choice. Decision theory is usually thought of as concerned with instrumental rationality (Buchak, 2014; Joyce, 1999).[4] That is, claims about what would be the best way for an individual to go about achieving their ends, whatever those may be. A theory of morality may be able to ignore a person's ends and still tell them what they ought to do; it does not matter whether you want to tell the truth or not, doing so is morally obligatory. Theories of instrumental rationality, however, require ends and means-end beliefs as inputs. *If* you aim to make a cup of coffee and believe that this requires you to heat some water, *then* it is instrumentally rational for you to do so. If you have no such aim, then instrumental rationality offers no such guidance. This idea of instrumental rationality is key to the diagnosis and cure of the three problems plaguing the analogical view that I propose. MacAskill, Bykvist, and Ord (2020) explicitly acknowledge this conception of rationality in the context of moral uncertainty: "*Rationality […] has to do with what one should do or intend,* given one's beliefs and preferences. *This is the kind of rationality that decision theory is often seen as invoking.*" (p.20).

Some hold the view that instrumental rationality is all there is to practical rationality. This is often called a "Humean" view, inspired by Hume's famous adage that '*reason is, and ought only to be the slave of the passions'* (Hume, 2007). Others, meanwhile, draw a distinction between structural and substantive rationality, the latter concerning whether one is suitably responsive to the reasons one has (Fogal, 2020; Fogal & Worsnip, 2021; Hooker & Streumer, 2003; Scanlon, 2007; Worsnip, 2021). The crucial role of instrumental rationality in the arguments below means that they will be

[4] See (Thoma, 2017) for a detailed critique of this view.

most appealing to those who hold the Humean view of rationality. Or, at the very least, those who think that the standards of instrumental rationality are those most relevant to choice under moral uncertainty. However, even those who deny that instrumental rationality is all there is to rationality tout court may find something of value here. An adequate account of instrumental rationality under moral uncertainty, may help to locate any perceived problems with the choices recommended by this account. For example, there maybe something substantively irrational about the ends or means-end beliefs that instrumental rationality takes as its input.

Expected utility theory, as a theory of instrumental rationality, gives a central role to agents' preferences; utility is a measure of the degree to which some states of affairs are preferred to others. However, despite being an attempt to emulate decision theories built on preferences and utility, MEC works with choiceworthiness values that represent the evaluation of alternatives according to different moral theories. Its proponents make the connection between moral choiceworthiness and individual preference with the notion of moral conscientiousness (MacAskill et al., 2020). The idea is that a morally conscientious agent will have a utility function that tracks the choiceworthiness function of whichever moral theory is being considered. But the relation between morality on the one hand and desire, preference, or utility, on the other is far more complex than this conception of moral conscientiousness allows.

There is a rich history of philosophical debate about this relation, and it has important implications for rational choice under moral uncertainty. Amartya Sen, for example, has criticised expected utility theory for its perceived inability to accommodate behaviour motivated by considerations other than agents' self-interest (Sen, 1977). He claims that each agent has their own, true preferences and that moral commitments motivate us to adopt different preference orderings in practice. This alternative preference ordering is arrived at through a process of compromise between an agent's preferences and their moral commitments. Sen therefore argues that morally

motivated actions are an example of counter-preferential choice, so cannot be accommodated by expected utility theory, which mandates and predicts choice in line with agents' own preferences.

Daniel Hausman, on the other hand, has defended rational choice theory from Sen's critique by arguing for a broader conception of preferences (Hausman, 2005). He claims that an agent's preferences should be thought of as all-things-considered evaluative judgements, which incorporate narrow self-interest, moral commitments, and anything else that is deemed relevant by the lights of the agent in question. He agrees with Sen's claim that models of rational choice should be sensitive to a broader range of considerations than merely narrow self-interest, but argues that this can be achieved by a single, richer conception of preference. In other words, the preferences that you end up adopting through compromise between your own self-interested preferences and your commitments simply *are* your preferences.

Rather than adjudicate on this debate here, I want to highlight a point of agreement between Sen and Hausman: that rational agents undertake some process of compromise between their own self-interest and their moral commitments when comparing alternatives, to arrive at the preference ordering on which they will act. For present purposes, nothing much hinges on whether we consider these preferences the agent's own, or think of this as a kind of counter-preferential choice.

Note also that a great deal is being left unsaid about the compromise between self-interest and moral commitment, a matter that is the subject of numerous contentious points in the theorisation of moral motivation. But recall that it is instrumental rationality that is at issue here. This is a form of structural rationality: a matter of whether one's attitudes and actions hang together in the right way. Structural rationality does not require agents to have any particular degree of moral conscientiousness. If I am less moved by certain moral considerations than another person, so end

up adopting overall preferences that are more closely aligned to my narrowly self-interested preferences, I am not being structurally irrational no matter how morally criticisable I may be.

With these ideas in mind, let us turn our attention back to MEC. It seems on the face of it as though its proponents have in mind an agent for whom moral commitments entirely determine the preference ordering that is adopted, with no ground ceded in a compromise between morality and self-interest. If an agent were certain of a particular moral theory, MEC says that they would be irrational unless they acted exactly as this theory prescribed. Recall that the proponents of MEC claim that it is the right account of rational choice for morally conscientious agents, where this is taken to mean that they *'prefer doing right to doing wrong and are indifferent between different right-doings'* (MacAskill et al., 2020). But, as I have suggested, this sort of moral motivation should not be taken to be a requirement of instrumental rationality, nor is it a realistic characterisation of living, breathing people. Moral conscientiousness is not a binary notion, but rather comes in degrees: it may be thought of as the degree of compromise an agent is willing to make between their self-interest and moral commitments, insofar as these diverge. The degree of moral conscientiousness between individuals is highly variable and we should neither assume, nor require that agents' motivations are solely moral. So, either MEC is reaching beyond its proper remit as a theory of instrumental rationality, by *requiring* agents to care only about morality, or it is unrealistic, by *assuming* that agents care only about morality.

Perhaps this is simply an idealising assumption, used to isolate the question of how to make choices under moral uncertainty from other complicating factors. However, even if we think only of agents who are completely morally conscientious, there remains another crucial difference between MEC and the approaches of Sen and Hausman to incorporating moral commitments in models of rational choice. While the latter use a utility function that represents the agent's preferences, the former uses choiceworthiness functions that represent the evaluations of the moral theories in

question. There is a difference between the choiceworthiness of an option according to a moral theory, and the desirability of an option to an agent, on the supposition that a particular theory is true.[5]

Decision theorists aim to model agents' preferences or desires with functions that assign values to options. If an agent is morally conscientious, then these values may track the values of the theory that they are considering. But the values nonetheless represent that agent's preferences. If an agent is less than fully conscientious, the moral theory may still play a crucial role in influencing their preferences. But this does not mean that the decision procedure that we adopt must take the values directly from the theory in question as its input. This is precisely what MEC does, and it is this feature that lies at the root of the aforementioned problems of limited scope, intertheoretic comparisons, and swamping effects.

The problem of limited scope is that not all theories can be accommodated in MEC, since not all theories can be numerically represented in the right way. Some theories do not provide any choiceworthiness values whatsoever and, of those that do, not all have the structure required to calculate expected choiceworthiness. The attempt to use values that represent the choiceworthiness of options according to theories is directly responsible for this problem, since no such values are available for some theories.

The problem of intertheoretic comparisons is that different theories have different choiceworthiness functions and these cannot always be compared, either because they refer to different kinds of quantity (the Reference Problem) or because they are unique only up to positive affine transformation and there is no way of fixing them on the same scale (the Scale Problem).

[5] For a detailed discussion of suppositional desirability, see (Bradley, 2017).

The attempt to use values that represent the choiceworthiness of options according to theories is responsible for this problem, because these values are not always comparable.

The problem of 'swamping' is that certain theories posit choiceworthiness differences that have an overwhelming effect on expected choiceworthiness, even when assigned very low credence. The attempt to use values that represent the choiceworthiness of options according to theories is responsible for this problem, because these values may be very large, but MEC is required to incorporate them.

There are two conclusions to be drawn from this section. The first is that MEC is only applicable for agents whose utility functions perfectly track the choiceworthiness functions that are taken to represent the evaluations of options according to the theories under consideration, i.e. those who are completely morally conscientious. Therefore, it is either unrealistic or overly demanding as a theory of instrumental rationality. The second is that the way in which this procedure takes these values as its input underlies three of the major problems facing the use of MEC under moral uncertainty.

5 The Cure

I have suggested that the problems facing MEC arise from assigning values to options according to choiceworthiness functions that represent answers to the question, '*how much value does this theory assign to this option?*' However, this is not the only way to instantiate the Analogical View. Instead, I propose a desirability function that represents an agent's preferences on the supposition that a given moral theory was true. That is, a preference ordering that is arrived at through a process of compromise between their self-interest and the recommendations of the moral theory under consideration, along with any other considerations they deem to be relevant to the choice at hand.

From this point onwards, much of the framework of MEC can be preserved. We can still represent a decision as comprising options $A_1,\ldots,A_m$, moral theories $T_1,\ldots,T_n$, a probability function P representing the agent's credence in those theories, and some values to be maximised in expectation. Now, however, the values should be thought of as a measure of the suppositional desires of the agent and are represented by a single desirability function. The desirability of A_1 on the supposition that T_1 is true is given by $d(A_1 \mid T_1)$. The expected value to be maximised is then given by:

$$ED(A_i) = \sum_{j=1}^{n} d(A_i|T_j) \cdot P(T_j)$$

To be clear, this desirability function should be taken to represent a real psychological quantity: desirability to the agent in question. This proposal, therefore, falls firmly under the umbrella of mentalist, rather than behaviouristic approaches to decision theory (Bermúdez, 2009; Buchak, 2013; Dietrich & List, 2016; Okasha, 2016; Pettit, 1991; Thoma, 2019). This should not come as any great surprise, since MEC itself is already a far stronger claim than the behaviouristic interpretation of expected utility theory. In fact, on this behaviourist interpretation, moral uncertainty would pose no particular challenge for rational choice, since this account does not require agents to assign values to alternatives at all, but merely to have preferences that satisfy the axioms of a representation theorem (Hicks, 2018). If anything, this might be seen as a limitation of behaviourism, since '*have transitive preferences*' is an utterly unhelpful response to the question, '*how should I act under moral uncertainty?*'

An important difference between MEC and the current proposal is that the former is only applicable to agents who are completely morally conscientious, but the latter neither assumes nor requires any particular degree of moral conscientiousness. When an agent is imagining that a given moral theory is true, they may be concerned only with morality, as MEC assumes, but they may

just as well hold a weaker moral commitment, reaching a different compromise with other considerations and assigning different values accordingly. The suppositional desires of an agent may perfectly align with the choiceworthiness functions of the moral theories under consideration. But they might also diverge quite significantly, if the recommendations of these theories are not all that matter to this individual's overall evaluation of the options. Therefore, while MEC is only applicable for agents who are morally conscientious in the precise way that MacAskill et al. (2020) suppose, the current proposal has no such restriction. The relaxation of complete moral conscientiousness will be relevant when considering the problems that undermine MEC, to which we will turn our attention next.

5.1 Incorporating Valueless Theories

The first problem for MEC was that its scope is limited to those choices in which an agent's degrees of belief are divided only between theories that can be represented by values on an interval scale. This excludes several prominent approaches to moral thought, to which many people give at least some credence. However, under the current proposal, there is nothing to restrict the scope in this way. No matter what structure a moral theory takes, an agent can imagine that it is true and then evaluate alternatives accordingly. Suppose I am considering whether to tell my friend the painful truth about their terrible singing voice, but am morally uncertain, giving some credence to the view that I should just do whatever makes my friend happiest overall and another view according to which there is an absolute moral prohibition against dishonesty (for example, the ninth Commandment: thou shalt not bear false witness). Even though the latter view provides no ordering or value assignment–it simply says that it is wrong to lie–I can entertain the supposition that this principle is correct, then consider how desirable the options are, all things considered, and assign values accordingly. In fact, something like this kind of reasoning takes place whenever someone makes a choice that requires them to make trade-offs between moral considerations and other factors that they care about. For example, when someone has to weigh up the environmental

impact of air travel versus its convenience, or how much time they are going to dedicate to volunteering, they can evaluate the all-things-considered desirability of the options to arrive at preferences that can be represented by a utility function, even if the relevant moral considerations do not themselves provide any ordering or value assignment.

This solution is not available to the proponents of MEC, since they require any moral theory in which the agent has some credence to provide the relevant choiceworthiness values and, as shown with the example of the Ten Commandments in §3.1, not all theories do this. Therefore, a theory built on subjective desirability, rather than moral choiceworthiness, provides a decision rule that does not suffer from the same scope limitation as MEC.

One might object that the move from choiceworthiness according to moral theories to suppositional desirability will not help, since preference orderings themselves might lack the requisite features. For example, one might follow Temkin (2012) and doubt that rational agents' preferences are always transitive, a necessary condition for representing them on an interval scale. This issue appears to be especially troubling in the context of moral uncertainty, since the problem arises even if one does not outright believe that it is rational to have intransitive preferences, but merely gives some credence to an intransitive moral theory. Indeed, this is a serious problem for MEC, given the conception of moral conscientiousness that its proponents adopt. The current proposal, however, drives a wedge between moral theories and suppositional desires, allowing room for agents to suppose that an intransitive moral theory is correct and still not have intransitive preferences. Of course, this does not entirely rule out the possibility of intransitive preferences, but rather avoids the conclusion facing MEC: that agents who have any credence in an intransitive moral theory are unable to use the decision procedure. In this sense, my proposal is in no worse a position in response to Temkin's arguments against transitivity than conventional expected value theory. This is still a challenge that must be met, but recall that the central claim of the Analogical

View is that *if* maximisation of expected value is the correct theory of rational choice under empirical uncertainty, *then* it is the correct theory of rational choice under moral uncertainty. This conditional cannot be rendered false if we deny the antecedent.

In fact, the current proposal is, if anything, a more natural realisation of the Analogical View than MEC. Proponents of MEC claim that the correct theory of rational choice under moral uncertainty should emulate the correct theory of rational choice under empirical uncertainty. It is odd, therefore, that they should devise a decision theory based on the choiceworthiness functions of moral theories, rather than a representation of agents' desires or preferences. Confining their proposal to 'morally conscientious agents' goes some way to explaining this, since such agents would, by definition, have utility functions that would track the choiceworthiness functions of the theory under consideration. But recall that it is not instrumentally irrational for an agent to care about things other than just morality. Incorporating agents' own preferences presents a more general theory and goes further towards providing an account of choice under moral uncertainty that is analogous to choice under empirical uncertainty.

5.2 Allowing Intertheoretic Comparisons

The second challenge facing MEC was the problem of intertheoretic comparisons: comparisons between different theories are required for calculations of expected choiceworthiness, but these comparisons cannot meaningfully be made, because the theories' choiceworthiness functions represent different kinds of quantity (the Reference Problem) and are not fixed to the same scale (the Scale Problem).

The Reference Problem does not apply to the current proposal, since the relevant measure represents a single psychological variable. We are not trying to compare units of, say, welfare and rights, like apples and oranges, but rather an agent's desires supposing one theory was true with

their desires supposing a different theory was true. In fact, one can remain neutral on the vexed question of whether intertheoretic comparisons of moral value or choiceworthiness are metaphysically possible; the current proposal simply has no need for them. To maximise expected desirability, one only needs to be able to compare the relative strength of desires.

Of course, this does still require making comparisons of desire across different suppositions. Richard Bradley (2017) has argued that attempts to make cross-suppositional comparisons involve a kind of confusion: we should think of a supposition not as a constituent of the object being evaluated, but rather as the standpoint one adopts when making the evaluation. Therefore, writes Bradley, "*we can meaningfully speak of* α *in comparison to* γ*, given that* β*, but not of* α *given that* β *in comparison to* γ *given that* δ." (p.100) However, it seems as though cross-suppositional comparisons of desirability are not really so mysterious. Indeed, Bradley opens his own discussion of conditional attitudes with the following example: "*I might find the prospect of going to the beach desirable, conditional on it being a sunny day, but not at all so, conditional on it being rainy.*" This example clearly involves making a comparison of desirability across different suppositions. Moreover, Bradley's claim that we should think of suppositions as standpoints of evaluation, rather than parts of the object being evaluated, applies to both conditional desires and conditional beliefs, but we have no trouble making comparisons of belief across different suppositions. It is not meaningless to ask whether it is more likely that it rains in London tomorrow, given that it rains in Leeds, than that a person has active tuberculosis, given that they had a positive Mantoux test. Likewise, it is not meaningless to ask whether an agent deems diverting the trolley to save 5 and kill 1, on the supposition that utilitarianism is true, more or less desirable than not diverting the trolley, on the supposition that Kantian deontology is true. These may be confusing or difficult questions to answer, because they involve a complex combination of possibilities and suppositions, but the fact that a comparison is difficult does not show that it is meaningless.

Next, the Scale Problem. Given that there is now only a single function, representing the suppositional desires of the agent, there are no different choiceworthiness functions to be commensurated. Note that this is closely related to what MacAskill et al. (2020) call a "universal scale" approach to intertheoretic comparisons. According to this approach, intertheoretic comparisons can meaningfully be made because different moral theories can be plotted onto a universal scale of choiceworthiness. It is the same scale for different theories, but exists independently of them. This is not quite the same as the current proposal, since the key idea here is to move away from measures of choiceworthiness, as understood by these authors. It is, however, a structurally similar approach, because the values attached to the options, which are inputs for the decision procedure, are already taken to be on a single scale. As MacAskill, et al. note, the two main challenges for universal scale accounts of intertheoretic comparisons are to provide (a) an adequate characterisation of this scale and (b) a justification for the claim that it exists. While open questions remain about whether these challenges can be met for the universal scale accounts offered by Ross (2006), Sepielli (2009), and MacAskill et al. (2020), there is nothing metaphysically suspect about the idea that agents have suppositional desires that vary in strength.

The move from functions representing choiceworthiness according to a theory, to functions representing desirability to an agent means that neither the Scale Problem nor the Reference Problem applies. The relevant comparative evaluations may be complex and difficult, but there is no theoretical barrier to overcome as there are with prior instantiations of the Analogical View.

5.3 Avoiding Swamping Effects

The last problem facing MEC is that calculations of expected choiceworthiness can be swamped by theories that posit such large choiceworthiness differences between some options that they dictate what the agent in question ought to do, even if that agent considers any such theories highly unlikely to be correct. This problem can be avoided on the current proposal by noting that one's

commitment to morality may depend on the moral theory that is under consideration. Under MEC, we have to use choiceworthiness values that represent the theories in question. If one of these theories considers the matter at hand sufficiently high-stakes, then the choice will be dictated by that theory, even if the agent has little credence in it. On the current proposal, however, we do not have to use whatever choiceworthiness values are provided by the theory in question. Rather, one assigns values based on the degree to which one would be swayed by a theory supposing it was true. Given that instrumental rationality does not require any particular degree of commitment to moral considerations, agents' desirability functions need not include differences that are large enough to swamp this choice procedure, even when considering theories that make outlandish claims about choiceworthiness differences between some options. There is nothing instrumentally irrational about one's commitment to a moral view being sensitive to what that moral view says.

It may seem overly permissive to have no further constraints on agents' moral motivation. But such constraints are simply not in the domain of instrumental rationality, which is the matter at issue here. The aim is to provide guidance to agents who want to know which option represents the best means to satisfy their preferences, given their credences. Of course, this does not guarantee that such agents will be free from criticism for the preferences they have and the choices they make as a result. Morally speaking, one should be willing to do whatever is morally right. Perhaps the same can be said in terms of substantive rationality, if one is convinced by the arguments for 'moral rationalism' from Korsgaard (2009), Portmore (2011), or Smith (1994). However, these views are not in conflict with the proposal I offer here for an account of instrumental rationality under moral uncertainty.

It may also be that swamping effects persist for a particular agent, if there are sufficiently large differences between their desires for different options under a given supposition. But the problem with swamping is not simply that there are *some* situations in which low-probability, high-stakes

views end up determining which choices are rational. This is a feature of any expectational decision theory. Rather, the problem for MEC is that *all* rational agents are subject to swamping as soon as they give any credence to a sufficiently high-stakes view. And, given that it is very difficult to be certain that all such views are false, these swamping effects will be pervasive. On my proposal, however, swamping will only occur for agents who have a very particular set of suppositional desires. This means that (a) swamping does not have the same ubiquity as for MEC and (b) in the cases in which it does occur, it does not seem so problematic: if an agent genuinely has such extreme suppositional desires, then it is not clear why their decisions should not be highly sensitive to them.

While avoiding the swamping problem, the account I have offered is still sensitive to stakes. Recall that there is something of a double-edged sword, with sensitivity to stakes on one side and swamping effects on the other. The current proposal captures the advantages of the former without succumbing to the exorbitant demands of the latter, since an agent's preferences can still be influenced by how much is at stake according to the different moral theories they are considering, but they are not required to assign whatever value differences these theories provide. Such an agent is sensitive to how much is at stake, but not sensitive enough to be unavoidably subject to swamping effects.

This reasoning might be subject to an objection from motivational internalism, the view that moral judgements are necessarily or intrinsically motivating. For the motivational internalist, if a rational agent were to judge a high-stakes theory to be true then they would have a corresponding motivation to pursue the action that it recommends. Swamping by such theories could not so easily be avoided by using judgements of subjective desirability rather than moral choiceworthiness, if the former tracked the latter as motivational internalism suggests. In response to this point, it will

be instructive to distinguish between two forms of motivational internalism.[6] Weak motivational internalism holds that moral judgements necessarily provide some degree of motivation. Judging that an action is right always provides *some* motivating reason to pursue it, but this can be overridden by countervailing reasons. This form of motivational internalism is consistent with the proposal for avoiding swamping. Although weak motivational internalism requires that an option that is judged to be morally right is pushed towards the top of one's preference ordering, the strength of motivation provided by a moral judgement need not be equivalent to exactly how much is at stake according to that view.

Strong motivational internalism, on the other hand, holds that moral judgements necessarily provide *overriding* motivation to act as they demand. This is a view that J. L. Mackie ascribed to Plato: of Plato's Forms, especially the Form of the Good, Mackie says, '*just knowing them or "seeing" them will not merely tell men what to do but will ensure that they do it, overruling any contrary inclinations*' (Mackie, 1977, p.23). It is true that strong motivational internalism, when coupled with the current approach, results in swamping. This means that there is conflict between three claims: (i) strong motivational internalism, (ii) the Analogical View, and (iii) the view that swamping is an unacceptable feature of an account of choice under moral uncertainty. But it is not immediately clear why the presence of this conflict should lead us to reject the Analogical View, rather than one of these other claims. This argument could just as well be used to show that strong motivational internalism entails an implausible conclusion, or that we should reconsider our aversion to swamping. In fact, strong motivational internalism and the assertion that swamping is a problem seem like the oddest bedfellows of the three. If one believes that moral judgements necessarily provide indefeasible motivation, then it is apt to think that the possibility of a high-stakes moral view should determine how one acts. So, just like the other problems before it,

[6] (Mason, 2008), drawing on (Brink, 1986), makes a similar distinction, but the views referred to here as strong and weak internalism are labelled by Mason as "weak internalism" and "weakest internalism". See also (C. M. Korsgaard, 1986) and (Smith, 1994) for detailed discussions of, and arguments for, related forms of internalism.

counter-intuitive swamping effects can be avoided by assigning values to options according to all-things-considered desirability, rather than moral choiceworthiness.

6 Conclusion

The question, '*how much value does this theory assign to this option?*' is not the same as the question, '*how much do I desire this option, supposing this theory is true?*' It is this method of value ascription via choiceworthiness according to theories, rather than desirability on the supposition that a theory is true, that I have argued is the root cause of three major problems for MEC. The current proposal can avoid these problems, while preserving MEC's advantageous features and thereby presents a superior instantiation of the Analogical View. Furthermore, it provides a role for agents' own preferences or desires that is conspicuous in its absence from MEC, and generalises the Analogical View to agents who do not care only about morality. Therefore, we can treat moral uncertainty as analogous to empirical uncertainty when modelling rational choice.

Wordcount: 7536

Acknowledgements: I am grateful to Richard Bradley, Campbell Brown, Johanna Thoma, Liam Kofi Bright, Krister Bykvist, and Andrew Sepielli for discussing earlier drafts of this work, and to two anonymous reviewers for *Ergo*.

References

Aboodi, R. (2022). Normative Uncertainty without Unjustified Value Comparisons. *Journal of Ethics and Social Philosophy*, *21*(3). https://doi.org/10.26556/JESP.V21I3.1492

Arrow, K. J. (1971). *Essays in the Theory of Risk-Bearing*. Markham Publishing Co.

Beckstead, N., & Thomas, T. (2023). A paradox for tiny probabilities and enormous values. *Noûs*, 1–25. https://doi.org/10.1111/nous.12462

Bermúdez, J. L. (2009). *Decision Theory and Rationality*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199548026.001.0001/acprof-9780199548026-chapter-3?print

Bostrom, N. (2011). Infinite Ethics. *Analysis and Metaphysics*, *10*, 9–59.

Bradley, R. (2017). *Decision Theory With a Human Face*. Cambridge University Press.

Brink, D. O. (1986). Externalist moral realism. *The Southern Journal of Philosophy*, *24*(1 S), 23–41. https://doi.org/10.1111/J.2041-6962.1986.TB01594.X

Broome, J. (2012). *Climate Matters: ethics in a warming world*. W. W. Norton & Company.

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

Buchak, L. (2014). Risk and Tradeoffs. *Erkenntnis*, *79*(S6), 1091–1117.

Bykvist, K. (2014). Evaluative Uncertainty, Environmental Ethics, and Consequentialism. In A. Hiller, R. Ilea, & L. Kahn (Eds.), *Consequentialism and Environmental Ethics* (pp. 122–135). Routledge.

Bykvist, K. (2017). Moral Uncertainty. *Philosophy Compass*, *12*(3).

Carr, J. R. (2020). Normative Uncertainty without Theories. *Australasian Journal of Philosophy*, *98*(4), 747–762. https://doi.org/10.1080/00048402.2019.1697710

Dietrich, F., & Jabarian, B. (2022). Decision under normative uncertainty. *Economics and Philosophy*, *38*(3), 372–394.

Dietrich, F., & List, C. (2016). Mentalism Versus Behaviourism in Economics: A Philosophy-of-Science Perspective. *Economics and Philosophy*, *32*(2), 249–281.

Fogal, D. (2020). Rational Requirements and the Primacy of Pressure. *Mind*, *129*(516), 1033–1070. https://doi.org/10.1093/MIND/FZZ038

Fogal, D., & Worsnip, A. (2021). Which Reasons? Which Rationality? *Ergo*, *8*(11). https://doi.org/10.3998/ERGO.1148

GiveWell. (2023, November). *Our Top Charities*. https://www.givewell.org/charities/top-charities

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

Gustafsson, J. E., & Torpman, O. (2014). In Defence of My Favourite Theory. *Pacific Philosophical Quarterly*, *95*(2), 159–174.

Harman, E. (2015). The Irrelevance of Moral Uncertainty. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics, Volume 10* (Vol. 10, pp. 53–79). Oxford University Press Oxford.

Hausman, D. (2005). Sympathy, Commitment, and Preference. *Economics and Philosophy*, *21*(1), 33–50.

Hedden, B. (2016). Does MITE Make Right? On Decision-Making under Normative Uncertainty. In R. Schafer-Landau (Ed.), *Oxford Studies in Metaethics Volume 11* (pp. 102–128). Oxford University Press.

Hicks, A. (2018). Moral Uncertainty and Value Comparison. In R. Shafer Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 13, pp. 161–183). Oxford University Press.

Hooker, B., & Streumer, B. (2003). Procedural and substantive practical rationality. In P. Rawling & A. R. Mele (Eds.), *The Oxford Handbook of Rationality* (pp. 57–74). Oxford University Press. https://doi.org/10.1093/0195145399.003.0004

Hudson, J. L. (1989). Subjectivization in Ethics. *American Philosophical Quarterly*, *26*(3), 221–229.

Hume, D. (2007). *A Treatise of Human Nature (1739)* (D. F. Norton & M. J. Norton, Eds.). Clarendon Press.

Joyce, J. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

Korsgaard, C. (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press.

Korsgaard, C. M. (1986). Skepticism about Practical Reason. *The Journal of Philosophy*, *83*(1), 25. https://doi.org/10.2307/2026464

Lockhart, T. (2000). *Moral Uncertainty and its Consequences*. Oxford University Press.

MacAskill, W. (2014). *Normative Uncertainty*. DPhil thesis, University of Oxford.

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

MacAskill, W., & Ord, T. (2020). Why Maximize Expected Choice-Worthiness? *Noûs*, *54*(2), 327–353.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. Penguin Books.

Mason, E. (2008). An argument against motivational internalism. *Proceedings of the Aristotelian Society*, *108*, 135–156. https://doi.org/10.1111/J.1467-9264.2008.00240.X

Nissan-Rozen, I. (2015). Against Moral Hedging. *Economics & Philosophy*, *31*(3), 349–369.

Nover, H., & Hájek, A. (2004). Vexing Expectations. *Mind*, *113*(450), 237–249. https://doi.org/10.1093/mind/113.450.237

Okasha, S. (2016). On the Interpretation of Decision Theory. *Economics and Philosophy*, *32*(3), 409–433.

Pettit, P. (1991). Decision Theory and Folk Psychology. In M. Bacharach & S. Hurley (Eds.), *Essays in the Foundations of Decision Theory* (pp. 147–175). Blackwell.

Portmore, D. (2011). *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford Univserity Press.

Riedener, S. (2019). Constructivism about Intertheoretic Comparisons. *Utilitas*, *31*(3), 277–290. https://doi.org/10.1017/S0953820819000165

Riedener, S. (2020). An axiomatic approach to axiological uncertainty. *Philosophical Studies*, *177*(2), 483–504. https://doi.org/10.1007/S11098-018-1191-7

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

Ross, J. (2006). Rejecting Ethical Deflationism. *Ethics*, *116*(4), 742–768.

Samuelson, P. A. (1977). St. Petersburg paradoxes: Defanged, dissected, and Historically Described. *Journal of Economic Literature*, *15*(1), 24–55. https://www.jstor.org/stable/2722712

Scanlon, T. (2007). Structural Irrationality. In G. Brennan, R. Goodin, & M. Smith (Eds.), *Common Minds: Themes From the Philosophy of Philip Pettit* (pp. 84–103). Clarendon Press.

Sen, A. (1977). Rational Fools: A Critique of the Behavioural Foundations of Economic Theory. *Philosoph & Public Affairs*, *6*(4), 317–344.

Sepielli, A. (2009). What To Do When You Don't Know What To Do. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics, Volume IV* (Vol. 4, pp. 5–28). Oxford University Press.

Sepielli, A. (2010). *"Along an Imperfectly-Lighted Path": Practical Rationality and Normative Uncertainty*. Doctoral Thesis, Rutgers University.

Sepielli, A. (2014). What to Do When You Don't Know What to Do When You Don't Know What to Do ... *Noûs*, *48*(3), 521–544.

Smith, M. (1994). *The Moral Problem*. Blackwell.

Tarsney, C. (2018a). Intertheoretic Value Comparison: A Modest Proposal. *Journal of Moral Philosophy*, *15*(3), 324–344.

Tarsney, C. (2018b). Moral Uncertainty for Deontologists. *Ethical Theory and Moral Practice*, *21*(3), 505–520. https://doi.org/10.1007/S10677-018-9924-4/TABLES/6

Temkin, L. S. (2012). *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford University Press.

Thoma, J. (2017). *Advice for the Steady: Decision Theory and the Requirements of Instrumental Rationality*. PhD Thesis, University of Toronto.

Thoma, J. (2019). Decision Theory. In R. Pettigrew & J. Weisberg (Eds.), *The Open Handbook of Formal Epistemology* (pp. 57–106). PhilPapers Foundation.

Weatherson, B. (2014). Running Risks Morally. *Philosophical Studies*, *167*(1), 141–163.

Weatherson, B. (2019). *Normative Externalism*. Oxford University Press.

Wilkinson, H. (2022). In Defense of Fanaticism. *Ethics*, *132*(2), 445–477. https://doi.org/10.1086/716869

Worsnip, A. (2021). *Fitting Things Together: Coherence and the Demands of Structural Rationality*. Oxford University Press.