In Defence of My Favourite Theory[⋆]

Johan E. Gustafsson & Olle Torpman[†]

ABSTRACT. One of the principles on how to act under moral uncertainty, My Favourite Theory, says roughly that a morally conscientious agent chooses an option that is permitted by the most credible moral theory. In defence of this principle, we argue that it prescribes consistent choices over time, without relying on intertheoretic comparisons of value, while its main rivals are either plagued by moral analogues of money pumps or in need of a method for making non-arbitrary intertheoretic comparisons. We rebut the arguments that have been levelled against My Favourite Theory and offer some arguments against intertheoretic comparisons of value.

Many people are uncertain what they morally ought to do. This might in some cases be due to descriptive uncertainty. For instance, a convinced utilitarian might be uncertain what to do because she is uncertain of the consequences of the available acts. Yet, in other cases, people are uncertain what to do because of moral uncertainty. For example, someone might be certain of all relevant descriptive facts but still be unsure about what to do since he finds both virtue ethics and Kantianism plausible and is certain of neither.[1] This second type of uncertainty is the topic of the present paper. We will defend an answer to the question of what a morally conscientious person (who is also minimally rational) would do in cases of moral uncertainty.

To start with, an agent acts under moral uncertainty if and only if

[⋆] Published in *Pacific Philosophical Quarterly* 95 (2): 159–174, 2014. The definitive version is available at www.blackwell-synergy.com.

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

[1] There could also be a third type of uncertainty involved. One could, for instance, be uncertain how to apply Kantianism in a situation, although certain of both Kantianism and the relevant descriptive matters. Yet we think that uncertainty about application can be reduced to moral uncertainty. If we replace Kantianism by an exhaustive set of specified versions of Kantianism such that the application of each version is clear, the uncertainty of how to apply Kantianism has been replaced by moral uncertainty (where the agent's credence is divided between the specified versions).

the agent has positive credence in more than one moral theory.[2] Several principles on what a morally conscientious person would do under moral uncertainty have been discussed in the literature. One of these principles, which is sometimes disparagingly labelled My Favourite Theory (MFT), says roughly that a morally conscientious agent chooses an option that is permitted by the most credible moral theory.[3] This principle is rejected by almost every author in the field since it has several fatal implications, they claim.[4] The chief aim of this paper is nonetheless to defend this principle. p. 160
Part of this defence will consist in rebuttals of the arguments that have been levelled against MFT so far.

Our main positive argument for MFT is that it provides consistent prescriptions over time without relying on intertheoretic comparisons of value, which its main rivals fail to do: They are either plagued by moral analogues of money pumps (due to inconsistent prescriptions over time) or in need of a method for making non-arbitrary intertheoretic comparisons. As we will argue in section 1, there does not seem to be any way of making non-arbitrary intertheoretic comparisons of value, and, as we will argue in sections 2 and 3, any adequate theory of moral conscientiousness needs to prescribe consistent choices over time—i.e. consistent given that one does not change one's credence in any moral theory.

Before we begin our investigation, we may explicate in more detail what a first tentative version of My Favourite Theory says:

> *My Favourite Theory: first tentative version (MFT1)*
> An option x is a morally conscientious choice for (a person) P in (a situation) S if and only if x is permitted by the moral theory that P in S has most credence in.[5]

Note that we take 'morally conscientious' to be like 'permissible' rather than 'obligatory' in its normative strength. Hence there can be more than

[2] Moral uncertainty does not, on this definition, entail that one does not know what to do, since all moral theories one has credence in may prescribe the same option in a situation.

[3] The name 'My Favourite Theory' is due to Lockhart (2000, p. 42).

[4] The only proponent of MFT that we have found is Gracely (1996).

[5] MFT1 is very similar to the position of Gracely (1996, p. 331), who claims that

> the proper approach to uncertainty about the rightness of ethical theories is to determine the one most likely to be right, and to act in accord with its dictates.

The objection raised below against MFT1 also affects Gracely, *mutatis mutandis*.

one morally conscientious choice in a non-dilemmatic situation.[6]

1. Intertheoretic comparisons of value

For the first objection to MFT, consider the following case, where you have the credences .51 to T_1 and .49 to T_2:[7]

Different Stakes

	T_1 (p = .51)	T_2 (p = .49)
a_1	slightly nasty	saintly
a_2	merely okay	terrible

To focus on the issues specific to moral uncertainty, we assume in all examples, unless otherwise stated, that one acts under descriptive certainty. In Different Stakes, it seems intuitive that the morally conscientious person chooses a_1, the option favoured by T_2, since the difference between the moral ranks of a_1 and a_2 according to T_2 seems greater than the difference in rank between a_1 and a_2 according to T_1. The problem is that MFT requires a_2 since it is required by the most credible theory, T_1. Nevertheless, this objection to MFT depends on intertheoretic comparisons of moral value, and it is far from obvious how such comparisons can be made.

p. 161

To make things clear, there are two main views as regards the possibility of intertheoretic comparisons: (i) *comparativism*, i.e. the view that they are possible—proposed by, e.g. Ted Lockhart (2000, pp. 84–89), Jacob Ross (2006b, pp. 761–765), Andrew Sepielli (2010, pp. 172–191), and William Crouch (2010, pp. 112–121)—and (ii) *non-comparativism*, i.e. the view that they are impossible—proposed by, e.g. James Hudson (1989, p. 224) and Edward J. Gracely (1996, pp. 330–331). Among the comparativists, some adhere to *strong comparativism*, i.e. the view that it is always possible to make intertheoretic comparisons (e.g. Lockhart and Sepielli), whilst others adhere to *weak comparativism*, i.e. the view that it is in at least some cases possible to make intertheoretic comparisons (e.g. Ross

[6] If one wants to allow for morally conscientious choices even when the most credible theory neither permits nor forbids any option, one could exchange *permitted by* for *not forbidden by* in MFT1. The same change could also be made in the other versions of MFT that we will discuss in this paper.

[7] Similar examples have been used by Hudson (1989, p. 224), Lockhart (2000, p. 84), and Sepielli (2013).

and Crouch). In this section, we will defend non-comparativism by arguing against the proposals that have so far been given for how intertheoretic comparisons of value can be made. There are three such proposals in the literature, viz. the principle of equity among moral theories, the reactive-attitude approach, and the common-ground approach. We will discuss them in turn. The upshot is that the Different Stakes objection fails since it seems that non-arbitrary intertheoretic comparisons of value cannot be made.

Lockhart was one of the first to propose a principle for normalizing different rankings on different moral theories, which he labels The Principle of Equity among Moral Theories (PEMT):

> The maximum degrees of moral rightness of all possible actions in a situation according to competing moral theories should be considered equal. The minimum degrees of moral rightness of possible actions in a situation according to competing moral theories should be considered equal unless all possible actions are equally right according to one of the theories (in which case all of the actions should be considered to be maximally right according to that theory).[8]

Nevertheless, this principle suffers from several fatal drawbacks. For instance, it is unable to yield the comparisons needed in cases such as Different Stakes. According to PEMT, the saintly a_1 on T_2 and the merely okay a_2 on T_1 have the same degree of rightness since they are the maximally right options in this situation on these theories. Likewise, the terrible a_2 on T_2 has the same degree of rightness as the slightly nasty a_1 on T_1 since they are the minimally right options in the situation on these theories. Hence, according to PEMT, the stakes are not different. Moreover, Ross (2006a, p. 27, fn. 4) argues that PEMT is incompatible with the fact that two moral theories can disagree concerning which of two choice situations is more morally significant.

Besides these problems, Sepielli (2013) points to several others, which together convincingly show that PEMT is unsatisfactory as a method of intertheoretic comparisons. His most general worry is that all versions of PEMT seem arbitrary. Why normalize the theories one way—for example, by equalizing the maximum and minimum value—rather than another? No version of PEMT seems to provide the needed non-arbitrary comparisons.

p. 162

[8] Lockhart (2000, p. 84).

We will now turn to the reactive-attitude approach. Sepielli suggests that a common theory of blame intervals can function as a conceptual link between moral theories. He writes:

> The relation between normative judgment and blame is not something that it makes sense to say varies from ranking to ranking. It is a feature that depends on the role in thought of normative concepts as such. Insofar as we say that my tendency to blame someone for doing an act is conceptually tied to the degree by which I believe that act falls short of the best act available, then two "blame intervals" […] must be of the same size.[9]

Nonetheless, there is the problem that the relation between normative judgement and blame does seem to vary between moral theories. For example, some moral theories (e.g. utilitarianism) allow for blameless wrongdoing.[10] Such an allowance seems very plausible in, for instance, so called Jackson-cases, which shares the following schematic form:[11]

	s_1 (p = .5)	s_2 (p = .5)
a_1	slightly suboptimal	slightly suboptimal
a_2	optimal	terrible
a_3	terrible	optimal

In this case, a_1 is wrong on utilitarianism since there is a better option under each possible state; i.e. a_2 under s_1, and a_3 under s_2. Nevertheless, as the agent is uncertain which of the states s_1 and s_2 will obtain, some utilitarians argue that she is not to be blamed for choosing a_1 (perhaps she is to be praised for it), although a_1 is still wrong.[12] Thus they do not regard normative judgement and blame as conceptually tied. Since the relation between normative judgement and blame varies between moral theories, an approach based on blame intervals cannot plausibly serve as a general method for making non-arbitrary intertheoretic comparisons of value.

The third approach, the common-ground approach, is to make intertheoretic comparisons of value via the common ground between moral theories. Ross writes:

[9] Sepielli (2010, p. 184). See also Ross (2006a, pp. 31–34).
[10] See, e.g. Parfit (1984, p. 32) and Tännsjö (1995).
[11] See Regan (1980, pp. 264–265) and Jackson (1991, pp. 426–463).
[12] See, e.g. Graham (2010, pp. 93–94) and Bykvist (2011).

> Suppose, for example, that I am uncertain what is the correct theory of rights. My credence is divided between two such theories, T_1 and T_2. Suppose, however, that I have a background theory, T_B, that evaluates my options in relation to all considerations other than those deriving from rights. And suppose I am fully confident that this background theory is true. Thus, my credence is divided among two complete ethical theories, the first, which we may call T_{B+1}, consisting in the conjunction of T_B and T_1, and the second, which we may call T_{B+2}, consisting in the conjunction of T_B and T_2. Now suppose there is a pair of options, *i* and *j*, such that, according to both T_1 and T_2, no one's rights are at stake in the choice between *i* and *j* (*i* and *j* might, e.g., be the options of giving either of two alternative gifts). Since no rights are at issue, T_B alone will suffice to evaluate these options, and so T_{B+1} and T_{B+2} will agree concerning their values. Therefore, these alternative ethical theories will agree concerning the difference between the values of these options. We may now define "one unit of value" as the magnitude of this difference. And having thus defined a common unit of value for the two theories, it will follow that so long as we can compare the value intervals within each of these theories, there will be no difficulty comparing value intervals between the two theories.[13]

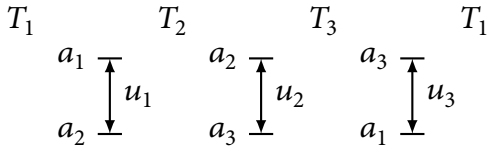
p. 163

If this approach is to work generally, however, there would always have to be a background theory common to all moral theories that is substantial enough to rank two options by itself. This seems implausible. Even in cases where all plausible theories rank options ordinally the same way (e.g. torturing or not torturing an innocent child for a small amount of pleasure), there is little reason to believe that they will agree on how much the options differ in value. Without such a common background theory, the approach will lead to inconsistent comparisons.

Moreover, one cannot rescue the common-ground approach by using different overlaps between different pairs of theories and compare them all via a chain of partial overlaps, since this may generate inconsistent comparisons. We will show this by a counter-example. Let T_1, T_2, and T_3 be three theories such that each pair of them shares a common background theory. The theoretical overlap between T_1 and T_2 is sufficient to rank option a_1 cardinally over option a_2. We define a common unit of value, u_1, between T_1 and T_2 as the difference in moral value between a_1 and a_2. Similarly, the theoretical overlap between T_2 and T_3 is sufficient

[13] Ross (2006b, pp. 764–765). See also Sepielli (2009).

to rank option a_2 cardinally over option a_3. We define a common unit of value, u_2, between T_2 and T_3 as the difference in moral value between a_2 and a_3. Finally, the theoretical overlap between T_3 and T_1 is sufficient to rank option a_3 cardinally over option a_1. We define a common unit of value, u_3, between T_1 and T_3 as the difference in moral value between a_1 and a_3.



Furthermore, suppose that the three theories rank the three options cardinally as follows:

	T_1	T_2	T_3
a_1	1	3	0
a_2	0	1	3
a_3	3	0	1

p. 164

Given these rankings, we can infer some relationships between the different units according to each moral theory:

$$T_1\colon u_3 = 2u_1$$
$$T_2\colon u_1 = 2u_2$$
$$T_3\colon u_2 = 2u_3$$

We hence have that $u_1 = u_2 = u_3 = 0$. But since, for example, T_1 ranks a_1 strictly higher than a_2, we have that $u_1 > 0$. Therefore, the common-ground approach might yield inconsistent prescriptions given partial credence in three internally consistent theories.

To make the example more concrete, suppose that a_1 is an act of lying, a_2 is an act of stealing, and a_3 is an act of adultery. T_1 and T_2 share a background theory that evaluates options in relation to all considerations other than those deriving from adultery, T_2 and T_3 share a background theory that evaluates options in relation to all considerations other than those deriving from lying, and T_1 and T_3 share a background theory that evaluates options in relation to all considerations other than those deriving from stealing. The theory shared by T_1 and T_2 yields that stealing is morally worse than lying, the theory shared by T_2 and T_3 yields that adultery is morally worse than stealing, and the theory shared by T_1 and T_3

yields that lying is morally worse than adultery. These theories seem internally consistent, and the theoretical overlaps seem as plausible as those in Ross's example.

More generally, this counter-example shows that the following three claims cannot all be true:

(1) If intertheoretic comparisons of value can be made between moral theories T and T', and a background theory common to both T and T' is sufficient to rank two options x and y cardinally, then the difference in value between x and y is the same on T and T'.

(2) A morally conscientious person may have positive credence in any combination of internally consistent moral theories.

(3) Intertheoretic comparisons of value are possible between all moral theories that yield cardinal rankings of value.

Of these claims, (3) is the least plausible. But even if one rejects (3), one might claim that intertheoretical comparisons are possible between at least some moral theories. Still, theories that are so similar that they even substantially overlap each other seem to be the best candidates for intertheoretic comparisons. And since it is such overlapping theories that give rise to the problems above, intertheoretical comparisons of value seem impossible between those theories where such comparisons are most p. 165 plausible. It is thus hard to see how any intertheoretic comparison of value whatsoever could be made. Hence the above problem seems to undermine not only strong comparativism but also weak comparativism.

Closing this section, it seems that none of the discussed proposals succeeds in making intertheoretic comparisons of value plausible. Therefore the Different Stakes objection, which depends on them, loses its punch against MFT.

2. My Favourite Option

Even if intertheoretic comparisons of value are granted impossible, another standard objection to MFT remains. This objection builds on a type of case where MFT requires the option that is most likely to be wrong. Consider:[14]

[14] This type of case has been used by Lockhart (2000, pp. 43–44), Ord and Bostrom (n.d., p. 5), and Crouch (2010, p. 26).

	T_1 ($p = .2$)	T_2 ($p = .1$)	T_3 ($p = .1$)	T_4 ($p = .1$)	T_5 ($p = .1$)	T_6 ($p = .1$)	T_7 ($p = .1$)	T_8 ($p = .1$)	T_9 ($p = .1$)
a_1	right	wrong	wrong	wrong	wrong	wrong	wrong	wrong	wrong
a_2	wrong	right	right	right	right	right	right	right	right

In this example, your credence in a_2 being right sums up to .8, while your credence in a_1 being right is merely .2. MFT requires a_1 since a_1 is required by the most credible theory, T_1. A common reaction to this case is nevertheless that only a_2, the option most likely to be right, is a morally conscientious choice. If one thinks so, one seems to abide by a principle like the following:

> *My Favourite Option (MFO)*
> An option x is a morally conscientious choice for P in S if and only if P in S has at least as high credence in x being right as in every other option.[15]

One might find this principle intuitively compelling. Nevertheless, a problem with this principle is that it can generate cycles. Consider a case structured like Condorcet's paradox:

	T_1 ($p = 1/3$)	T_2 ($p = 1/3$)	T_3 ($p = 1/3$)
a_1	2	0	1
a_2	0	1	2
a_3	1	2	0

Here, the number for each outcome represents the ranking of the outcome with regard to moral value. Furthermore, T_1, T_2, and T_3 requires maximizing moral value. In this case, MFO will lead to cyclic pairwise choices, which in turn leads to inconsistent choices over time. To see this, consider the following example. You first face a choice between a_1 and a_2. In this first situation, a_2 is required by both T_2 and T_3. Therefore, you choose a_2 as required by MFO. You are then faced with the opportunity to revoke your decision upon a_2 in favour of a_3. In this second situation, a_3 is required by both T_1 and T_2. Therefore, you choose a_3 as required by MFO. Finally, you are faced with the opportunity to revoke your decision upon a_3 in favour of $a_1 - \epsilon$, where $a_1 - \epsilon$ is a_1 with a small moral sacrifice such that $a_1 - \epsilon$ is worse than a_1 on each of T_1, T_2, and T_3. This sacrifice is so small, however, that T_1 and T_3 will still require $a_1 - \epsilon$, since

p. 166

[15] Lockhart (2000, p. 26) discusses a similar principle.

a_1 beats a_3 with some margin on these theories. So you decide, guided by MFO, upon $a_1 - \epsilon$. Nonetheless, MFO has now led you through a series of steps to a certain moral loss: you chose $a_1 - \epsilon$ when you could have chosen a_1, which would have been morally better according to all moral theories in which you have some credence. If you had rejected the opportunity to revoke your choices and avoided the certain moral loss, you would not have been morally conscientious according to MFO, which seems counter-intuitive.

Another related problem is that MFO violates a version of the principle of independence of irrelevant alternatives:

> *The Independence of Irrelevant Alternatives*
> If x is a morally conscientious choice from the set of options U and x belongs to the set of options V contained in U, then x is also a morally conscientious choice from V (given that the credences for all moral theories are fixed relative to U and V and, on all moral theories with a positive credence, the moral value of the options in V given a choice from V is the same as their moral value given a choice from U).[16]

In other words, if an option is a good enough choice from one set of options, it should still be a good enough choice even after one has removed some of its rival options from the set. So long as the moral value of the options does not change when the set is contracted, this seems like a plausible principle. To see that MFO violates The Independence of Irrelevant Alternatives, note that a_1 in the above example is a morally conscientious choice from the set of options $\{a_1, a_2, a_3\}$ but a_1 is not a morally conscientious choice from the set of options $\{a_1, a_2\}$.

One might object that the principle of independence of irrelevant alternatives has been subject to a number of counter-examples. In a typical example, due to Amartya Sen, an agent is offered a choice at a dinner between the last remaining apple and having nothing. Since she does not want to violate good behaviour, she does not take the one apple. Even so, she would have taken the apple if there had also been a further apple on

[16] This condition is analogous to a principle for decision under descriptive uncertainty proposed by Roy Radner and Jacob Marschak (1954, p. 63), which in turn is based on a condition for solution points in bargaining situations by John Nash (1950, p. 159). Radner and Marschak's principle should not be confused with the principle of the same name that was employed by Kenneth J. Arrow in his impossibility theorem. See Arrow (1951, p. 27). The two versions are logically independent, see Ray (1973).

offer.[17] A standard reply to this kind of example is that *having the last apple and leave nothing for the other guests* is less preferable or worse than *having the next to last apple and leave one for the other guests* and hence a different option. Thus, in this type of case, the agent chooses differently between the options in the smaller set when they are supplemented by the third option, since they are then valued differently. MFO's violation of The Independence of Irrelevant Alternatives above, however, is not of this type, since neither a_1 nor a_2 is valued differently in $\{a_1, a_2, a_3\}$ than in $\{a_1, a_2\}$ by T_1, T_2, or T_3. Hence there seems to be no reason to rank a_1 and a_2 in $\{a_1, a_2, a_3\}$ differently than in $\{a_1, a_2\}$.

p. 167

3. Tie breaking and inconsistency

Although MFT has survived the threats so far, there are some more formal objections left, which call for some minor revisions of the view. One such objection is due to Ross. He objects that MFT1 runs into trouble when more than one theory has the highest credence:

> [I]n the absence of intertheoretic value comparisons, one might simply follow the theory one finds most plausible, so that if one regarded the traditional morality as ever so slightly more plausible than Singer's theory, one would order the veal cutlet, and if one regarded the two theories as equally plausible, one would flip a coin. But […] this hardly seems like a rational solution.[18]

Ross's objection, however, is based on the assumption that one must break ties. A more straightforward solution is to grant all choices that are permitted by at least one of the most credible moral theories as morally conscientious. While the first tentative version of MFT did not cover the possibility of ties, the following modified version does:

> *My Favourite Theory: second tentative version (MFT2)*
> An option x is a morally conscientious choice for P in S if and only if x is permitted by one of the moral theories that P in S has most credence in.

Nevertheless, MFT2 does not seem to solve the tie-breaking problem in a satisfactory way. If two equally credible theories give starkly different

[17] Sen (1993, p. 501).
[18] Ross (2006b, p. 762, fn. 11).

prescriptions (*Start a mink farm and make furs!* vs. *Set the minks free!*), then MFT2 yields that both options are morally conscientious choices and that one may start a mink farm on Monday in order to provide winter clothing, whereas on Tuesday one sets all the minks free following some animal-liberation theory, which seems counter-intuitive. There is something unsatisfactory with such an agent. A more conscientious approach would be to commit to one of the theories and be consistent over time.[19]

There is also another problem with MFT2 regarding consistency over time, which is similar to the one we raised above against MFO. Namely, it allows agents to, in a series of steps, choose as to achieve a certain moral loss. Consider again this table of equally credible theories T_1, T_2, and T_3 p. 168 that all require maximizing moral value, where the number for each outcome represents the ranking of the outcome with regard to moral value:

	T_1 ($p = 1/3$)	T_2 ($p = 1/3$)	T_3 ($p = 1/3$)
a_1	2	0	1
a_2	0	1	2
a_3	1	2	0

Suppose that you first face a choice between a_1 and a_2. In this first situation, a_2 is permitted by T_2 and T_3. You may therefore choose a_2 in accordance with MFT2. You are then faced with the opportunity to revoke your decision upon a_2 in favour of a_3. In this second situation, a_3 is permitted by T_1 and T_2. Hence you may choose a_3 in accordance with MFT2. Suppose, finally, that you are faced with the opportunity to revoke your decision upon a_3 in favour of $a_1 - \epsilon$, where $a_1 - \epsilon$ is a_1 with a small moral sacrifice such that $a_1 - \epsilon$ is worse than a_1 on each of T_1, T_2, and T_3. This sacrifice is so small, however, that T_1 and T_3 will still permit $a_1 - \epsilon$, since a_1 beats a_3 with some margin on T_1 and T_3. So you decide, in accordance with MFT2, upon $a_1 - \epsilon$. Nevertheless, MFT2 has now allowed you to choose, through a series of steps, a certain moral loss: you chose $a_1 - \epsilon$ when you could have chosen a_1, which would have been morally better according to all moral theories in which you have some credence. This inconsistent series of choices should not be granted as morally conscientious.

In order to avoid these problems regarding consistency over time, we revise MFT2 accordingly:

[19] We thank Gustaf Arrhenius for this point.

My Favourite Theory: third tentative version (MFT3)
An option x is a morally conscientious choice for P in S if and only if x is permitted by a moral theory T such that

(a) T is in the set U of moral theories that are at least as credible as every moral theory for P in S and

(b) P in S has not violated T more recently than any other moral theory in U.

MFT3 avoids the mink problem, since it does not allow you to follow the animal-liberation theory on Tuesday if you violated it on Monday in order to act in accordance with the fur-making theory. Furthermore, MFT3 avoids granting the choice of a certain moral loss as morally conscientious in the above example, since it does not allow the final step where one revokes the decision upon a_3 in favour of $a_1 - \epsilon$. This revocation is not allowed, since you have violated T_1 and T_3 more recently than T_2, and therefore you must follow T_2, which requires a_3. p. 169

MFT3 does not just avoid the moral money-pump problem in the above case, it circumvents them generally, at least given that all moral theories in which the agent has most credence yield transitive rankings. To see this, note that as long as you follow MFT3 there will always be one moral theory, among those in which you have the highest credence, that you never violate. Moreover, the problem with these moral versions of money pumps is that they lead to a certain moral loss. But then there cannot be one moral theory you have positive credence in that permits you to go along with every step in the pump, again given that the moral theories yield transitive rankings. So, if you go along with every step in the pump, you must violate every moral theory in which you have positive credence. In that case, however, you do not follow MFT3, since it requires that there is at least one moral theory in which you have positive credence that you never violate. So, as long as you follow MFT3, you will avoid a certain moral loss.

Still, one may wonder why one should not just pick one of the most credible theories and follow it consistently over time. MFT3 does allow for such a strategy but does not require it. Instead, MFT3 requires only that one does not act on a theory that is inconsistent with the choices one has made so far. And there seems to be no reason to demand any particular strategy for meeting that requirement.

4. Dominance

There is yet another objection that threatens MFT3: It violates a plausible version of the dominance principle.[20]

> *Dominance*
> An option x is not a morally conscientious choice if there is an option y such that there is at least one positively credible moral theory that permits y but not x, and no positively credible moral theory permits x but not y.

To see this, consider the following example where you have credence in two theories; on the most credible one both a_1 and a_2 are right, while on the other only a_1 is right.

	T_1 ($p = .6$)	T_2 ($p = .4$)
a_1	right	right
a_2	right	wrong

Since T_2 requires a_1 and neither T_1 nor T_2 requires a_2, Dominance yields that a_2 is not a morally conscientious choice. But, according to MFT3, both a_1 and a_2 are morally conscientious choices since they are both permitted by the most credible theory.

p. 170

Furthermore, it seems that in cases where the most credible theory permits more than one option, the second most credible theory should be taken into account even though no option is dominated. Consider, for instance, the following case:

	T_1 ($p = .5$)	T_2 ($p = .4$)	T_3 ($p = .1$)
a_1	right	right	wrong
a_2	right	wrong	right

Here, both a_1 and a_2 are morally conscientious, according to MFT3, but a_1 is permitted by the two most credible theories whereas a_2 is not. In order to make MFT better handle cases of this type and comply with Dominance, we revise MFT3 by adding a second condition accordingly:

[20] Crouch (2010, pp. 27–29).

My Favourite Theory: fourth and final version (MFT4)
An option x is a morally conscientious choice for P in S if and only if

1. x is permitted by a moral theory T such that
   (a) T is in the set U of moral theories that are at least as credible as every moral theory for P in S and
   (b) P in S has not violated T more recently than any other moral theory in U, and
2. there is no option y and no moral theory T' such that
   (a) T' permits y and T' does not permit x and
   (b) there is no moral theory T'' such that T'' is at least as credible as T' for P in S and T'' permits x and T'' does not permit y.

By this revision, we have made MFT lexical by taking into account not only the most credible moral theories, but also the second most credible (and so on) theories, in cases where the more credible theories yield ties. In the last example, a_2 is not morally conscientious according to MFT4, since there is the option a_1 and the theory T_2 that permits a_1 but does not permit a_2, and none of the moral theories that are at least as credible as T_2 (that is, T_1 and T_2) permits a_2 but not a_1.

5. Individuation of theories

A final objection is that MFT's prescriptions may depend on how moral theories are individuated. The upshot of the objection is that if the individuation of moral theories is arbitrary, so are MFT's prescriptions.[21] For instance, consider the following case:

p. 171

	Deontology	Consequentialism version 1	Consequentialism version 2
	($p = .4$)	($p = .3$)	($p = .3$)
a_1	right	wrong	wrong
a_2	wrong	right	right

[21] Ord and Bostrom (n.d., p. 4).

If the two versions of consequentialism are distinct moral theories, MFT will follow the prescriptions of deontology; that is, it will require a_1. But if the two versions of consequentialism are versions of a single moral theory with a .6 credence, consequentialism is the most credible theory and MFT requires a_2 instead:

	Deontology ($p = .4$)	Consequentialism ($p = .6$)
a_1	right	wrong
a_2	wrong	right

To solve this problem, it suffices to find a principle for how to individuate moral theories when one applies MFT. This is the approach we adopt. The principle we propose, however, is only supposed to be an interpretation rule for 'moral theory' in the formulation of MFT; it should not be taken as a general account for individuation of moral theories.[22] We suggest the following:

> 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.

The rationale behind this principle is to individuate moral theories so that MFT yields non-arbitrary and consistent recommendations over time. If T_1 is regarded as the same theory as T_2 and this is the theory in which one has the highest credence, it is arbitrary which of T_1 and T_2 one follows. Thus in order to avoid arbitrary recommendations, T_1 and T_2 must yield the same recommendations in every situation, otherwise the recommendations would depend on the arbitrary choice between T_1 and T_2.

A less fine-grained individuation principle could yield that two theories, which you think might yield different prescriptions in some situation, should be regarded as the same theory. In such situations, this would result in arbitrary prescriptions depending on which version of the theory is adopted. On the other hand, an even more fine-grained individuation principle would just be overkill.

One might object, however, that this individuation principle yields implausible results when combined with MFT. Suppose you are trying to decide whether to lie. You have .99 credence that Kantianism is true, and .01 credence that utilitarianism is true. There is only one version of utili- p. 172

[22] Cf. Bergström (1966, pp. 12–14).

tarianism in which you have any credence, which implies that you should lie. By contrast, you have slightly less than .01 credence in 100 versions of Kantianism, which all agree that lying is wrong. Still, the versions of Kantianism disagree about other issues, such as the rights of animals and the unborn, sexual morality, the morality of waging war, and so on. For all you know, you might one day be in a situation in which any one of the differences between these theories might be relevant. But, on the supposition that Kantianism is true, you are absolutely certain that lying is wrong. Yet, since all the versions of Kantianism will count as distinct theories on our individuation principle and you regard each of those theories as less plausible than utilitarianism, MFT requires that you lie. This might seem implausible.

This objection, however, seems to require a non-arbitrary way of individuating moral theories. If there is no non-arbitrary individuation principle, we have to drop the claim that the 100 versions of Kantianism are in a non-arbitrary way versions of a single moral theory. Without this claim, it seems that they are just theories that happen to give the same recommendation in this case. And if we take not lying to be the only morally conscientious choice because it is the option that is most likely to be right, then we seem to rely on MFO, which we rejected in section 2. If, on the other hand, there is a non-arbitrary principle for individuating moral theories, this principle could yield that the 100 versions of Kantianism are indeed in a non-arbitrary way versions of one single moral theory. This would conflict, however, with the main premise of the individuation objection to MFT, that is, that the individuation of moral theories is arbitrary. If there is a non-arbitrary way of individuating moral theories, MFT could be applied with theories individuated in that way. So, in that case, the individuation objection would not get off the ground.

In conclusion, My Favourite Theory seems to have an advantage over its main rivals since it yields consistent prescriptions over time—and hence avoids moral analogues to money pumps—without relying on problematic intertheoretic comparisons. Moreover, the objections that have been levelled against My Favourite Theory seem to be far less threatening than has been suggested so far.

Thanks to Gustaf Arrhenius, William MacAskill, Nicolas Espinoza, Marc Fleurbaey, Sven Ove Hansson, Jonas Olson, Toby Ord, Michael Otsuka, Torbjörn Tännsjö, Martin Peterson, Alex Voorhoeve, Nicolas Olsson-Yaouzis, and an anony-

mous referee for valuable comments. Thanks also to the audiences at the *Workshop on Moral Uncertainty*, The Stockholm Centre for Healthcare Ethics (CHE), April 6, 2011 and at the *Economics and Philosophy Workshop*, Princeton, November 26, 2012. Financial support for Johan E. Gustafsson from the Franco-Swedish Program in Economics and Philosophy, Fondation Maison des sciences de l'homme, and Riksbankens Jubileumsfond is gratefully acknowledged.

References

Arrow, Kenneth J. (1951) *Social Choice and Individual Values*, New York: Wiley.

Bergström, Lars (1966) *The Alternatives and Consequences of Actions: An Essay on Certain Fundamental Notions in Teleological Ethics*, Stockholm: Almqvist & Wiksell.

Bykvist, Krister (2011) 'How to Do Wrong Knowingly and Get Away with It', in Rysiek Sliwinski and Frans Svensson, eds., *Neither/Nor: Philosophical Papers Dedicated to Erik Carlson on the Occasion of His Fiftieth Birthday*, no. 58 in Uppsala Philosophical Studies, pp. 31–47, Uppsala: Uppsala University.

Crouch, William (2010) 'Moral Uncertainty and Intertheoretic Comparisons of Value', BPhil thesis, University of Oxford.

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

Graham, Peter A. (2010) 'In Defence of Objectivism about Moral Obligation', *Ethics* 121 (1): 88–115.

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

Jackson, Frank (1991) 'Decision-Theoretic Consequentialism and the Nearest and Dearest Objection', *Ethics* 101 (3): 461–482.

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

Nash, John F. Jr. (1950) 'The Bargaining Problem', *Econometrica* 18 (2): 155–162.

Ord, Toby and Nick Bostrom (n.d.) 'Decisions under Moral Uncertainty', unpublished.

Parfit, Derek (1984) *Reasons and Persons*, Oxford: Clarendon Press.

Radner, Roy and Jacob Marschak (1954) 'Note on Some Proposed Decision Criteria', in R. M. Thrall, C. H. Coombs, and R. L. Davis, eds., *Decision Processes*, pp. 61–68, New York: Wiley.

Ray, Paramesh (1973) 'Independence of Irrelevant Alternatives', *Econometrica* 41 (5): 987–991.
Regan, Donald (1980) *Utilitarianism and Co-Operation*, Oxford: Clarendon Press.
Ross, Jacob (2006a) *Acceptance and Practical Reason*, Ph.D. thesis, Rutgers.
— (2006b) 'Rejecting Ethical Deflationism', *Ethics* 116 (4): 742–768.
Sen, Amartya (1993) 'Internal Consistency of Choice', *Econometrica* 61 (3): 495–521.
Sepielli, Andrew (2009) 'What to Do When You Don't Know What to Do', in Russ Shafer-Landau, ed., *Oxford Studies in Metaethics: Volume Four*, pp. 5–28, Oxford: Oxford University Press.
— (2010) *Along an Imperfectly-Lighted Path*, Ph.D. thesis, Rutgers.
— (2013) 'Moral Uncertainty and the Principle of Equity among Moral Theories', *Philosophy and Phenomenal Research* 86 (3): 580–589.
Tännsjö, Torbjörn (1995) 'Blameless Wrongdoing', *Ethics* 106 (1): 120–127.