IS VALUE UNDER HYPOTHESIS VALUE?

ITTAY NISSAN-ROZEN

Hebrew University of Jerusalem

In the context of two recent yet distinct philosophical debates—over choice under conditions of moral uncertainty and over transformative choices—several philosophers have implicitly adopted a thesis about how to evaluate alternatives of uncertain value. The thesis says that the value a rational agent ought to attach to an alternative under the hypothesis that the value of this alternative is x , ought to be x . I argue that while in some contexts this thesis trivially holds, in the context of the two debates in which the thesis has been adopted, it does not. I also discuss several implications of this failure.

THE claim that rational agents should maximize expected value in their choices has been misunderstood in some cases. While it is true that a rational agent should choose in a way such that, given her degrees of belief, there will be some value function with respect to which she maximizes expected value, for any specific value function, $f(\cdot)$ it is false that rationality requires that she maximizes expected value with respect to $f(\cdot)$.

However, in some cases other normative considerations (i.e., other than the standard demands of practical rationality) may lead a rational agent to maximize the expectation of a specific value function.

All this is relatively uncontroversial. However, lately, in the context of two different philosophical debates, a more controversial claim has been advanced. This claim is about choice situations in which a rational agent is certain that she should (on the basis of some set of normative considerations) maximize the expectation of a specific value function (let us call this function “the correct value function”) but is uncertain which one it is.

We can represent such choice situations in the following way. Let $A = \{a_1 \dots a_n\}$ be the set of acts available to the agent in the choice situation she faces. Let $H = \{h_1 \dots h_n\}$ be a set of all the hypotheses regarding which value function is the

Contact: Ittay Nissan-Rozen <ittay.nissan@mail.huji.ac.il>

correct value function, to which the agent attaches a positive credence. Each h_i is itself a value function that attaches a certain value to each act in A (so each h_i is a function from A to the set of real numbers). Let $c(\cdot)$ be a probability function which is defined over H and represents the agent's degrees of belief regarding which h_i is the correct value function.

Now, let $v(\cdot)$ be the function that attaches to every ordered pair of the form (a_j, h_i) , the value h_i attaches to a_j . In other words, $v(\cdot)$ is the function that attaches to each one of the possible acts available to the agent, in every state in which a given hypothesis about the correct value function holds, the value that the relevant hypothesis attaches to the act.

The controversial claim is that in such cases the agent must maximize the expectation of $v(\cdot)$ relative to $c(\cdot)$.¹

Let us call this claim, the "Value Under Hypotheses is Value" thesis (VUV thesis). In this paper I argue that the truth of the VUV thesis depends on the type of normative considerations (determining which value function is "the correct one") involved. While in some normative contexts the VUV thesis trivially holds, in others it does not. I also argue that in the context of the two philosophical debates mentioned above, which are the two philosophical debates in which the VUV thesis has been (implicitly) suggested, it does not hold.

The first debate in which the VUV thesis has been put forward concerns moral decision making in face of moral uncertainty. The second debate concerns decision making in face of expected experience that is both personally and epistemically transformative.

In both cases it has been argued against the use of the VUV thesis that it must involve acting from the wrong kind of motivation. In the first debate such an argument has been advanced by Brian Weatherson (2014). Weatherson argues that following a decision rule for choice under conditions of moral uncertainty that is sensitive to the agent's uncertainty regarding which value function is the correct one must involve acting from a *de dicto* moral motivation.

In the second debate such an argument has been advanced by Laurie Paul

1. This representation of the VUV thesis fits decision theoretic frameworks in which the agent's credence function, the agent's utility function, and the agent's preferences are defined over different sets of mathematical objects (states, outcomes and acts, correspondingly), such as Savage's (1972) one. However, the thesis can also be formulated in decision theoretic frameworks in which this is not the case. For example, a natural way to represent the thesis in Richard Jeffrey's decision theoretic framework is the following one. Let $d(\cdot)$ be a rational agent's desirability function, let A be a proposition, and let A_i^* be the proposition "the value of A is x_i ". The VUV thesis can be represented in the following way: for every A_i^* , $d(AA_i^*) = x_i$. It then follows from Jeffrey's desirability axiom (applied to the partition A_i^*) that $d(A) = \sum x_i c(A_i^* | A)$. In his discussion of the "Desire as Beliefs Thesis", David Lewis seems to commit himself to this formulation of the thesis (see Lewis 1988: 332) and so do others that responded to Lewis (see, e.g., Price 1989; Broome 1991). See Nissan-Rozen (2015a) and Bradley and Stefansson (2016) for deeper discussions of this formulation. I thank an anonymous referee for pointing this out to me.

(2015). Paul argued that Richard Pettigrew's (2015) suggestion regarding how to choose in the face of expected transformative experience, a suggestion implicitly committed to the VUV thesis, must lead to choices that are unauthentic in a relatively well-defined sense.

I find Weatherson's and Paul's arguments compelling. However, I also think that there are good objections to their arguments.² In this paper I aim to strengthen their arguments by pointing to a type of case in which the conflict between the VUV thesis and the demand to act out of "the right" kind of motivation is particularly salient. Specifically, I show that in both cases, respecting the VUV thesis must sometimes involve either a violation of the standard rationality principle of Dominance or evaluating outcomes in a way that is *completely insensitive* to any of their first-order features.

My argument achieves this in a way that is invulnerable to the objections to Paul's and Weatherson's original arguments (discussed in Section 2), by concentrating on the possibility of buying information that the agent finds valuable. Introducing this possibility makes it possible for me to construct cases in which the conflict between the VUV thesis and the demand to act out of "the right kind" of motivation, the conflict to which both Paul and Weatherson have pointed, arises in its most extreme form.

The general lesson I wish to draw is that we should be extremely careful when moving from the relatively weak demand to maximize expected value in one's choices to the much stronger demand to maximize the expectation of a specific value function. Examining the value one must attach to information in case one is committed to the latter demand helps to see what this demand really amounts to.

The rest of this paper is organized in the following way. In Section 1 I go over some necessary background: I present the two suggested applications of the VUV-thesis and distinguish clearly between the VUV-thesis and the demand to maximize expected value. In Section 2 I briefly present Weatherson's and Paul's arguments and discuss their limitations. In Section 3 I present my argument and show how it manages to overcome the limitations of Paul's and Weatherson's arguments.

The VUV Thesis and Maximization of Expected Value

Orthodox normative decision theory says that rational agents that have to make a choice under conditions of uncertainty should choose in a way *consistent with*

2. I discuss them in Section 2. For other discussions see Sepielli (2016) and Aboodi (2017), with respect to Weatherson's argument, and Bykvist and Stefansson (2017) and Pettigrew (2015) with respect to Paul's.

the following procedure. First, they should assign a certain numerical value to each of the possible outcomes of their choice. These numbers are called utilities and they represent how much the decision makers value each outcome. Second, they should assign a certain numerical value, x_i , to each one of the possible states they think the world might be in, such that $\sum x_i = 1$. These numbers represent the decision makers' degrees of belief that each state is in fact the true state. Third, using these values, they should calculate the expected utility of each one of the acts available to them and choose the one with the highest expected utility.³

For example, consider the following case:

Medical Procedure (MP)

You suffer from a medical condition that causes you mild physical discomfort. You can either do nothing about it or you can go through a certain medical procedure that—if successful—will cure you. However, if the procedure fails you will suffer much greater physical discomfort. You are uncertain whether the procedure will succeed.

Orthodox normative decision theory demands that whatever the mental process you go through when deciding whether to take the procedure actually is, it must be consistent with the following. First you decide how much you value each one of the three possible outcomes involved in the situation (i.e., “suffering mild discomfort”, “suffering severe discomfort”, and “suffering no discomfort”). You do this by assigning a utility value to each outcome. Second, you assign a probability value to each of the two possible states of the world involved (i.e., “in cases I go through the procedure, it succeeds” and “in case I go through the procedure, it fails”). Third, you calculate the expected utility of each of the two acts available to you (i.e., “go through the procedure” and “do nothing”) and then choose the one with the higher expected utility.

So if—for example—you assign utility 1 to your condition without going through the procedure, utility 2 to your position after going through a successful procedure, utility 0 to your position after going through an unsuccessful procedure, and believe to degree 0.4 that the procedure will succeed in case it is taken, you should not go through the procedure because, while the expected utility of

3. When I say that orthodox normative decision theory demands that rational agents choose in a way “which is consistent” with the procedure described in the main text, this is what I mean: there must be a pair of a utility function defined over the set of possible outcomes involved in the situation, and a probability function, defined over the set of possible states of the world, such that, with respect to this pair of functions, it is true for any two acts involved in the situation that if the agent is disposed to choose one of them over the other in case these two acts are the only acts available to her, the expected utility of the first is greater than that of the second.

not going through the procedure is 1, the expected utility of going through the procedure is $0.4 \cdot 2 + 0.6 \cdot 0 = 0.8 < 1$.

Now, orthodox decision theory does not demand that you use this procedure, only that you choose in a way consistent with it. But it does allow you to use this procedure. Following Pettigrew (2015), we can call usages of this procedure, instances of application of “the deliberative conception of decision theory”.

Even in case you do follow this procedure, orthodox normative decision theory says nothing about what should be the utility values you attach to the different outcomes. In some cases, though, other considerations may push you to adopt certain utility functions. For example, suppose you are an investor in the stock market and the only thing you care about is how much money you will end up earning by the end of the year. In such a case you may be disposed to adopt a utility function that assigns to each outcome in each decision you face its monetary value.

In some cases, some *normative* considerations may push you to adopt certain utility functions. For example, suppose you act as the guardian of another person and have to make choices for him. It seems plausible that, at least in some cases, you should (either morally or legally but—in any case—in terms of some normative system) adopt that person’s utility function.

What happens if you find yourself in a situation in which you are certain there is a utility function that it is normatively required of you to adopt (call it the “correct” utility function), but you are uncertain which one it is? For example, what happens in the following case?

Medical Procedure (MP*)*

You have to choose for another person whether to go through a given medical procedure (maybe he is unconscious). There is no uncertainty regarding the outcome of the procedure involved. However, you are uncertain about how this person evaluates (or would evaluate) the two possible outcomes involved in the decision (i.e., the outcome of going through the operation and the outcome of doing nothing). Suppose you are certain that—when making the decision—you have to use this person’s utility function (i.e., the values the expectation of which that person would have maximized if he was the one making the decision). However, while you are certain that this person would attach a utility value of 1 to the (certain) outcome of not going through the procedure, you are uncertain which utility value he would attach to the outcome of going through the procedure. You believe to degree 0.4 that he would attach a value of 2 to this outcome and believe to degree 0.6 that he would attach a value of 0 to this outcome. How should you choose?

The immediate answer that comes to mind is the following. You should avoid sending this person through the procedure because of the following consideration. In MP* you have two available acts: Do nothing (D) or Go through the procedure (G). You assign a positive credence to two hypothesis about the relevant utility function: h_1 ("the value of D is 1 and the value of G is 2") and h_2 ("the value of D is 1 and the value of G is 0"). Let $c(\cdot)$ be your credence function over the set $\{h_1, h_2\}$. By stipulation $c(h_1) = 0.4$ and $c(h_2) = 0.6$.

Now, let $v(\cdot)$ be the function that assigns to (G, h_2) the value 0, to (G, h_1) the value 2, and to (D, h_1) and (D, h_2) the value 1. In other words, $v(\cdot)$ assigns to each act in each state in which the utility of the act is x , the value x . It is easy to see that maximizing the expectation of $v(\cdot)$ relative to $c(\cdot)$ leads to choosing D. While the expected utility of G is $0.4 \cdot 2 + 0.6 \cdot 0 = 0.8$, the expected utility of D is 1.

It seems to me that it is the structural similarity between *medical procedure* and *medical procedure** that is responsible for the widely held judgement that in *medical procedure** it is normatively required to choose in the way specified above. However, notice that if it is indeed normatively required to choose in such a way, it is not so in virtue of the standard norms of practical rationality. The standard norms of practical rationality say nothing about which utility function one must adopt. They only require that one chooses in a way consistent with maximizing the expectation of some utility function with respect to some probability function. So if it is the case that you must choose in the way specified above in *medical procedure**, it is in virtue of whatever normative system it is that makes it normatively required of you to adopt the other person's utility function when making choices for him,⁴ not in virtue of the standard decision theoretic norms of practical rationality.

It might be helpful to represent both *medical procedure* and *medical procedure** using a single simple table.

In the case of MP, we have *stipulated* that you attach a utility value of 1 to "mild pain", a utility value of 2 to "no pain" and a utility value of 0 to "severe pain". These numbers were not dictated to us by any decision theoretic norm. Similarly, if it is true that in MP* you are normatively required to attach a utility value of 1 to "other person's utility is 1", a utility value of 2 to "other person's utility value is 2" and a utility value of 0 to "other person's utility value is 0", it is none of the standard decision theoretic norms that requires you to do so. Rather it is some other norm (maybe a moral norm).

Let us call this type of norm *Numerical-Value norms* (NV norms). An NV norm

4. There is a further issue here regarding whether or not you are required to adopt a certain attitude to risk when making your choice in *medical procedure**. However, the discussion here is independent of the question of attitude to risk and so, for the sake of simplicity, I will assume, throughout, neutrality to risk. Nothing in the paper depends on this assumption. For a discussion of this issue see Nissan-Rozen (2015b).

Table 1

	State 1 (probability 0.4)	State 2 (probability 0.6)
	<ul style="list-style-type: none"> • MP: successful procedure. • MP*: other person's utility from going through the procedure is 2. 	<ul style="list-style-type: none"> • MP: unsuccessful procedure. • MP*: other person's utility from going through the procedure is 0.
Do nothing	<ul style="list-style-type: none"> • MP: mild pain • MP*: other person's utility is 1 	<ul style="list-style-type: none"> • MP: mild pain • MP*: other person's utility is 1
Go through the procedure	<ul style="list-style-type: none"> • MP: no pain • MP*: other person's utility is 2 	<ul style="list-style-type: none"> • MP: severe pain • MP*: other person's utility is 0

is always defined with respect to some function that assigns numerical values to possible outcomes. It says that in a state of the world in which this function assigns to a given outcome a numerical value of x , the decision maker must attach a utility of x to this outcome.

Now, of late two different philosophical problems that involve choice situations structurally similar to *medical procedure** have been discussed in the literature and in both cases part of the debate was about whether rational agents should obey certain NV norms.⁵

The first problem is that of decision making in face of expected experience that is both personally and epistemically transformative (in the sense introduced by Lauri Paul in Paul 2014). An experience which is epistemically transformative is an experience following which the agent gains some knowledge that she is unable to acquire without going through this experience (or in any case her epistemic state changes in a given way for which going through the experience is a necessary condition). An experience which is personally transformative is an experience following which the agent's fundamental preferences change in a significant way.

Cases in which people go through experiences which are both epistemically and personally transformative are not uncommon. In many of these cases the transformations involve changes in the phenomenal character of certain experiences. Paul's paradigmatic example is the experience of becoming a parent. She argues that becoming a parent is epistemically transformative because it typically involves gaining—through the experience of the phenomenal character of interactions with one's child—knowledge that cannot be gained in other ways. However, it is also personally transformative, because it is typically accompanied by deep changes in one's basic preferences and values.

5. This is not the way these debates have been framed in the literature but the framing I use here is helpful for the purposes of this paper.

Paul argues that in cases in which people have to make a choice whether to go through an experience which is both epistemically and personally transformative (such as the choice of whether to become a parent) the decision theoretic procedure described above cannot be used. This is so because in such cases the decision maker is unable to assign utilities to the different outcomes involved in the decision problem in a non-arbitrary way. However, and as will be discussed below, some philosophers have rejected this claim.

It will be helpful to construct a decision problem of this sort which is formally similar to MP and MP*:

Transformative Medical Procedure (Transformative MP)

You have to make a choice whether to go through a certain medical procedure that will have a deep impact on your personality (and no other effect). Every person who has gone through this procedure in the past has reported that he has changed dramatically in ways he cannot articulate. The behavioral impact of the procedure is also indisputable. All the 10,000 people who have gone through this procedure were asked to assess their general life satisfaction afterwards compared to their general life satisfaction before, when they were all instructed to use a cardinal scale and the benchmark of life satisfaction to degree 1 as a representation of their life satisfaction before the procedure. Exactly 4000 people answered that the degree of their general life satisfaction after the procedure is 2 and exactly 6000 people answered that it is 0. Should you go through the procedure?

The second problem is that of decision making under conditions of moral uncertainty. In such cases, morally motivated rational agents have to make a choice when they are uncertain what the moral value of each of their possible choices is. For example, consider the following case:

Medical Procedure under Moral Uncertainty (MP under Moral Uncertainty)

You have to choose for another person whether to go through a given medical procedure. There is no uncertainty regarding the outcome of the procedure involved. However, you are uncertain what the moral values of the two possible outcomes involved in the decision (i.e., the outcome of going through the operation and the outcome of doing nothing) are. While you are certain that the moral value of not sending that person through the procedure is 1, you are uncertain what the moral value of sending him through the procedure is. You believe to degree 0.4 that it is 2 and to degree 0.6 that it is 0. How should you choose?

With respect to both of these cases, several scholars have argued that the right decision rule to adopt is maximization of expected value while also implicitly being committed to certain NV norms.

Let us start with transformative MP. Pettigrew (2015), Bykvist and Stefansson (2017), and others have argued that in cases like this the rational way⁶ to choose is to maximize expected value relative to one's uncertainty regarding one's future preferences. Of course, in the absence of a NV norm, this decision rule is incomplete. Although none of these scholars explicitly acknowledges that such a norm is needed, all implicitly commit themselves to such a norm. Without discussing the (interesting) differences between the NV norms each of these scholars implicitly adopts, we can present the general idea all seem to share.

Having gone through the transformative experience she expects to go through, if the agent is rational (and let us assume this is the case—the problem of an expected transformative experience that might make one irrational is very interesting but a different problem) then there must be a value function the expectation of which she maximizes. In the context of expected transformative experience, this value function arguably constitutes the correct value function. Now, although the agent—at the time she makes her decision—is uncertain which value function is the correct one, she can—on the basis of scientific evidence—have justified degrees of belief in different hypotheses regarding what this value function is and so can maximize expected value relative to this uncertainty and based on the NV norm that says that the value attached to each outcome in a state in which a given value function is the correct value function is the value this function attaches to the outcome. In other words, the claim is that in the Transformative MP case what I referred to in the introduction as the VUV thesis holds.

Now, as just presented, this suggestion is still incomplete. What is missing is a procedure for making value comparisons between different value functions. While it is true that a rational agent always maximizes the expectation of some value function, this function is unique only up to affine transformation. Thus, the NV norm specified does not point to a unique value function but rather to a set of value functions, which are all affine transformations of each other. Using different value functions out of this set when applying the VUV thesis may lead the agent to choose differently and thus a way to determine which value function should be used is needed.

However, this problem—the problem of value comparisons across different possible value functions (a problem that arises also in the case of moral uncertainty soon discussed)—is not the problem Paul has pointed to in her criticisms

6. Or at least “a rational way”; this is not always clear, but a careful reading of both papers mentioned leads me to interpret them as committed only to the weaker “a” claim. I thank an anonymous referee for pointing this out to me

of applying the VUV thesis to cases such as Transformative MP. There is no doubt that it is a serious problem that might be unsolvable (see Briggs 2015 for a useful discussion), but many philosophers believe that, at least in some cases, it is solvable (for an interesting recent suggestion see Pettigrew discussion in Pettigrew 2019) and so in this paper—following Paul—I will ignore it.

For example, in the case of Transformative MP, if—although you are uncertain how the procedure will change you—you are certain that all that you will care about (and, let us assume, all that you care about now) is the level of your life satisfaction as determined by self-reflection, then the problem of value comparisons across value functions does not seem to arise. The suggested NV norm tells you to attach to each outcome the value the relevant hypothesis regarding the level of life satisfaction you will judge yourself to be in attaches to this outcome. Surely, many people care about other aspects of the outcomes, not only about life satisfaction, but this does not significantly change the picture, for we can easily replace “life satisfaction” with a more complicated index that is also sensitive to other aspects of the outcomes. So long as some way of overcoming the problem of value comparison across different value functions is available, some version of the suggested NV norm is applicable.

Consider next the MP under Moral Uncertainty case. Several scholars (especially Lockhart 2000 and Sepielli 2013) have argued that in this kind of case the right decision rule to adopt is maximization of expected value with respect to one’s moral uncertainty. Again, in order to apply this decision rule one must first adopt a certain NV norm and find a method (for exactly the same reasons specified above) for inter-theoretical comparison of moral value. While the problem of value comparisons across different hypotheses regarding the correct value functions is arguably even more serious here (see Nissan-Rozen 2015b for a discussion), it is—once again—not the problem that Weatherson makes use of in his criticism of the application of the VUV thesis to cases such as MP under Moral Uncertainty,⁷ and there are many scholars (such as Sepielli 2013) that believe it is solvable. Thus, in this paper I ignore it.

If the problem of inter-theoretical value comparisons is solvable a commitment to the VUV thesis in cases such as MP under Moral Uncertainty amounts to a commitment to the NV-norm that says that the value of an outcome in a state in which a given moral theory is true is the moral value of this outcome according to that theory.

Now, both these applications of the VUV thesis seem straightforward and unproblematic. If all that you care about when making your decision is the utility of another person, then the NV-norm that tells you to value each outcome in

7. Weatherson actually argues against a much wider set of decision rules. I think, however, that his argument works only against the VUV thesis. I discuss this further below.

each state in which the other person's utility from this outcome is x , to degree x , seems the correct norm to adopt. Replacing "the utility of another person" with "your own life satisfaction" or with "moral value" in the last sentence does not seem to change anything substantial. However, or so I (following Weatherson and Paul) argue, it does. There is a problem with the application of the VUV thesis that arises in the cases of moral uncertainty and transformative choices that does not arise in the case of making choices for someone else. This problem is discussed in the next section.

The VUV Thesis and Acting Out of the Wrong Kind of Motivation

In both cases the problem is that accepting the relevant NV-norm entails acting from the wrong kind of motivation.

Paul argues that accepting the VUV thesis in cases such as Transformative MP amounts to choosing unauthentically. Paul's idea is the following. Since in Transformative MP the medical procedure is personally transformative, then from your point of view at the time of making the decision, although the person who will enjoy the life satisfaction achieved by the decision to go through the procedure is metaphysically identical to you, he is not the person whose point of view is the relevant one for assessing the outcomes of your choice.

When making choices the outcomes of which will be experienced only in the future, the relevant point of view for the assessment of these outcomes is the point of view of one's psychological future self, not the person one will evolve to be (for Paul, "being the same person" is a different relation from "being one's future self"; see Paul 2015: Footnote 9). The person you will physically become after going through the procedure is not your psychological future self. If you will go through the procedure your psychological future self will not exist. Thus, attaching to an outcome, at the time of making the decision, a value which is equal to how much you (or your metaphysical future self) would value this outcome is unjustified. It amounts to conflating one's physical future-self and one's psychological future-self.

However, one might argue, although the last consideration indeed leads to the conclusion that deferring to one's metaphysical future self's preferences when evaluating the outcomes in cases such as Transformative MP is not the way to go, it does not lead to the conclusion that the relevant value function one should defer to is inaccessible. One can evaluate one's physical future self's preferences from one's current point of view. In doing so one must attach values to complex outcomes that include a reference not only to one's physical, social and psychological states, but also to the way one evaluates these states. One must,

that is, attach values to states such as “I will be in a state, *S*, while valuing being in *S* to degree *x*”, that is, one must form what is usually called in the literature “extended preferences” (see Adler 2014; Broome 1998; and Greaves & Lederman in press for some good discussions). However, one might argue, it is perfectly possible to form such extended preferences in cases that involve expected transformative experience, and if this is so Paul’s objection fails.

Well, I understand Paul as arguing that forming such extended preferences is problematic because what is important to us when evaluating outcomes of the form “I will be in state, *S*, while valuing being in *S* to degree *x*” is not, on the one hand, any of the features of the state *S* itself, and not, on the other hand, the value we will end up attaching to being in *S*. Rather, it is the *phenomenal character* of valuing being in *S* to a certain degree, and this phenomenal character is inaccessible to us.⁸

We do not care about any of the features of *S* itself, because we will not experience any of these features from our current point of view, and we do not care about valuing being in *S* to a certain degree in the future because we value being in *S* differently now. What we do care about is what it feels like to value *S* to degree *x*, and this is exactly what we lack access to (as the experience is epistemically transformative).

Thus, as Paul notes (2015: 798), the kind of unauthenticity she points to is not that discussed by existentialist philosophers. It is an epistemic and metaphysical unauthenticity, not a psychological or social one. It is not a matter of having preferences that were formed in a way that is sensitive to the wrong variables; it is a matter of being unable to form one’s preferences in a way that is sensitive to the right variables.

However, notice that Paul’s argument is based on an empirical claim. It is based on the claim that what is indeed important to people in cases such as Transformative MP is the phenomenal character of experiencing a given state in a given way (and see Footnote 8 for a qualification of this claim). While this assumption is surely true with respect to some people in some cases, it seems that in many cases that involve expected transformative experiences the assumption does not hold with respect to at least some people.

Indeed, Bykvist and Stefansson (2017) argue that this assumption does not

8. An anonymous referee remarked that Paul does not argue that what we really care about is the phenomenal character of being in a given state in itself, but rather that we care about the *experience* of being in the given state. Experiencing a given state might have a content beyond the phenomenology of being in it. However, this content is learnt via learning the phenomenology of being in the state. I agree with the referee that this is an important distinction that one might be able to exploit in order to present an objection to Bykvist’s and Stefansson’s argument. However, as Bykvist and Stefansson ignore this distinction and as my argument in this paper is independent of it, I will ignore it as well in this paper. My argument holds even in case all that an agent cares about is the phenomenological character of a given experience.

hold with respect to *them* in the context of what Paul herself takes to be the paradigmatic case of expected transformative experience, that is, becoming a parent. Bykvist and Stefansson go even further and argue that they take the type of person with respect to which Paul's argument works, that is, the type of person who cares so much about the phenomenal character of the experience that she cannot evaluate in any useful way (i.e., not even assign a set of possible values) outcomes that involve the experience, to be "a very special type of person" (see 2017: 128), one who they call "the texture fetishist".

I think the criticism of Bykvist and Stefansson succeeds in the sense that it indeed shows that the argument from the inaccessibility of the phenomenal character of transformative experiences has a limited scope. However, I also believe that Paul's general claim that applying the VUV thesis in cases of expected transformative experience leads to choices that are unauthentic in the sense explicated above is correct and can be supported by another argument, one that does not apply only to texture fetishists. The argument in the next section applies, actually, to all people, except to those with the very specific type of preferences one might want to call "fetishistic".

Moving to the case of moral uncertainty, Weatherson's argument against the application of the VUV thesis to cases such as MP under Moral uncertainty relies too on a claim about the kind of motivation that such an application requires. Weatherson (2014) argues that obeying any decision rule for choice under conditions of moral uncertainty that is sensitive to one's moral uncertainty must involve acting out of a *de dicto* moral motivation. He further argues (following Smith 2002) that such a motivation is fetishistic and thus unvirtuous and so it cannot be true that morality demands that one respect such a decision rule.

Although there is much to be said about Weatherson's argument (see Sepielli 2016 for a comprehensive discussion), my interest here is only with the first part of his argument, that is, with the claim that any decision rule that is sensitive to moral uncertainty must involve acting out of a *de dicto* moral motivation.

A *de dicto* moral motivation is a motivation to do the right thing, *as such*, that is, irrespectively of what "the right thing" is. In a nutshell, such a motivation seems problematic to some philosophers (such as Smith 1994; Weatherson 2014; and also myself) because of the thought that a virtuous person should care about the reason-giving features of the world (such as *this* instance of suffering of *this* person or *that* limitation of freedom of *that* person), not about the title "the right thing to do". Now, admittedly, many philosophers (see, e.g., Carbonell 2013) reject this latter claim and argue that—at least in many cases—there is nothing wrong with acting out of a *de dicto* moral motivation. Other philosophers (see, e.g., Olson 2002) argue that there is nothing wrong in having a *de dicto* moral motivation as long as it is accompanied by a corresponding *de re* motivation. I will not enter into this debate. The type of *de dicto* moral motivation to which I

point in the next section is—in a clear sense—the worst kind possible from the point of view of those who take a *de dicto* moral motivation to be problematic (those who do not accept the mere claim that at least in some cases there is something problematic about a *de dicto* moral motivation will not be impressed by my argument, I suspect).

Now, Weatherson does not give a general argument for his claim. Instead he argues by an analogy with cases in which an agent is uncertain about what constitutes his own welfare. In the cases Weatherson considers an agent is uncertain whether some type of experience, *E*, a type which in fact (so we are asked to assume) plays no constitutive part in the agent's welfare, actually plays such a role. In such cases, if the agent uses a decision rule that is sensitive to her uncertainty regarding whether *E* is a constituent of her welfare, it must be because she cares about increasing her welfare as such (since by assumption, *E* is not a constituent of welfare). However, such a motivation, a motivation to increase one's own welfare, even if this increase is accompanied by no change either in one's environment or in one's physical or mental state with respect to which one has any positive attitude, seems fetishistic.

Similarly, argues Weatherson, if an agent is uncertain whether some feature of reality, *F*, is morally valuable, then using a decision rule which is sensitive to this uncertainty must be the result of a motivation to do "the morally right thing" as such (i.e., independently of what the morally right thing actually is).

However, this argument by analogy is incomplete. First, it only applies to cases in which the uncertainty is about whether some feature of reality is morally significant (by analogy from the case in which the uncertainty is about whether some type of an experience is a constituent of welfare). It does not apply to cases in which there is no uncertainty regarding which features of reality are morally significant, only about which one of the possible acts available to the agent is the most morally valuable. This latter type of cases is analogous to cases in which one is certain that two different types of experience contribute to one's welfare but is uncertain which one of them contributes more. There seems to be nothing fetishistic or problematic in any way in taking such uncertainty into account when deciding how to act.

Second, there is an important difference between the welfare case and the morality case even when the uncertainty is about the mere moral significance of some feature of reality. When it comes to welfare, most plausible accounts of welfare (as well as most people's intuitive judgments regarding it) accept that happiness, joy, and any other kind of positive subjective experience are constituents of welfare. They might not be the only constituents of welfare, maybe not even the most important ones, but they are constituents of welfare. However, in Weatherson's examples we are asked to imagine that although the agent is uncertain whether *E* is a constituent of welfare, *E* is not. Thus, we are asked to

imagine a case in which although the agent assigns some probability to E being a constituent of welfare, in terms of her subjective experience she is completely indifferent to E.

The moral uncertainty case is importantly different in this respect. In the moral uncertainty case, even when F is not morally significant it can be significant to the agent in non-moral ways. Thus, although it might be the case, for example, that there is nothing morally wrong about eating animals, an agent who is uncertain whether eating animals is wrong may avoid eating animals not because she has a *de dicto* desire to “avoid doing morally impermissible things”, but rather because she cares about animals in a first order, non-moral, *de re* way.

Indeed (and as Sepielli 2016 argues) it seems odd that while an agent can act out of a *de re* moral motivation when she is certain that F is morally significant, she loses this ability once she become ever so slightly uncertain that F is morally significant. It seems both psychologically and conceptually more plausible to argue that acting out of a *de re* motivation to bring about F is also possible for agents who are uncertain whether F is morally significant. Of course, if F is not morally significant then the *de re* motivation the agent has is not a moral motivation, but this does not imply that an agent who uses a decision rule that is sensitive to her moral uncertainty is acting out of a *de dicto* moral motivation. Rather it is consistent with the claim that such an agent is either acting out of a *de re* moral motivation or out of a *de re* motivation which is non-moral.

Thus, I agree with Sepielli that Weatherson’s strong claim—that by using any decision rule that is sensitive to one’s moral uncertainty one must act on a *de dicto* moral motivation—is false. I do, however, think that the weaker claim, according to which using the decision rule of maximization of expected moral value relative to one’s moral uncertainty, necessarily involves acting on a *de dicto* moral motivation, and, or so I argue, in the worst possible way.

The Value of Information and the Wrong Kind of Motivation

The argument takes the form of a dilemma. I present—in an informal way—a decision situation in which an agent can buy some information that can help him make the decision. The decision situation (and more specifically—the possible outcomes involved in it) can be individuated in two ways. Individuated one way, following the VUV thesis amounts to a violation of a standard decision theoretic axiom, the axiom of Dominance. Individuated the other way, following the VUV thesis amounts to a violation of the demand to act out of the “right kind” of motivation (when the latter term is interpreted differently in moral uncertainty cases than in transformative choice cases).

Here is the case. A decision maker, Vahan, is about to go through a trans-

formative experience. He is uncertain how the experience will affect his preferences. Specifically, he is uncertain which one of two possible utility functions (fully comparable, or so—for the sake of the argument—I assume in this paper) he will end up having, U_1 or U_2 . The two possible utility functions are represented in table 2.

Table 2

Outcome	U_1	U_2
A	2	0
B	0	2
C	1.5	-11.5
D	-11.5	1.5

Vahan has to choose between two possible acts, a and b. Act a brings outcome A with certainty and act b brings outcome B with certainty. Suppose Vahan assigns a probability value of 0.5 to each of the two possibilities, ending up having U_1 and ending up having U_2 , so his expected utility of both acts—*calculated using the VUV thesis*—is 1.

God pays a visit to Vahan and offers him the following deal: she will tell Vahan which utility function he will end up having and in exchange Vahan will pay her 0.5 unit of utility (measured on the same scale Vahan used when he made the comparison between U_1 and U_2).

God then specifies the practical details of the deal: after she tells Vahan which utility function he will end up having, she will replace outcome A or outcome B in the choice Vahan faces with another outcome. If Vahan ends up having U_1 , she will replace A with C and if Vahan ends up having U_2 she will replace B with D. In any case, Vahan can be certain he will face a choice between an act that brings an outcome of 1.5 units of utility and an act that brings an outcome of 0 units of utility *according to the utility function he will end up having*. Thus, no matter how the experience will change Vahan, he will be able to choose an act that brings 1.5 units of utility according to the utility function he will end up having.⁹

Vahan, then, has to choose between three possible acts: rejecting God's offer and choosing act A, rejecting God's offer and choosing act B and accepting God's offer. In case Vahan chooses to accept God's offer he is certain—let us assume—

9. The argument does not rest on the assumption that such an offer can be made in a credible way (or on the existence of God). A similar argument can be made using completely realistic assumptions. A realistic example might be a case in which Vahan can receive some information that although it will not make him raise his degree of belief that his future utility function is either U_1 or U_2 to 1, will make him raise his degree of belief in one of these hypothesis to some extent. In such a case the expected utility calculations would become a bit less straightforward, but nothing substantial in the argument changes.

that his future self will choose rationally (i.e., will choose the act that brings the outcome that has—according to the utility function Vahan will end up having—a utility of 1.5). Thus, Vahan’s choice can be represented as in table 3 below.

Table 3

	U1 (probability 0.5)	U2 (probability 0.5)
Act a	A (utility: 2)	A (utility: 0)
Act b	B (utility: 0)	B (utility: 2)
Act l	C (utility: 1.5)	D (utility: 1.5)

It is straightforward to see that that if Vahan obeys the VUV thesis he should accept God’s offer. However, by accepting the deal Vahan in fact accepts an equal probability lottery between outcome C and outcome D. The expected utility of such a lottery *according to both U₁ and U₂* is -5, which is much lower than the expected utility of acts a and b (which are 0 and 2 or 2 and 0 correspondingly, depending on which utility function, U₁ or U₂, is the correct one).

Thus, on the face of it, this example shows that the VUV thesis can be in direct conflict with one of the most fundamental conditions of rationality, the Dominance condition. Dominance requires that if an agent prefers one lottery, l, to another lottery, m, if some event, E, is true, and also prefers l to m if E is false, he must prefer l to m even if he is uncertain whether E is true. In our example, E is the event “Vahan will end up having U₁”. If this event is true Vahan prefers both acts a and b to an equal probability lottery between outcome C and outcome D, and the same is true if this event is false (i.e., if Vahan ends up having U₂). It seems, then, that the VUV thesis *must* lead in some cases to violations of one of the most fundamental principles of rationality and so cannot be the correct decision rule for choices in cases of expected transformative experience.¹⁰

At this point, one might argue that the description above is based on a mischaracterization of the choice situation Vahan faces. God’s offer to Vahan does not amount to an equal probability lottery between C and D. Rather it amounts to an equal probability lottery between C and D *such that whatever the outcome of the lottery is, it is the preferred outcome*. Thus, the correct way to describe God’s offer, one might argue, is as an equal probability lottery between “C and C’s utility for future Vahan is 1.5” to “D and D’s utility for future Vahan is 1.5”, and the utility of “C and C’s utility for future Vahan is 1.5” might be different from the utility of C.

10. Notice that even if one rejects the event-wise Dominance principle presented in the main text and instead commits oneself only to the weaker state-wise Dominance principle (according to which if the outcome of one act is preferred to the outcome of another act, *in every state*, the first should be preferred to the second) the violation of Dominance still holds. I thank an anonymous referee for pointing this out to me.

If this line of reasoning is correct, however, then what the demand to maximize expected utility amounts to in the example is a demand to assign an expected utility of at least 1 to an equal probability lottery between “C and C’s utility for future Vahan is 1.5” and “D and D’s utility for future Vahan is 1.5” (otherwise the conflict with Dominance re-arises). This demand rules out as irrational many utility assignments that intuitively seem rationally permitted. Specifically, it does not allow Vahan to be indifferent to (or value only to a low degree) the mere fact that his future-self assigns a certain utility level to a certain outcome. It does not allow, for example, that a rational Vahan will be indifferent between C and “C and C’s utility for future Vahan is 1.5”. Such an attitude seems, however, not only permitted by rationality (as there is nothing irrational in caring only about the first-order features of a given state of affairs), but also—one might argue—dictated in some cases (i.e., in cases in which Vahan does not care at all about the utility function of his future-self) by the demand to choose authentically.

Notice that the argument does not depend on the specific re-individuation of the outcomes presented in the previous paragraph. Any other re-individuation of the outcomes that refers to the utility of the outcomes according to the utility function of future Vahan will do. This is so since, by stipulation, there is no first-order difference between C and the re-individuated-C. The only difference between them is the utility they are assigned by Vahan’s future utility function (and similarly for D).¹¹

A comparison with the case of choosing for someone else might help here. Suppose now U_1 and U_2 represent not hypotheses about Vahan’s own future utility function, but rather hypotheses about the utility function of Amal, Vahan’s unconscious friend. Suppose further that Vahan has to choose for Amal and that in doing so he is only interested in maximizing Amal’s utility (and notice that we *stipulate this*). In such a case it seems that respecting the NV-norm, which says that the value Vahan should assign to any outcome of the form “M and Amal’s utility from M is x ” is x , is required (out of a conceptual necessity: this is what it means, so it seems, “to be interested only in maximizing Amal’s utility”).

Moving back to our original interpretation, in which U_1 and U_2 are hypotheses about Vahan’s own future utility function, to claim that Vahan should use the NV-norm, according to which the value of “M and future-Vahan’s utility from M is x ” is x , is to claim that Vahan should only be interested in maximizing Vahan’s own future utility. Why, however, assume that this is the case? Surely, Vahan is rationally permitted to care solely about the utility function of his future self, but in the same vein, Vahan is also rationally permitted to care solely about the utility function of another agent (e.g., about Amal’s utility function).

11. Notice also that the argument does not depend on the price God sets for her offer (of 0.5 unit of utility). Nothing in the argument would change if God would demand a higher price as long as it is lower than 1 unit of utility.

As Paul explained, there is nothing special about the utility function of Vahan's metaphysical future self compared to the utility functions of other agents when it comes to cases in which Vahan is certain that he is going to go through a transformative experience.¹²

Recall that Bykvist and Stefansson's objection to Paul's argument was that it applies only to a special kind of person, one whom they called the texture fetishist. Now we see, however, that when it comes to the argument presented here it is exactly the other way around: there is only one kind of person with respect to whom our argument does not apply, a person who cares solely about the utility function of his metaphysical future self. Many people, however, are not such people and so—for those people—the demand to respect the VUV thesis in cases like Vahan's amounts to either a violation of Dominance or to a demand to attach to outcomes values that do not represent what these people really care about. The latter possibility is, of course, an unwelcomed type of unauthenticity.

It is worth emphasizing that according to the line of reasoning presented above, it is not only that Vahan should care about his metaphysical future self's utility. This line of reasoning leads to the much stronger conclusion that the value Vahan assigns to outcomes is *determined solely* by Vahan's expectation regarding his future utility function. The features of C that Vahan cares about, the features that makes Vahan value C in the way he does, are completely irrelevant to the way he evaluates "C and C's utility for future Vahan is 1.5". The value he attaches to this outcome must be—in order to avoid the conflict with Dominance—at least 1 independently of which first-order features C has (and according to the VUV-thesis it must be exactly 1.5).¹³

One might argue that all that this argument show is that *there are cases*, cases that involve the possibility of buying more information about one's future preferences, in which a commitment to the VUV must lead to the dilemma just specified. However, this reply misses the point. The possibility of buying information only reveals what a commitment to the VUV thesis in cases that involve expected transformative experience always amounts to. It amounts to caring solely about the utility function one's metaphysical future self will turn out to have. For even if the possibility of buying information is absent in a given choice situation, in the (possibly hypothetical) case in which it will become available, a

12. A similar—but possibly weaker—point can be made with respect to any NV-norm that refers to Vahan's future utility function (e.g., made with respect to the norm according to which the value of "M and future-Vahan's utility function from M is x" is a weighted average of x and the utility of M according to Vahan's current utility function, the set of people against which the charge of unauthenticity cannot be made is larger). However, the VUV-thesis—against which I argue here—is based on the NV-norm mentioned in the main text.

13. And I could have, of course, used any number greater than 1. There is nothing special about 1.5.

deal such as the deal God offers in the example can be constructed and give rise to the dilemma.

A very similar point holds in the case of choices under conditions of moral uncertainty. To demonstrate this we can use the same example, now interpreting U_1 and U_2 not as hypotheses regarding Vahan's future utility function but rather as two moral theories that specify the moral value of each outcome.

Under this interpretation, in order to resist the conclusion that a morally motivated rational agent *must* in some cases violate Dominance, the VUV advocate must differentiate between C and "C and C's moral value is 1.5" and between D and "D and D's moral value is 1.5". In other words, the VUV advocate must rule out as normatively impermissible many assignments of moral value that intuitively seem permitted. Specifically, he must demand that Vahan will not be morally indifferent to (or value only to a low degree) the mere fact that the moral value of some outcome is 1.5. He does not allow, for example, that a rational and morally motivated Vahan will be indifferent between C and "C and C's moral value is 1.5".

Such an attitude seems, however, not only permitted by both morality and rationality, but also—one might argue—dictated by the demand not to be motivated by a *de dicto* moral motivation. Since, by stipulation, C and "C's moral value is 1.5" are identical in every first order property, valuing them differently must be due solely to the mere fact that in the case of the latter it is true that C's moral value is 1.5. Thus, if one does value them differently one must have a *de dicto* moral motivation, a motivation to maximize moral value, independently of the first order properties in which this moral value is grounded.

Notice that unlike the case of Weatherson's original argument, the argument here does not apply to any decision rule for choices under conditions of moral uncertainty that is sensitive to the moral uncertainty. The argument here targets only the demand to maximize expected moral value relative to one's moral uncertainty accompanied by the NV-norm that says that the value attached to an outcome in each state in which the moral value of the outcome is x , should be x .

It might be possible to present a similar argument with respect to other decision rules for choices under conditions of moral uncertainty that make use of inter-theoretical comparisons of moral value. However, the argument is clearly not extendable to decision rules that are only sensitive to moral uncertainty but not to the degrees of moral value under the different hypotheses.

I take this to be a welcome result. One of Sepielli's (2016) arguments against Weatherson is the following. Even if a *de dicto* moral motivation is unvirtuous, since—according to Weatherson's own argument—all decision rules for choices under conditions of moral uncertainty involve a *de dicto* moral motivation and since choosing on the basis of no decision rule in such cases is unvirtuous too (so argues Sepielli), the demand that the correct decision rule does not necessar-

ily involve unvirtuousness cannot be satisfied and thus Weatherson's argument against decision rules that are sensitive to moral uncertainty, which is based on this demand, fails.

If the argument here is correct then there are decision rules for choice under conditions of moral uncertainty that do not involve a *de dicto* moral motivation (i.e., decision rules that are not sensitive to the degree of moral value under the different hypotheses but are sensitive to the moral uncertainty¹⁴), and thus one can argue that one of these decision rules is the one we should adopt.

Conclusion

The VUV thesis says that when one is uncertain which utility function is the correct one, one should maximize the expectation of the value function that assigns to each outcome, under each hypothesis regarding the correct utility function, the utility that the hypothesis assigns to the outcome. I have argued that the VUV thesis is false, both in the case of moral uncertainty and in the case of expected transformative experience.

Although the arguments for the two cases are based on the same formal phenomenon, the interpretations of the phenomenon and their philosophical significance are different in the two cases. Moreover, the VUV thesis does hold in some cases (e.g., in the cases of choosing for someone else). Thus, the lesson of this paper is a cautionary one: the demand to maximize expected value, in cases in which one is uncertain which value function is the correct one, must involve either implicit or explicit reference to a specific NV-norm. However, such a commitment is very demanding. Before adopting it in a given context one should make sure one is willing to commit oneself to all of its implications.

Acknowledgments

I thank Ron Aboodi, Jonathan Barzilai, David Enoch, Preston Werner and two anonymous referees for their very helpful comments.

References

- Adler, Matthew (2014). Extended Preferences and Interpersonal Comparisons: A New Account. *Economics and Philosophy*, 30(2), 123–162.
 Aboodi, Ron (2017). One Thought Too Few: Where *De Dicto* Moral Motivation Is Necessary. *Ethical Theory and Moral Practice*, 20(2), 223–237.

14. For two suggestions see Gustafsson and Torpman (2014) and Nissan-Rozen (2012).

- Bradley, Richard and Hylinur Orri Stefansson (2016). Desire, Expectation and Invariance. *Mind*, 125(499), 691–725.
- Briggs, Rachel (2015). Transformative Experience and Interpersonal Utility Comparisons. *Res Philosophica*, 92(2), 189–216.
- Broome, John (1991). Desire, Belief and Expectation. *Mind*, 100(2), 265–267.
- Broome, John (1998). Extended Preferences. In Christoph Fehige, Georg Meggle and Ulla Wessels (Eds), *Preferences* (271–287). De Gruyter.
- Bykvist, Krister and Hylinur Orri Stefansson (2017). Transformative Experience and Rational Choice. *Economics and Philosophy*, 33(1), 125–138.
- Corbonell, Vanessa (2013). De Dicto Desires and Morality as Fetish. *Philosophical Studies*, 163(2), 459–477.
- Greaves, Hilary and Harvey Lederman (in press). Extended Preferences and Interpersonal Comparisons of Well-Being. *Philosophy and Phenomenological Research*.
- Gustafsson, E. Johan and Olle Torpman (2014). In Defence of My Favorite Theory. *Pacific Philosophical Quarterly*, 95(2), 159–174.
- Lewis, David (1988). Desire as Belief. *Mind*, 97(387), 323–332.
- Lockhart, Ted (2000). *Moral Uncertainty and Its Consequences*. Oxford University Press.
- Nissan-Rozen, Ittay (2012). Doing the Best One Can: A New Justification for the Use of Lotteries. *Erasmus Journal for Philosophy and Economics*, 5(1), 45–72.
- Nissan-Rozen, Ittay (2015a). A Triviality Result for the “Desire by Necessity” Thesis. *Synthese*, 192(8), 2535–2556.
- Nissan-Rozen, Ittay (2015b). Against Moral Hedging. *Economics and Philosophy*, 31(3), 349–369.
- Olson, Jonas (2002). Are Desires De Dicto Fetishistic? *Inquiry: An Interdisciplinary Journal of Philosophy*, 45(1), 89–96.
- Paul, Lauri A. (2014). *Transformative Experience*. Oxford University Press.
- Paul, Lauri A. (2015). Précis of Transformative Experience and Reply to Symposiasts Elizabeth Barnes, John Campbell, and Richard Pettigrew. *Philosophy and Phenomenological Research*, 91(3), 794–813.
- Pettigrew, Richard (2015). Transformative Experience and Decision Theory. *Philosophy and Phenomenological Research*, 91(3), 766–774.
- Pettigrew, Richard (2019). *Choosing for Changing Selves*. Under contract with Oxford University Press. Retrieved from: <https://richardpettigrew.com/books/choosing-book/>
- Price, Huw (1989). Defending Desire-as-Belief. *Mind*, 98(389), 119–127.
- Savage, Leonard Jimmie (1972). *The Foundations of Statistics*. Dover Publications.
- Sepielli, Andrew (2013). Moral Uncertainty and the Principle of Equity among Moral Theories. *Philosophy and Phenomenological Research*, 86(3), 580–589.
- Sepielli, Andrew (2016). Moral Uncertainty and Fetishistic Motivation. *Philosophical Studies*, 173(11), 2951–2968.
- Smith, Michael (1994). *The Moral Problem*. Blackwell.
- Smith, Michael (2002). Evaluation, Uncertainty and Motivation. *Ethical Theory and Moral Practice*, 5(3), 305–320.
- Weatherson, Brian (2014). Running Risks Morally. *Philosophical Studies*, 167(1), 141–163.