Second Thoughts about My Favourite Theory[⋆]

Johan E. Gustafsson[†]

ABSTRACT. A simple way to handle moral uncertainty is to follow the moral theory in which one has the most credence. This approach is known as My Favourite Theory. One, alleged, advantage of this approach is that, unlike some rival views, it is immune to value pumps. In this paper, I argue that My Favourite Theory is vulnerable to similar dynamic problem. I argue that My Favourite Theory prescribes, sequentially, choices that are worse in expected moral value according to each moral theory you have any credence in. That is, there are situations where one must, in order to follow My Favourite Theory, follow a plan that has a worse expectation of moral value than some other available plan according to every moral theory in which one has some credence. Moreover, this argument generalizes to other approaches that avoid intertheoretic comparisons of value, such as My Favourite Option, the Borda Rule, and the Principle of Maximizing Expected Normalized Moral Value.

There are lots of moral theories. Like most philosophers, you probably have a favourite among them. But you're not sure that theory is correct: You find that some of its rivals have at least some plausibility. Suppose, then, that you have at least some credence in two moral theories. What do you do when these theories prescribe different, incompatible courses of action? For instance, it might be that, according to one of these theories, you *ought* to do some act, but, according to another of these theories, you *ought not* to do that act.

A simple way to deal with this kind of moral uncertainty is to just follow the moral theory in which you have the most credence. This approach is known, disparagingly, as

⋆ Forthcoming in *Pacific Philosophical Quarterly*.

† I would be grateful for any thoughts or comments on this paper, which can be sent to me at *johan.eric.gustafsson@gmail.com*.

> *My Favourite Theory* An option is a morally conscientious choice for a person P in a situation S if and only if this option is permitted in S by a moral theory such that there is no moral theory in which P in S has more credence.[1]

My Favourite Theory captures the way that many people deal with moral uncertainty: They identify with the theory they like best (they call themselves Kantians, libertarians, utilitarians, and so on) and they aim to follow the dictates of that theory.

A standard objection to My Favourite Theory is that it's insensitive to the relative sizes of the stakes in terms of moral value on different moral theories. Suppose, for instance, that you only have credence in two moral theories, T_1 and T_2. And you have slightly more credence in T_1 than in T_2. More precisely, there is some positive constant ϵ less than $1/4$ such that your credence in T_1 is $1/2+\epsilon$ and your credence in T_2 is $1/2-\epsilon$. Consider the following case:[2]

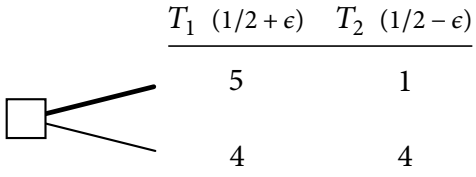
[1] Gracely 1996, p. 331. The name 'My Favourite Theory' is due to Lockhart (2000, p. 42). Gustafsson and Torpman (2014, pp. 167–170) put forward some more complicated versions of My Favourite Theory to better handle different kinds of ties between options and between theories. These complications, however, won't matter for our discussion. This so, because, in the cases we will consider, there is never any tie between theories (at each choice node, there is a theory that you have more credence in than any other theory) and there is never any tie between options (at each choice node, the theory with the highest credence prescribes a single option). So these complications cannot block the objection that's raised in this paper, because Gustafsson and Torpman's more complicated versions prescribe the same options in the cases we will consider as the simpler version stated here. Tarsney (2017, pp. 215–219) puts forward a similar view, called 'Course-Grained My Favourite Theory', which says that one should

> Aggregate theories as far as content-based comparisons will allow, and when no further content-based comparisons are possible, do what the most plausible comparability class of theories says, in aggregate.

Assuming that all moral theories belong to different comparability classes in the cases we shall discuss, the argument of this paper also works against Course-Grained My Favourite Theory.

[2] Hudson 1989, p. 224, Lockhart 2000, p. 84, and Gustafsson and Torpman 2014, p. 160.

Case One



Here, the square represents a choice, where one can either go up or go down. Suppose that T_1 and T_2 are maximizing moral theories.[3] Going up has a value of 5 according to T_1 and a value of 1 according to T_2, whereas going down has a value of 4 according to both theories.

The nerve of the different-stakes objection is that, since there is much more at stake on T_2 (the 3 unit difference of 4 and 1) than on T_1 (the 1 unit difference of 5 and 4), the higher stakes on T_2 should outweigh the slightly higher credence in T_1. Hence, according to this line of thought, the morally conscientious choice in Case One is to go down. My Favourite Theory, however, prescribes going up, because you have the most credence in T_1 and, according to T_1, going up is better than going down. (This prescription of My Favourite Theory is marked by the thicker line.)

The main rival to My Favourite Theory is

> *The Principle of Maximizing Expected Moral Value* An option is a morally conscientious choice for P in S if and only if this option has at least as great expected moral value for P in S than any other option in S.[4]

This approach yields a more compelling result in Case One. The expected moral value of going up is $(1/2+\epsilon)\cdot 5+(1/2-\epsilon)\cdot 1 = 3+4\epsilon$, and the expected moral value of going down is $(1/2+\epsilon)\cdot 4+(1/2-\epsilon)\cdot 4 = 4$. Since ϵ is less than 1/4, the Principle of Maximizing Expected Moral Value prescribes going down in Case One.

The different-stakes objection may seem fatal for My Favourite Theory and to be strong evidence in favour of the Principle of Maximizing Expected Moral Value. But there is a problem with this objection: The claim that the stakes are higher on T_2 than on T_1 relies on *intertheoretic comparisons of value*, that is, comparisons of the values of options according to one moral theory with the values of options according to another

[3] Moreover, we shall assume that every moral theory in the examples we shall discuss measure moral value on an interval scale; see Roberts 1979, pp. 64–65.

[4] A variation of this approach was put forward by Lockhart (2000, p. 82).

moral theory. Such comparisons seem arbitrary because moral theories typically don't say how their evaluations compare to those of other theories.[5] So having credence in two or more moral theories does not seem to commit us to any particular exchange rate between the units of moral value in these theories.[6] If we cannot make sense of the claim that the stakes are higher on T_2 than on T_1, then the different-stakes objection to My Favourite Theory cannot get off the ground.

Part of the appeal of My Favourite Theory is that it doesn't need any intertheoretic comparisons of value. The reliance on such intertheoretic comparisons is the main drawback of the Principle of Maximizing Expected Moral Value. Without intertheoretic comparisons of value, we can't calculate the intertheoretic expectations of moral value which the Principle of Maximizing Expected Moral Value relies on.

A further, *alleged*, advantage of My Favourite Theory is that, unlike many of its rivals, it is immune to value pumps.[7] In this paper, I will argue that My Favourite Theory is vulnerable to a similar dynamic problem. I will argue that My Favourite Theory prescribes, sequentially, choices that are worse in expected moral value according to each moral theory you have any credence in. That is, there are situations where one must, in order to follow the approach, follow a plan that has a worse expectation of moral value than some other available plan according to every moral theory in which one has some credence.[8] What's more, the argument gener-

[5] Consider, for example, Total and Average Utilitarianism. Equating the difference of one unit of average well-being with the difference of one unit of total well-being is implausible, since it would make Average Utilitarianism count for almost nothing compared to Total Utilitarianism for most choices; see Broome 2012, p. 185. And any other exchange rate between the two theories seems completely arbitrary and even more implausible. For discussions of potential solutions to the problem of intertheoretic comparisons of value, see Ross 2006, pp. 761–765 and Gustafsson and Torpman 2014, pp. 160–165.

[6] The problem is not whether there is an overarching scale of moral value to which we could convert the moral value from all moral theories. The problem is merely whether there are any non-arbitrary exchange rates between the theories. Consider an analogy with monetary currencies. The exchange rates between national currencies do not rely on any conversions to some overarching international currency.

[7] Gustafsson and Torpman 2014, pp. 160, 172. MacAskill et al. (2020, p. 104) questions the cogency of value-pumps arguments, pointing approvingly to Ahmed's (2017) self-regulation response to value pumps. Yet Ahmed's approach can be rebutted with a very minimal form of backward induction; see Gustafsson and Rabinowicz 2020, p. 585n13.

[8] Most of the other commonly raised objections to My Favourite Theory aren't, I think, very worrying. The most discussed problem is that My Favourite Theory is sensitive to the individuation of moral theories. This problem can be solved by combining the approach with a principle of theory individuation, such as the following (suggested

alizes to other approaches that avoid intertheoretic comparisons of value, such as My Favourite Option, the Borda Rule, and the Principle of Maximizing Expected Normalized Moral Value.

1. The problem of future moral progress

Suppose that you start off with a 1/2 credence in each of two mutually exclusive moral theories, T_1 and T_2. That is, you're certain that one of these theories is correct, but you find them equally likely to be so. These theories are maximizing theories; they prescribe, in each situation, the option with the best outcome. Suppose, in addition, that you know that you will soon make moral progress in the sense that you will soon learn something new that will make one of T_1 and T_2 seem more credible than the other but you currently don't know which. You find it equally likely that the news you are about to receive will favour T_1 as that it will favour T_2. Let ϵ be the size of this foreseen change in your credences and suppose that the shift in your credences between T_1 and T_2 will be less than 1/4, that is, we suppose that $0 < \epsilon < 1/4$.

Now, consider the dynamic case depicted in the following diagram,

---

in Gustafsson and Torpman 2014, p. 171):

> *The Principle of Fine-Grained Individuation* Regard moral theories T and T' as versions of the same moral theory if and only if you are certain that you will never face a situation where T and T' yield different prescriptions.

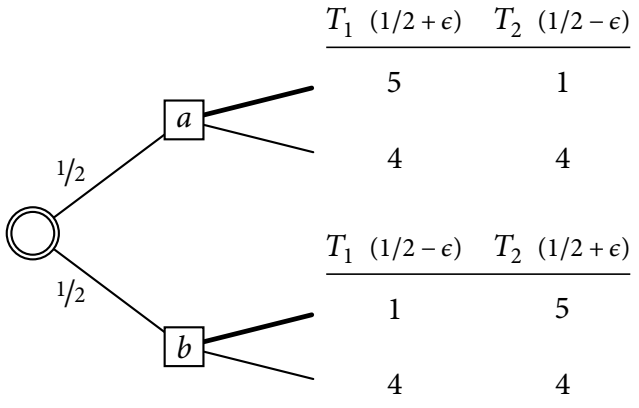
MacAskill (2014, p. 25; 2016, p. 975n18; 2020, p. 9) proposes a counter-example where one is almost certain in prioritarianism but has some credence in utilitarianism. One is unsure about the shape of the relevant concave function, so one's credence is split between a lot different versions of prioritarianism, each with less credence than utilitarianism, and these versions might make different prescription in some future choices. Hence these versions of prioritarianism should be treated as different theories. Suppose that all versions of prioritarianism recommend one option and utilitarianism another, the above approach would still recommend that one follows utilitarianism.

But this objection implicitly relies on either (i) My Favourite Option—defined later—or (ii) the intuition that these version of prioritarianism should be regarded as the same theory. If the objection relies on (i), it runs into the same value-pumps as My Favourite Option; see Gustafsson and Torpman 2014, pp. 165–166. If the objection relies on (ii), it suggests that there is a favoured way of individuating theories and hence that there must be some other non-arbitrary principle of individuation And then My Favourite Theory could be combined with that non-arbitrary principle of individuation instead.

Most other objections to My Favourite Theory rely on, seemingly arbitrary, intertheoretic comparisons of value—see, for example, Hudson 1989, p. 224, Lockhart 2000, p. 84, Hedden 2016, p. 106, and MacAskill and Ord 2020, p. 11.

where the tables (one for each of the two choice situations) give the value of each outcome according to each of the moral theories, with your credence for each theory in that situation in parenthesis:

Case Two



The double circle represents a learning node, and the squares represent choice nodes. The initial learning node models the uncertainty about which theory you will come to have more credence in. A *learning node* differs from a standard chance node in that, at a learning node, no random event occurs—the agent merely learns a piece of information from a set of alternative pieces of information. But the agent has credences in advance about which of the alternative pieces of information they will receive, just like agents have credences about how standard chance nodes will resolve.

At the initial learning node, there is, for each of the two choice nodes, a 1/2 chance that you will face that choice node. At each of the choice nodes, you have a choice between going up and going down. At choice node *a*, you have slightly more credence in T_1 than in T_2; more precisely, you have a $1/2 + \epsilon$ credence in T_1 and a $1/2 - \epsilon$ credence in T_2. And, at choice node *b*, you have slightly more credence in T_2 than in T_1; more precisely, you have a $1/2 + \epsilon$ credence in T_2 and a $1/2 - \epsilon$ credence in T_1.

Note that the learning node just represents your uncertainty about which theory you will get evidence for. The learning node does not depend on a random event. There is a probabilistic dependence between how the learning node resolves and the two moral theories. Suppose that you know from reliable sources that a newly published paper has made a breakthrough in the debate regarding T_1 and T_2 and contains a new, compelling argument in favour of one of T_1 and T_2, but you don't know

which. The learning node just represents your uncertainty that you regard it as equally likely that the new argument favours T_1 as that it favours T_2. When the learning node resolves, you learn which theory the new argument supports and you adjust your credence in the two theories accordingly. At the initial node, your conditional credence in T_1 given that T_1 *is supported by the new argument* is $1/2 + \epsilon$. And your conditional credence in T_2 given that T_2 *is supported by the new argument* is also $1/2 + \epsilon$. But your unconditional credence in each of T_1 and T_2 is still $1/2$.

Let a plan at a node n be a specification of what to choose at each choice node that can be reached from n. Let us say that one follows a plan at node n' if and only if, for each choice node n'' that can be reached from n', one would choose in accordance with the plan if one were to face n''. Finally, let us say that a plan is available at a node n if and only if the plan can be followed in n.

In order to follow My Favourite Theory, one must follow *the Up-Up Plan*, that is, the plan of going up at choice node a and going up at choice node b. At choice node a, you have the most credence in T_1 and going up has a value of 5 and going down has merely a value of 4 according to T_1. Accordingly, My Favourite Theory prescribes going up at choice node a. And, at choice node b, you have the most credence in T_2 and going up has a value of 5 and going down has a value of 4 according to T_2. Accordingly, My Favourite Theory also prescribes going up at choice node b. (Like before, the prescriptions of My Favourite Theory are marked by the thicker lines.)

Consider the expectation of moral value at the initial learning node of the Up-Up Plan conditional on each theory. Let *Learn Up* denote that the learning node resolves upwards, that is, that you get information that favours T_1. And let *Learn Down* denote that the learning node resolves downwards, that is, that you get information that favours T_2. The expectation of moral value for the Up-Up Plan conditional on T_1 is $P(\text{Learn Up} \mid T_1) \cdot 5 + P(\text{Learn Down} \mid T_1) \cdot 1$. We define the conditional credence of C given A, where $P(A) > 0$, in the usual way:

$$P(C \mid A) =_{\text{df}} \frac{P(A \,\&\, C)}{P(A)}.$$

So we have

$$P(T_1 \,\&\, \text{Learn Up}) = P(T_1 \mid \text{Learn Up}) \cdot P(\text{Learn Up}).$$

Since $P(T_1 \mid \text{Learn Up}) = 1/2 + \epsilon$ and $P(\text{Learn Up}) = 1/2$, we have that $P(T_1 \ \& \ \text{Learn Up}) = (1/2+\epsilon)\cdot 1/2$. And, since $P(T_1) = 1/2$, we then have

$$\begin{aligned} P(\text{Learn Up} \mid T_1) &= \frac{P(T_1 \ \& \ \text{Learn Up})}{P(T_1)} \\ &= \frac{(1/2 + \epsilon) \cdot 1/2}{1/2} \\ &= 1/2 + \epsilon. \end{aligned}$$

Note that, while the unconditional credence $P(\text{Learn Up})$ is $1/2$, the conditional credence $P(\text{Learn Up} \mid T_1)$ is, as we have just seen, $1/2 + \epsilon$. Similarly, since $P(T_1 \mid \text{Learn Down}) = 1/2 - \epsilon$ and $P(\text{Learn Down}) = 1/2$, we have that $P(T_1 \ \& \ \text{Learn Down}) = (1/2 - \epsilon) \cdot 1/2$. So

$$\begin{aligned} P(\text{Learn Down} \mid T_1) &= \frac{P(T_1 \ \& \ \text{Learn Down})}{P(T_1)} \\ &= \frac{(1/2 - \epsilon) \cdot 1/2}{1/2} \\ &= 1/2 - \epsilon. \end{aligned}$$

Hence the expectation of moral value for the Up-Up Plan conditional on T_1 is $(1/2 + \epsilon) \cdot 5 + (1/2 - \epsilon) \cdot 1 = 3 + 4\epsilon$. Because of symmetry, the expectation of moral value for the Up-Up Plan conditional on T_2 is the same.

Compare these theory-conditional expectations of the Up-Up Plan with the theory-conditional expectations of the opposite plan, namely, *the Down-Down Plan*—that is, the plan of going down at choice node *a* and going down at choice node *b*. The expectation of moral value for the Down-Down Plan conditional on T_1, or conditional on T_2, is $(1/2 + \epsilon) \cdot 4 + (1/2 - \epsilon) \cdot 4 = 4$. So, from the above calculations, we have that the expectations of moral value at the initial node for each of the theories must be the following:

Table One

	T_1	T_2
The Up-Up Plan	$3 + 4\epsilon$	$3 + 4\epsilon$
The Down-Down Plan	4	4

Since ϵ is less than $1/4$, we have that the Up-Up Plan has a worse expec-

tation conditional on each moral theory with positive credence than the Down-Down Plan. Therefore, since following My Favourite Theory requires following the Up-Up Plan, following My Favourite Theory forces you to violate

> *The Weak Principle of Theory-Conditional Plan Dominance* If X and Y are two available plans for a person P in situation S and, for each moral theory that P in S has some credence in, X has a worse expectation of moral value in S than Y conditional on that theory being true, then P does not follow X in S.[9]

The problem with violating this principle is that, when you violate it, you're certain that, regardless of which moral theory turns out to be correct, your plan will have a worse expectation of moral value than some other available plan. An adequate approach to moral uncertainty shouldn't lead to a lower expected moral value according to *every moral theory* in which you have some credence. Consider, once more, Table One. Your moral uncertainty consists in not knowing whether the moral expectations of the plans are those in the T_1 column or those in the T_2 column. Your moral uncertainty does not extend to how the expectations compare within each column. Your moral uncertainty does not prevent you from knowing that, for each moral theory, the moral expectations are greater for the Down-Down Plan than for the Up-Up Plan. Hence an adequate approach to moral uncertainty shouldn't require that one follows the Up-Up Plan, since your moral uncertainty doesn't blind you to the fact that the Up-Up Plan has a worse moral expectation than the Down-Down Plan. But following My Favourite Theory requires following the Up-Up Plan. So we should reject My Favourite Theory.

[9] This is a weak variant of the following, logically stronger, principle:

> *The Strong Principle of Theory-Conditional Plan Dominance* If X and Y are two available plans for a person P in situation S and, for each moral theory that P in S has some credence in, X has a at least as bad expectation of moral value in S as Y conditional on that theory being true and, for some moral theory that P in S has some credence in, X has a worse expectation of moral value in S than Y conditional on that theory being true, then P does not follow X in S.

While this principle is also plausible, it is stronger than necessary for the argument against My Favourite Theory. The Strong Principle of Theory-Conditional Plan Dominance may be the closest we can get to the Principle of Maximizing Expected Moral Value without relying on intertheoretic comparisons of value.

Note that this objection to My Favourite Theory does not rely on intertheoretic comparisons of value. The expectations of moral value in Table One don't rely on comparisons of value between moral theories. These expectations are calculated conditional on each theory being correct. So, when we calculate each of these expectations, we assume that one particular moral theory is true. Hence these expectations only rely on *intra*theoretic comparisons of value.

It may be objected that we could avoid these violations of the Weak Principle of Theory-Conditional Plan Dominance if we adopt resolute choice. *Resolute choice* is an approach to dynamic choices where one adopts a plan at the initial node and then, at later nodes, one sticks to this plan even if it's no longer optimal when it's evaluated at the later nodes.[10] At the initial node, the agent divides her credence between T_1 and T_2. If we apply My Favourite Theory to the available plans at the initial node, we see that neither of these theories permit the Up-Up plan: T_1 requires the Up-Down plan, and T_2 requires the Down-Up plan. Thus the Up-Up plan is not permitted by any of the moral theories in which the agent has some credence. So, far from being the plan one must follow in order to follow this resolute version of My Favourite Theory, the Up-Up plan would be prohibited.

There are, however, two problems with this resolute response. First, a version of My Favourite Theory which is combined with this resolute approach would, in Case Two, require that you ignore the new moral evidence you receive between the initial learning node and the choice nodes. It does not seem morally conscientious to ignore moral evidence. Second, on the resolute approach, there would be one point in time that is special in the sense that your plans are always calculated relative to that time. It's hard to see why one point in time would have this special status. Why would the expectation of a plan calculated relative to an earlier node or time have any special significance at a later choice node?

It may next be objected that it's strange to consider different plans at an initial learning node where you have no immediate choice to make. Note, however, that we could add an earlier choice between the decision tree in Case Two and another decision tree just like it. Then, at the new initial choice node, the plans that involve going up at each choice node after the initial choice node would still be dominated by the plans that involve going down at each choice node after the initial choice node. Still,

[10] McClennen 1990, p. 13.

My Favourite Theory would require that you follow one of the plans that involves going up at each choice node after the initial choice. Hence we still get the same problem.

2. Conditionalization versus imaging

It might seem strange that we calculate the theory-conditional expectations using conditional credence for chance going up (and for chance going down) at the initial node, rather than the unconditional credence, that is, 1/2. (If this doesn't seem strange to you, you are hereby permitted to skip ahead to the next section.) It may be objected that we should use imaging rather than conditionalization. To get the image of a credence distribution P on A, we transfer the credence of each world W where A is false to the world closest to W where A is true.[11] The crucial thing for our purposes is that the credences of the moves of the chance and learning nodes in our decision trees are unaffected by imaging on any of the moral theories in which one has some credence. So, for example, the credence of chance going up at the initial learning node of Case Two after imaging on T_1 is 1/2 (the same as the unconditional credence).

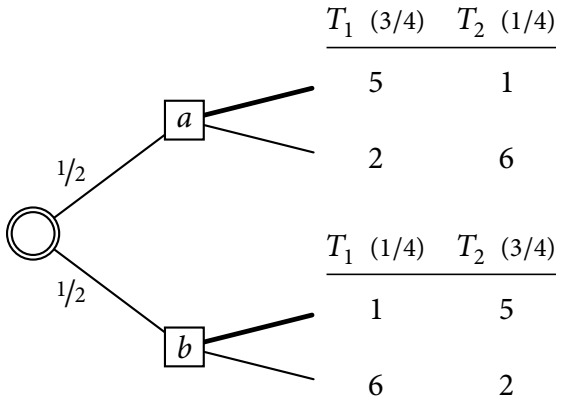
If we replace conditionalization with imaging in the Weak Principle of Theory-Conditional Plan Dominance, we get

> *The Weak Principle of Theory-Imaged Plan Dominance* If X and Y are two available plans for a person P in situation S and, for each moral theory that P in S has some credence in, X has a worse expectation of moral value in S than Y after imaging on that theory being true, then P does not follow X in S.

But this revised principle is implausible. It rules out the Principle of Maximizing Expected Moral Value, which is a plausible approach to moral uncertainty at least given that non-arbitrary intertheoretic comparisons of value can be made. To see this, assume that non-arbitrary intertheoretic comparisons of value can be made and consider

[11] Lewis 1976, pp. 310–311.

Case Three



Like before, the double circle represents a learning node, and the squares represent choice nodes. You start off with 1/2 credence in each of T_1 and T_2. You think that, at the learning node, it's equally likely you will get information that favours T_1 (the node resolves upwards) as that you will get information that favours T_2 (the node resolves downwards). After the learning node, your credence in the theory that is favoured by the evidence rises to 3/4.

At choice node *a*, the Principle of Maximizing Expected Moral Value prescribes going up, because the expected moral value of going up is $3/4 \cdot 5 + 1/4 \cdot 1 = 4$, whereas the expected moral value of going down is merely $3/4 \cdot 2 + 1/4 \cdot 6 = 3$. Similarly, the Principle of Maximizing Expected Moral Value also prescribes going up at choice node *b*. Hence following the Principle of Maximizing Expected Moral Value requires following the Up-Up Plan, that is, the plan of going up at choice node *a* and up at choice node *b*. (The prescriptions of the Principle of Maximizing Expected Moral Value are marked by the thicker lines.)

The expectation of the Up-Up Plan after imaging on T_1, or imaging on T_2, is $1/2 \cdot 5 + 1/2 \cdot 1 = 3$. And the expectation of the Down-Down Plan (that is, the plan of going down at choice node *a* and down at choice node *b*) after imaging on T_1, or conditional on T_2, is $1/2 \cdot 2 + 1/2 \cdot 6 = 4$. Hence the Principle of Maximizing Expected Moral Value violates the Weak Principle of Theory-Imaged Plan Dominance.

Note that the Principle of Maximizing Expected Moral Value does not violate the Weak Principle of Theory-Conditional Plan Dominance. In Case Three, the expectation of the Up-Up Plan conditional on T_1, or conditional on T_2, is $3/4 \cdot 5 + 1/4 \cdot 1 = 4$, And the expectation of the Down-Down Plan conditional on T_1, or conditional on T_2, is merely $3/4 \cdot$

$2 + 1/4 \cdot 6 = 3$. So following the Up-Up Plan doesn't violate the Weak Principle of Theory-Conditional Plan Dominance.

3. Other approaches without intertheoretic comparisons of value

Do other approaches in the literature which avoid intertheoretic comparisons of value fare any better than My Favourite Theory in Case Two? And do they satisfy the Weak Principle of Theory-Conditional Plan Dominance? As we shall see, they do not.

MY FAVOURITE OPTION

First, consider

> *My Favourite Option* An option is a morally conscientious choice for person P in situation S if and only if P in S has at least as high credence in this option being right as in every other option.[12]

Basically, My Favourite Option prescribes doing what is most likely to be right. In Case Two, the option that is most likely to be right at choice node a is to go up, since going up has a $1/2 + \epsilon$ chance of being right whereas going down merely has a $1/2 - \epsilon$ chance of being right. Likewise, the option that is most likely to be right at choice node b is to go up, since going up has a $1/2 + \epsilon$ chance of being right and going down has merely a $1/2 - \epsilon$ chance of being right. Hence following My Favourite Option requires following the Up-Up Plan. So, just like My Favourite Theory, My Favourite Option violates the Weak Principle of Theory-Conditional Plan Dominance.

THE BORDA RULE AND THE PRINCIPLE OF MAXIMIZING EXPECTED NORMALIZED MORAL VALUE

Next, we shall consider two approaches that compare options relative to a certain *Comparative Set*. We could specify this set in a number of ways. We start with the

[12] Lockhart (1992, pp. 35–36) defends a similar principle. The name 'My Favourite Option' is due to Gustafsson and Torpman (2014, p. 165).

The Availability Specification The Comparative Set in a situation is the set of available options in that situation.

Given this specification of the Comparative Set, consider

The Borda Rule

The *Borda Score* of option A in situation S according to theory T is equal to the number of available options in the Comparative Set which are, according to T, morally worse than A minus the number of options in the Comparative Set which are, according to T, morally better than A.

The *Credence-Weighted Borda Score* of an option A for person P in situation S is the sum, for all theories T, of the Borda Score of A according to T multiplied by the credence that P in S has in T.

An option is a morally conscientious choice for a person P in a situation S if and only if the option has an at least as high Credence-Weighted Borda Score for P in S as any alternative option in S.[13]

The idea is that each moral theory assigns a Borda Score to each option in the Comparative Set—that is, giving a score of 1 to the worst option, a score of 2 to the second worst, and so on. Then these scores are multiplied by the credence of the theory, and then these credence weighted scores are added up for each option. Finally, the Borda Rule prescribes the option with the greatest Credence-Weighted Borda Score, that is, the sum total of the credence weighted scores.

To see how this works, consider Case Two.

At choice node a, T_1 gives a value of 1 to going down and a value of 2 to going down, whereas T_2 gives a value of 1 to going up and a value of 2 to going down. The Credence-Weighted Borda Score for going up is then $1 \cdot (1/2 + \epsilon) + (-1) \cdot (1/2 - \epsilon) = 2\epsilon$. And the Credence-Weighted Borda Score for going down is then $(-1) \cdot (1/2 + \epsilon) + 1 \cdot (1/2 - \epsilon) = -2\epsilon$. So the Borda Rule prescribes going up at choice node a.

At choice node b, T_1 gives a value of 1 to going up and a value of 2 to going down, whereas T_2 gives a value of 1 to going down and a value of 2 to going up. The Credence-Weighted Borda Score for going up is then $(-1) \cdot (1/2 - \epsilon) + 1 \cdot (1/2 + \epsilon) = 2\epsilon$. And the Credence-Weighted Borda

[13] MacAskill 2016, p. 989 and MacAskill et al. 2020, p. 73.

Score for going down is then $1 \cdot (1/2 - \epsilon) + (-1) \cdot (1/2 + \epsilon) = -2\epsilon$. So the Borda Rule prescribes going up at choice node *b*.

Hence following the Borda Rule requires following the Up-Up Plan in Case Two. So the Borda Rule violates the Weak Principle of Theory-Conditional Plan Dominance.

Next, consider

> *The Principle of Maximizing Expected Normalized Moral Value*
>
> Normalize the value scales for each moral theory with positive credence so that the best option in the Comparative Set according to each theory is equally good as the best option in the Comparative Set according to the other theories and the worst option in the Comparative Set according to each theory is equally good as the worst option in the Comparative Set according to the other theories.
>
> An option is a morally conscientious choice for a person *P* in a situation *S* if and only if this option has at least as great expected normalized moral value for *P* in *S* as any other option in *S*.[14]

Basically, the idea is to first normalize the scales of moral value for each moral theory so that the difference between the best and the worst option is the same on all theories. Then, given this normalization, the Principle of Maximizing Expected Normalized Moral Value prescribes the option with the greatest expected moral value.

Let us see how this works in Case Two, using 1 for the maximum value on each theory after the normalization and 0 for the minimum value.

At choice node *a*, T_1 gives going up a normalized value of 1 and going down a normalized value of 0, while T_2 gives going down a normalized value of 1 and going up a normalized value of 0. Then the expected normalized moral value of going up is $(1/2 + \epsilon) \cdot 1 + (1/2 - \epsilon) \cdot 0 = 1/2 + \epsilon$. And the expected normalized moral value of going down is $(1/2 + \epsilon) \cdot 0 + (1/2 - \epsilon) \cdot 1 = 1/2 - \epsilon$. Hence the Principle of Maximizing Expected Normalized Moral Value prescribes going up at choice node *a*.

[14] This approach is a variation of Lockhart's (2000, p. 581) Principle of Equity among Moral Theories. Lockhart's principle also takes care of cases where all options are equally good according to some theories with positive credence but not according to some others. This complication doesn't matter for the argument of this paper.

Similarly, at choice node b, T_1 gives going down a normalized value of 1 and going up a normalized value of 0, and T_2 gives going up a normalized value of 1 and going down a normalized value of 0. Then the expected normalized moral value of going up is $(1/2 - \epsilon) \cdot 0 + (1/2 + \epsilon) \cdot 1 = 1/2 + \epsilon$. And the expected normalized moral value of going down is $(1/2 - \epsilon) \cdot 1 + (1/2 + \epsilon) \cdot 0 = 1/2 - \epsilon$. Hence the Principle of Maximizing Expected Normalized Moral Value prescribes going up at choice node b.

Thus following the Principle of Maximizing Expected Normalized Moral Value requires that one follows the Up-Up Plan in Case Two. So, like earlier approaches, the Principle of Maximizing Expected Normalized Moral Value violates the Weak Principle of Theory-Conditional Plan Dominance.

At this point, it may be objected that both the Borda Rule and the Principle of Maximizing Expected Normalized Moral Value could avoid this problem if they were revised so that they took into account more options than those that are available in a choice situation. One could revise the Borda Rule by replacing the Availability Specification with

> *The Possibility Specification* The Comparative Set in a situation is the set of all possible options in all possible situations.[15]

A problem with this revision is that there seems to be infinitely many possible options. So the revised Borda Score would be undefined for most options, being equal to infinity minus infinity. Hence this revision of the Borda Rule breaks down.

In the same way, we could revise the Principle of Maximizing Expected Normalized Moral Value so that we normalize the best and worst possible options across all moral theories with positive credence.[16] A problem with this revision is that, on some moral theories, there's no upper limit to how good or bad possible options could be. Consider, for instance, utilitarianism: For every possible option that realizes a certain sum total of happiness, there is another possible option which realizes an even greater sum total of happiness. Hence there would be no best option among all possible options. And then this revision of the Principle of Maximizing Expected Normalized Moral Value breaks down.[17]

To avoid these problems with infinity, we could let the Comparative Set be all possible options but all potential options in the decision tree

[15] MacAskill 2014, pp. 123–125.
[16] Sepielli 2010, p. 163; 2013, p. 588.
[17] Sepielli 2010, pp. 163–164; 2013, p. 588.

(rather than the available options). That is, the idea is to adopt the following specification of the Comparative Set:

> *The Resolute Specification* The Comparative Set in a situation is the set of all options that are available in any choice node that could possibly be reached from a certain privileged node.

Given that we take the privileged node to be the initial learning node in Case Two, neither the Borda Rule nor the Principle of Maximizing Expected Normalized Moral Value violates the Weak Principle of Theory-Conditional Plan Dominance, since following them then requires the Down-Down Plan.

With this revision, however, these approaches both violate

> *The Principle of Separability* Whether an option is a morally conscientious choice in a situation does not depend on options that are no longer feasible in that situation.

The revised versions of the Borda Rule and the Principle of Maximizing Expected Normalized Moral Value, with the Resolute Specification and the initial learning node as the privileged node, violate the Principle of Separability because what option is a morally conscientious choice at either choice node in Case Two depends in part on what options are available at the other choice node. It's implausible that options that can no longer be reached at a choice node should matter for what a morally conscientious person would choose at that node. Moreover, there seems to be no plausible reason why the initial learning node should have any special significance at later choice nodes.

OTHER APPROACHES

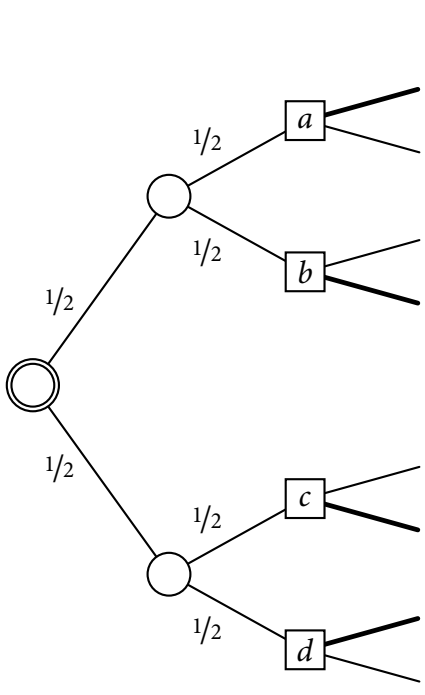
We have seen that My Favourite Theory and its rivals that don't rely on intertheoretic comparisons of value violate either the Weak Principle of Theory-Conditional Plan Dominance or the Principle of Separability. Can we do better without relying on intertheoretic comparisons of value? I doubt it. To see a more general problem, consider again Case One. If we don't rely on intertheoretic comparisons of value, all we can say about Case One is that there are two options and two moral theories, which gives opposite prescriptions, and one of the moral theories has slightly more credence than the other. Once we describe Case One this way,

it's hard to see how a morally conscientious person could do anything else in that case than to follow the slightly more credible theory—that is, to go up. But any approach that satisfies the Principle of Separability and prescribes going up in Case One must also prescribe going up at choice node *a* of Case Two. Then, by symmetry, the approach also needs to prescribe going up at choice node *b* of Case Two. But then we have that following the approach requires following the Up-Up Plan in Case Two, which violates the Weak Principle of Theory-Conditional Plan Dominance.

4. Choice-independent stakes

A strange feature of Case Two is that you will face very different stakes depending on how your moral credences will change. Although possible in principle, this may seem unrealistic. This worry, however, can be sidestepped with a variation of Case Two. This variation is more complex. Suppose, like before, that T_1 and T_2 are two maximizing moral theories. And suppose, again, that you know that you will soon learn something that will make one of T_1 and T_2 seem more credible than the other but currently you don't know which. And let ϵ be the size of this foreseen change in your credences and suppose that the shift in your credences between T_1 and T_2 will be less than $1/3$. That is, we suppose that $0 < \epsilon < 1/3$. Now, consider

Case Four



T_1 $(1/2+\epsilon)$	T_2 $(1/2-\epsilon)$
7	1
6	6
1	7
6	6

T_1 $(1/2-\epsilon)$	T_2 $(1/2+\epsilon)$
7	1
6	6
1	7
6	6

Here, the double circle represents a learning node, the circles represent standard chance nodes, and the squares represent choice nodes. Just like in Case Two, you start off with 1/2 credence in each of T_1 and T_2, but you think that, at the learning node, it is equally likely you will get information that favours T_1 (the node resolves upwards) as that you will get information that favours T_2 (the node resolves downwards). After the learning node, your credence in the theory that were favoured by the evidence rises to $1/2 + \epsilon$. After the learning node, you reach one the two standard chance nodes, which depend on the same random event that you think is equally likely to resolve upwards or downwards.

In Case Four, following My Favourite Theory requires that one follows the plan of going up at choice node *a*, down at choice node *b*, down at choice node *c*, and up at choice node *d*; let us call it *the Up-Down-Down-Up Plan*. (The prescriptions of My Favourite Theory are marked by the thicker lines.)

Likewise, following My Favourite Option also requires that one follows the Up-Down-Down-Up Plan, and so do both the Borda Rule and the Principle of Maximizing Expected Normalized Moral Value given the Availability Specification. Consider the expectation of moral value at

the initial learning node of the Up-Down-Down-Up Plan conditional on each theory. The expectation of moral value for the Up-Down-Down-Up Plan conditional on T_1, or conditional on T_2, is $(1/2 + \epsilon) \cdot (1/2 \cdot 7 + 1/2 \cdot 6) + (1/2 - \epsilon) \cdot (1/2 \cdot 6 + 1/2 \cdot 1) = 5 + 3\epsilon$. Compare this expectation of the Up-Down-Down-Up Plan with the expectation of one of the alternative plans, namely, *the Down-Down-Down-Down Plan*—that is, the plan of going down at all four choice nodes. The expectation of moral value for the Down-Down-Down-Down Plan conditional on T_1, or conditional on T_2, is $(1/2 + \epsilon) \cdot (1/2 \cdot 6 + 1/2 \cdot 6) + (1/2 - \epsilon) \cdot (1/2 \cdot 6 + 1/2 \cdot 6) = 6$. So we have that the expectations of moral value at the initial node for each of the theories must be the following:

Table Two

	T_1	T_2
The Up-Down-Down-Up Plan	$5 + 3\epsilon$	$5 + 3\epsilon$
The Down-Down-Down-Down Plan	6	6

Since ϵ is less than 1/3, we have that the Up-Down-Down-Up Plan has a worse expectation conditional on each moral theory with positive credence than the Down-Down-Down-Down Plan. Following My Favourite Theory, My Favourite Option, or either of the Borda Rule or the Principle of Maximizing Expected Normalized Moral Value given the Availability Specification requires following the Up-Down-Down-Up Plan. Hence these approaches all violate the Weak Principle of Theory-Conditional Plan Dominance. And, as we have seen in this section, we can show this without assuming that the stakes one will face in the future depends on ones moral credences.

5. The arbitrariness of the Principle of Maximizing Expected Moral Value

As we have seen, approaches that avoid such comparisons either violate the Weak Principle of Theory-Conditional Plan Dominance or some other plausible requirement. But, as mentioned earlier, the trouble with the Principle of Maximizing Expected Moral Value is its need for intertheoretic comparisons of value. If intertheoretic comparisons of value are arbitrary, then the prescriptions of the Principle of Maximizing Expected Moral Value are also arbitrary. But it's hard to see how non-arbitrary intertheoretic comparisons of value could be made. In the rest

of this section, we shall consider two proposal for how to make these comparisons.[18]

HARSANYI-BASED APPROACHES

It may be objected that the intertheoretic comparisons needed for the Principle of Maximizing Expected Moral Value can be established via a variation of John C. Harsanyi's social-aggregation theorem.[19] Let *the overall moral value* of an option be the agent's overall all moral evaluation of the option given their moral uncertainty. Then the idea is that, if the agent's judgements about overall moral value (or choice-worthiness) under moral uncertainty satisfies the axioms of Expected-Utility Theory and a compelling dominance condition, then there is a unique expected-utility representation of these value judgements.[20] The problem with this argument is that, given the moral theories you have some credence in, there are lots of potential judgements about overall moral value that would satisfy the axioms of Expected-Utility Theory and the dominance condition. Different sets of these judgements about overall moral value would have different expected-utility representations, which may prescribe different choices. Since the choice between these different sets of judgements about overall moral value seem arbitrary, the prescriptions of the Principle of Maximizing Expected Moral Value would still be arbitrary.[21]

[18] I will not cover the common-ground approach and the reactive-attitude approach. For a critical discussion of Ross's (2006, pp. 764–765) common-ground approach and Sepielli's (2010, p. 184) reactive-attitude approach, see Gustafsson and Torpman 2014, pp. 162–164 and MacAskill 2014, pp. 142–149.

[19] Riedener 2020, based on Harsanyi 1955. See also Ross 2006, p. 763.

[20] The dominance condition is, roughly, that options are equally good in terms overall moral value if the options are equal in moral value according to all moral theories with some credence, and an option is better in terms of overall moral value than another option if the first option has at least as much moral value as the second option according to all theories with some credence and the first option has more moral value according to some theory with some credence.

[21] Another problem, put forward by MacAskill (2014, p. 146), is that we would like to know what a morally conscientious person would choose under moral uncertainty. Judgements about the choice-worthiness of options under moral uncertainty is what we would like an approach to moral uncertainty to provide. But, on the Harsanyi-based approach, these judgements are the input to the theory rather than the output.

It may next be objected that there could be a universal scale for moral value. The idea is that the moral theories we have credence in assign moral value (or choice-worthiness) to options on one and the same scale. So these moral theories are all theories about the same absolute moral value quantities (or choice-worthiness quantities).[22] And, once we have a universal scale for moral value, we have a way to make non-arbitrary intertheoretic comparisons of value, since different moral theories can then use the same universal scale.

It is far from clear, however, that there is a universal intertheoretic scale for moral value.[23] Nevertheless, in following, I shall grant for the sake of the argument that there is such a scale. And I shall focus on a new objection. For this objection, we need to introduce some technical notions. Let $V_T(x)$ be the moral value of option x according to moral theory T. A theory T' is has the same *cardinal structure* as theory T'' if and only if, for all possible options x, $V_{T'}(x) = kV_{T''} + c$, where c and k are constants such that $k > 0$. And a theory T' is an *amplified* variant of theory T'' if and only if, for all possible options x, $V_{T'}(x) = kV_{T''} + c$, where c and k are constants such that $k > 1$.[24]

Suppose we grant that all moral theories in which we have some credence all grade options in terms of moral value on the same universal scale. Then intertheoretic comparisons of value are non-arbitrary. So far so good. The trouble is that we have traded the old problem of the arbitrariness of intertheoretic comparisons of value for a new problem of the arbitrariness of our distribution of credence between theories with same cardinal structure.

To see what I mean, note first that, on the universal-scale view, there could be two versions of utilitarianism, call them Utilitarianism 1 and Utilitarianism 2, such that Utilitarianism 2 is an amplified variant of Utilitarianism 1. The problem is that there seem to be no reason to have any more credence in one of these versions of utilitarianism than the other. The standard arguments for utilitarianism, such as Harsanyi's social-aggregation theorem, do not give support for utilitarianism with any specific amplitude of moral value.[25] So these arguments cannot give

[22] MacAskill (2014, pp. 149–157) calls this view 'absolutism about choice-worthiness'.
[23] MacAskill 2014, p. 154.
[24] MacAskill 2014, p. 136.
[25] Harsanyi 1955.

us any reason to adopt any specific distribution of credence between Utilitarianism 1, Utilitarianism 2, and other versions with the same cardinal structure. This seems to be a general problem: Plausible arguments for moral theories in moral philosophy do not mention any specific positive absolute cardinal amplitudes of moral value on some universal scale. And it is hard to see how there could ever be any plausible argument for favouring one version of a moral theory rather than an amplified variant. The upshot is that, if the distribution of credence between moral theories that only differ in amplitude is arbitrary, then the expectations of moral value would still be arbitrary, even if we had non-arbitrary intertheoretic comparisons of value. And, if so, the prescriptions of the Principle of Maximizing Expected Moral Value would be arbitrary.

6. Living with arbitrariness

Nevertheless, even if intertheoretical comparisons of value are arbitrary, following the Principle of Maximizing Expected Moral Value still guarantees that one satisfies the Weak Principle of Theory-Conditional Plan Dominance and the Principle of Separability. Or, more precisely, it does so as long as one relies on the same intertheoretic comparisons throughout. Hence, if one follows the Principle of Maximizing Expected Moral Value, it's guaranteed that one satisfies the Weak Principle of Theory-Conditional Plan Dominance as long as there is no change in the exchange rates between units of moral value from different moral theories. Even if the intertheoretic comparisons are arbitrary, they do impose a certain structure to our choices when they are combined with the Principle of Maximizing Expected Moral Value. And this imposed structure helps us avoid violations of the Weak Principle of Theory-Conditional Plan Dominance and the Principle of Separability. So, even if its recommendations are to some extent arbitrary, the Principle of Maximizing Expected Moral Value might still be our best approach to moral uncertainty.

I wish to thank Gustaf Arrhenius, Gunnar Björnsson, Krister Bykvist, Hilary Greaves, Anders Herlitz, William MacAskill, Erik Malmqvist, Julia Mosquera, Daniel Ramöller, and H. Orri Stefánsson for valuable comments.

References

Ahmed, Arif (2017) 'Exploiting Cyclic Preference', *Mind* 126 (504): 975–1022.

Broome, John (2012) *Climate Matters: Ethics in a Warming World*, New York: Norton.

Gracely, Edward J. (1996) 'On the Noncomparability of Judgments Made by Different Ethical Theories', *Metaphilosophy* 27 (3): 327–332.

Gustafsson, Johan E. and Wlodek Rabinowicz (2020) 'A Simpler, More Compelling Money Pump with Foresight', *The Journal of Philosophy* 117 (10): 578–589.

Gustafsson, Johan E. and Olle Torpman (2014) 'In Defence of My Favourite Theory', *Pacific Philosophical Quarterly* 95 (2): 159–174.

Harsanyi, John C. (1955) 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility', *The Journal of Political Economy* 63 (4): 309–321.

Hedden, Brian (2016) 'Does MITE Make Right? On Decision-Making under Normative Uncertainty', in Russ Shafer-Landau, ed., *Oxford Studies in Metaethics*, vol. 11, pp. 102–128, Oxford: Oxford University Press.

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

Lewis, David (1976) 'Probabilities of Conditionals and Conditional Probabilities', *The Philosophical Review* 85 (3): 297–315.

Lockhart, Ted (1992) 'Professions, Confidentiality, and Moral Uncertainty', *Professional Ethics* 1 (3–4): 33–52.

— (2000) *Moral Uncertainty and Its Consequences*, New York: Oxford University Press.

MacAskill, William (2014) *Normative Uncertainty*, Ph.D. thesis, University of Oxford, URL http://www.williammacaskill.com/s/MacAskill-Normative-Uncertainty-tnps.pdf.

— (2016) 'Normative Uncertainty as a Voting Problem', *Mind* 125 (500): 967–1004.

MacAskill, William, Krister Bykvist, and Toby Ord (2020) *Moral Uncertainty*, Oxford: Oxford University Press.

MacAskill, William and Toby Ord (2020) 'Why Maximize Expected Choice-Worthiness?', *Noûs* 54 (2): 327–353.

McClennen, Edward F. (1990) *Rationality and Dynamic Choice: Foundational Explorations*, Cambridge: Cambridge University Press.

Riedener, Stefan (2020) 'An Axiomatic Approach to Axiological Uncer-

tainty', *Philosophical Studies* 177 (2): 483–504.
Roberts, Fred S. (1979) *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*, Reading, MA: Addison-Wesley.
Ross, Jacob (2006) 'Rejecting Ethical Deflationism', *Ethics* 116 (4): 742–768.
Sepielli, Andrew (2010) *Along an Imperfectly-Lighted Path*, Ph.D. thesis, Rutgers, URL https://doi.org/doi:10.7282/T3B56JWG.
— (2013) 'Moral Uncertainty and the Principle of Equity among Moral Theories', *Philosophy and Phenomenal Research* 86 (3): 580–589.
Tarsney, Christian (2017) *Rationality and Moral Risk: A Moderate Defense of Hedging*, Ph.D. thesis, University of Maryland, URL https://doi.org/10.13016/M2K931684.