The Right Version of 'the Right Kind of Solution to the Wrong Kind of Reason Problem'

LARS SAMUELSSON

Umeå University

In a recent article in *Utilitas*, Gerald Lang suggests a solution to the so called 'wrong kind of reason problem' (WKR problem) for the buck-passing account of value. In two separate replies to Lang, Jonas Olson and John Brunero, respectively, point out serious problems with Lang's suggestion, and at least Olson concludes that the solution Lang opts for is of the wrong *kind* for solving the WKR problem. I argue that while both Olson and Brunero have indeed identified considerable flaws in Lang's particular suggestion for solution to the WKR problem, they have not provided sufficient grounds for dismissing the *kind* of solution that Lang proposes. I show how a version of this kind of solution can be formulated so as to avoid both Olson's and Brunero's objections. I also raise some worries concerning an alternative solution to the WKR problem suggested by Sven Danielsson and Jonas Olson.

I. INTRODUCTION

According to T. M. Scanlon's influential buck-passing account of value (BPV), being valuable 'is not a property that itself provides a reason to respond to a thing in certain [positive] ways. Rather, to be…valuable is to have other properties that constitute such reasons'.[1] Thus, when it comes to the role of providing reasons for responding to a valuable thing, the buck is passed from the value-properties of that thing to some of its other, lower-order properties – hence the name of the account: 'the buck-passing account of value'.

Of the several challenges that have been posed for BPV, the one most extensively discussed is no doubt the so called 'wrong kind of reason problem' (WKR problem).[2] The source of this problem is that there are cases in which we seem to have reason to respond in a

positive way to a thing despite the fact that this thing lacks value. Let us call such cases 'WKR cases'. The WKR problem is raised by providing some WKR case. My discussion in this contribution proceeds from a WKR case formulated by Wlodek Rabinowicz and Toni Rønnow-Rasmussen: Suppose that a powerful evil demon threatens to inflict severe pain on us unless we admire him. This threat seems to provide a strong reason to admire the demon, but it certainly does not make him valuable.[3]

According to BPV, the property of a thing of being valuable is simply the property of possessing some other, lower-order property that provides reason to respond to it in certain positive ways. The demon possesses such a lower-order property: his determination to punish us unless we admire him. How, then, can proponents of BPV account for the fact that the demon lacks value? This is the WKR problem. Somehow the reason to admire the demon must be of the wrong kind for corresponding to a value of the demon.[4] The challenge that the WKR problem poses for buck-passers is hence to formulate BPV in such a way that reasons of the wrong kind can be convincingly distinguished from reasons of the right kind.

In his recent article 'The Right Kind of Solution to the Wrong Kind of Reason Problem', Gerald Lang attempts to solve the WKR problem. After having discussed and (correctly, I believe) dismissed several versions of BPV taken by their adherents to provide solutions to the WKR problem, Lang formulates his own version of BPV (BPV6), which he takes to be resistant to all WKR cases:

BPV6

X is good if and only if *X* has properties (other than its being good) that give us reason to have a certain positive attitude towards *X*, just as long as those properties of *X* that give us reason to have that attitude towards *X* would still be reason-giving in the absence of the benefits to us of having that attitude towards *X*.[5]

This version of BPV resists the demon case, since the reason to admire the demon is provided precisely by our benefits of doing so, namely our avoidance of severe pain.

It may seem that BPV6 only covers non-instrumental value, and not instrumental value: a thing that benefits us when we adopt a positive attitude towards it is instrumentally valuable, but would not be so (on account of this feature) in the absence of these benefits. However, I think that BPV6 actually works for instrumental value as well. On a basic level there are two ways to conceive of instrumental values: (1) instrumental values are not really values at all, but merely means to something else that is valuable; (2) at least some instrumental values are indeed genuine values, albeit values based on the property of being a means to something else that is valuable.[6] If we opt for (1), BPV *should* not cover instrumental value, since it is an account of *genuine* value. If, on the other hand, we opt for (2), it seems to me that BPV6 does indeed hold even for genuine instrumental value. If *X* is *really* valuable because of its property (*p*) of being a means to something else that is valuable, then there is some reason to have some positive response (*r*) towards *X* because of its possessing *p* (otherwise it is hard to see in what sense *X* itself is *really* valuable on account of its possessing *p*), and this reason does not depend on any benefits of responding with *r* towards *X*.[7]

In two separate replies to Lang, Jonas Olson and John Brunero, respectively, point out serious problems with Lang's suggestion, and at least Olson concludes that the solution Lang proposes is of the wrong *kind* for solving the WKR problem.[8] In this article I argue that while both Olson and Brunero have indeed identified considerable flaws in Lang's particular suggestion for solution to the WKR problem, they have not provided sufficient grounds for dismissing the *kind* of solution that he opts for. I offer a version of this kind of solution that is resistant to both Olson's and Brunero's objections, and which we reach by making some highly plausible modifications of BPV6.

The *kind* of solution to the WKR problem to which I take Lang's suggestion to belong is neatly characterized by Sven Danielsson and Jonas Olson in the following passage:

> In the demon scenario…the reason to have the pro-attitude (of favouring the demon)…[is] provided by the desirable consequences of having the relevant attitudes. Intuitively, one is inclined to think that this is the heart of the WKR problem: reasons of the right kind are not provided by the consequences of taking up the relevant attitude.[9]

Despite finding this kind of solution intuitive, Danielsson and Olson reject it. But it is unclear on what grounds they reject it. They do not offer any arguments of their own against it, but refer instead to two works which they take to show that it is untenable:[10] Rabinowicz and Rønnow-Rasmussen, 'Buck-Passing and the Right Kind of Reasons',[11] and Lang, 'The Right Kind of Solution'. However, none of these works contains any argument against this kind of approach in general.[12] As we have seen, Lang even defends a version of it.

Whatever their reason for rejecting this kind of solution to the WKR problem, Danielsson and Olson provide their own innovative and interesting alternative to it; what they call 'a Brentano-style solution' (drawing, as it does, on Franz Brentano's notion of *correctness*). Although this proposed solution is among the ones that Lang discusses and rejects, there is reason to pay some additional attention to it here. The reason is that in his reply to Lang, Olson does not only criticize Lang's suggestion for solution to the WKR problem, he also defends the Brentano-style solution against Lang's objections.[13] Thus, if one takes Olson's reply to Lang to be convincing, my defence of the kind of solution to the WKR problem that Lang opts for does not, by itself, provide reason to prefer this kind of solution to the Brentano-style solution. I therefore end this article by raising some worries regarding the latter.

II. OLSON'S OBJECTION

While I do think that BPV6 points in the direction of the right *kind* of solution to the WKR problem, it is not the right version of this kind of solution. As Olson has shown, it is possible to formulate counter-examples to BPV6, i.e. WKR cases which it cannot handle.[14] BPV6 excludes, as reasons of the right kind, such reasons whose presence depends on the benefits that we would receive if we were to respond in accordance with these reasons. Olson's strategy is to show that these are not the only reasons that should be excluded. He writes:

> Suppose for instance that an evil demon threatens to inflict severe pain on people on the other side of Earth, who are all strangers to us, unless we favour him. That seems to provide reason for us to favour the demon even though we would not benefit from favouring him.[15]

This problem can be easily taken care of, as Olson notes, by modifying BPV6 so that it also excludes – from being of the right kind – such reasons whose presence depends on the benefits that others would receive if we were to respond in accordance with them. But further problems are looming:

> Suppose that an evil demon threatens to create a number of individuals who suffer greatly unless we favour him. If we favour him he will not bring these suffering individuals into existence. That seems to provide reason for us to favour the demon even though no one would benefit from our favouring him. This is because one cannot benefit individuals who do not and will not exist.[16]

In order to deal with this problem we need to modify BPV6 once again, so that it excludes – from being of the right kind – such reasons whose presence depends on the fact that we would prevent occurrences of suffering if we were to respond in accordance with them.

We need not go through all the modifications of BPV6 (BPV7-BPV10) that Olson tests and rejects (by way of providing counter-examples involving an evil demon); the structure is the same throughout. For each new version of BPV a new counter-example is introduced in order to show that there are possible reason-giving consequences of favouring/not favouring the demon that this version cannot handle. In order to deal with those consequences we need to modify BPV again by adding new restrictions on reasons (or reason-giving properties) of the right kind, and so it continues. Thus Olson concludes:

> But it is obvious that only lack of imagination stops us from conjuring up demons that threaten to eradicate from collective memory the works of Plato, Aristotle, Dostoevsky, and what have you. In order to account for this possibly infinite series of counterexamples the reformulation of BPV10 would have to take the form of a hopelessly complex conjunction.[17]

Olson is certainly right that we cannot save the kind of BPV that Lang defends by adding more and more conjuncts to the requirement on reasons (or reason-giving properties) of the right kind (and even if we could, the resulting version of BPV would presumably come out as hopelessly *ad hoc*). However, this is not the way to save it. Lang's mistake was to formulate his requirement on reasons (or reason-giving properties) of the right kind too narrowly to begin with: he should not have restricted the relevant consequences to benefits, and he should not have required that the consequences (or benefits) be consequences *for* (or benefits *to*) someone or something. If a thing is valuable, then the corresponding reason to respond to it (in the way corresponding to its value) does not depend on *any* consequences of so responding. Differently put: in order to explain why there is reason to respond to a valuable thing, we need not refer to any result of having the relevant response.

Without Lang's restrictions we get the following version of BPV, which is resistant to all of Olson's counter-examples:

BPV11

X is good if and only if *X* has properties (other than its being good) that give us reason to have a certain positive attitude towards *X*, just as long as those properties of *X* that give us reason to have that attitude towards *X* would still be reason-giving in the absence of any consequences of having that attitude towards *X*.

I take the requirement on good-making properties (properties providing reasons of the right kind) expressed in BPV11 to be essential to the very idea of a *genuine value*.[18] If, for *every* positive response that we think there might be reason to have towards a certain thing *X*, we need to ask what consequences having that response towards *X* would have in order to find out whether there is *some* reason to respond to *X* in that way, then *X* is simply not the kind of thing that we take to possess a genuine value.[19] We may have many reasons to respond to *X* that do depend on the consequences of so responding, but if all of our reasons to respond to it are like that, then *X* is just not the kind of thing that we take to be genuinely valuable. We may express this idea by saying that our reason to respond in the appropriate way to a valuable thing (on account of the properties that make it valuable) does not depend on the consequences of so responding.

The consequences that first come to mind when one considers the WKR problem are probably causal consequences. For instance, the relevant consequence of admiring the demon is a causal consequence, broadly conceived; if we admire him we will not be punished. But it may be that we have to count among the consequences relevant to BPV also non-causal ones, since it is possible to formulate WKR cases involving consequences that at least appear not to be purely causal (consequences that seem to involve a conceptual component). Let me illustrate this point by providing two such cases. (If the consequences appealed to in these cases can be understood as purely causal after all, buck-passers might not have to bother

about non-causal consequences. But if we do need to exclude – from being able to provide reasons of the right kind – also non-causal consequences, I think that is only to be expected: As stated above, if we take a thing to have value, we should not have to enter into *any* considerations regarding what will *follow* if we respond to it (in the way that we take to correspond to its value) in order to find out whether we have *some* reason to respond to it in that way.)

(1) Consider a particular version of hedonistic rule-consequentialism. According to this theory, the only thing that has genuine (positive) value is the occurrence of pleasure, and the only legitimate *ultimate* ethical goal is to maximize the occurrence of pleasure (and minimize the occurrence of pain). However, we ought not to aim to maximize the occurrence of pleasure *directly*, since doing so is taken to be counterproductive. Instead we ought to adopt the strategy of following a certain set of rules, *S*, the general adoption of which is taken to have better consequences in the long run – in terms of maximizing the occurrence of pleasure (and minimizing the occurrence of pain) – than the consequences of adopting any other available strategy.

Suppose now that among the rules in *S* is a rule *R* according to which we ought to have a certain positive attitude *p* towards a certain thing *X* (where *X* is not pleasure or the occurrence of pleasure). On this version of rule-consequentialism, we ought to adopt *p* towards *X* in each case in which we encounter *X*, irrespective of any causal consequences of doing so in the particular case at hand, because having *p* towards *X* – in combination with following the other rules in *S* – is taken to have the best consequences in the long run (in terms of maximizing the occurrence of pleasure and minimizing the occurrence of pain).

Now, since *X* does not have genuine value according to this version of rule-consequentialism, there cannot be a reason of the right kind to have *p* towards *X*. Yet, according to this theory, there is a reason to have *p* towards *X* which does not seem to depend

on any causal consequences of having *p* towards *X* (the reason being that this is what *R* prescribes (given that *R* belongs to *S*, and so on)).[20] How, then, can this reason be of the wrong kind in the light of BPV11? The answer is that it depends on what appears to be a non-causal consequence of adopting *p* towards *X*, namely that one thereby follows a rule belonging to the set of rules (*S*), the general adoption of which has the best consequences. (This seems to be a conceptual consequence; it follows by virtue of what this version of rule-consequentialism – and in particular *R* – states.) In the absence of this consequence (and any other consequence that may give us this reason), there is no reason – from the point of view of this version of rule-consequentialism – to adopt *p* towards *X*. Our reason for having *p* towards *X* is therefore of the wrong kind for corresponding to a genuine value of *X*, according to BPV11 (which is the desired result).

(2) Suppose that some now deceased person *P* expressed the following preference regarding a certain object (of which we assume that it does not possess genuine value) before she passed away: 'I want this object *X*, which has meant a lot to me, to be preserved and cared for even after my death'. Some people think there is reason to satisfy a preference of this kind for the sake of the person who once held it. If that is the case, this reason to care for *X* seems to obtain irrespective of any purely causal consequences of caring for *X*. (*P* is not affected by our caring for *X*, since she is dead, and this particular reason is based solely on her preference, so effects on other parties than *P* are irrelevant as far as this reason is concerned.) But the reason does not obtain irrespective of *any* consequences of caring for *X*. One supposedly non-causal consequence of caring for *X* is that *P*'s preference is thereby satisfied. (This seems to be a conceptual consequence; it follows by virtue of what *P*'s preference expressed.) In the absence of this consequence there is no reason to care for *X*. Hence, the reason is of the wrong kind according to BPV11 (which is the desired result).

At least if we accept that non-causal consequences belong among the consequences on which reasons of the right kind are not allowed to depend, BPV11 also gives the right result with respect to a kind of WKR case suggested by Roger Crisp:

> Imagine that I say to someone, in normal circumstances, 'I promise to φ'. The fact that I have said these words seems to give me a reason to choose to φ. And because φ-ing has the property that I have promised to do it, we might think φ-ing is therefore valuable.[21]

Crisp's point is that one may well accept that my promising to φ gives me a reason to choose to φ without having to accept that φ-ing is valuable. However, BPV11 correctly deems this reason a reason of the wrong kind, since it would not obtain in the absence of a certain (seemingly non-causal) consequence of responding in accordance with it, namely the consequence that I thereby keep a promise.

(But what about my general reason to choose to keep promises? Would not that reason be independent of the consequences of so choosing? I guess it would, but if some entity *promise-keeping* is the kind of thing that can be good (that can be a value-bearer) I do not see why someone who thinks that I have a general reason to choose to keep promises (even in the absence of the consequences of so choosing) would not want to admit that *promise-keeping* is good. An alternative challenge would be to claim that I have a reason to choose to follow the principle according to which I ought to keep promises, and that BPV11 therefore implausibly implies that this principle is valuable. But this principle is itself a principle that states what I have reason to do, and it is odd to claim that I have further reasons to choose to follow principles that state what I have reason to do. It is like saying that I have reason to choose to do what I have reason to do, which is close to tautological. Then we could just as well say that I have reason to choose to follow this new principle (i.e. the principle that I have reason to choose to do what I have reason to do), and so on.)

Of course, one may also hold that actions of promise-keeping (such as φ-ing in Crisp's example) are indeed valuable (or perhaps more plausibly, that actions of promise-breaking are bad), but whether or not one holds such a view is reasonably related to one's opinion on our reasons to respond to such actions. If one believes, for instance, that we have reason to adopt negative attitudes to actions of promise-breaking (even in the absence of any consequences of adopting these attitudes), that seems tantamount to believing that such actions are bad. Crisp does not seem to agree, however. A few pages down from the passage quoted above he writes: 'I may admire your keeping your promises because I admire duty's being performed for its own sake, not because I think there is anything of value at stake'.[22] But the relevant question from the perspective of BPV is whether I *have reason to admire* duty's being performed for its own sake, and I have difficulties seeing how I can have reason to admire duty's being performed for its own sake (where this reason does not depend on the consequences of admiring duty's being performed for its own sake) if duty's being performed for its own sake is not admirable. And being admirable, of course, is a way of being valuable – a way of having value.

To my knowing, all WKR cases so far developed (including Olson's variants of the evil demon case) point towards some reason that is provided by some consideration of some consequence of responding in accordance with that reason – indeed, this feature seems to be what characterizes a WKR case. Thus it is difficult to see how one could construct a WKR case which lacks this feature. Accordingly, it is also difficult to see how one could provide a WKR case which is a counter-example to BPV11. I therefore think we can safely assert that BPV11 manages to exclude all reasons of the wrong kind. But perhaps BPV11 suffers from another flaw; perhaps it excludes too much. I shall now turn to two such worries. The first is raised by Brunero, who argues that Lang's formulation of BPV is incompatible with consequentialism, and the second is the worry that there may be some reasons of the right

kind that BPV11 incorrectly dismisses as reasons of the wrong kind. I begin with Brunero's objection.

III. BRUNERO'S OBJECTION

According to Brunero, BPV6 is incompatible with consequentialism and should therefore be rejected as a formal account of value. While I agree with Brunero on this point, it seems to me that he has not managed to adequately explain why BPV6 fails to comprise consequentialism. Once we explicitly spell out the correct explanation of this shortcoming of BPV6 – which it shares with BPV11 – it becomes clear how both BPV6 and BPV11 can be modified in a way that makes them immune to Brunero's objection. This modification is also plausible in its own right.

The explanation of why BPV6 and BPV11 are incompatible with at least many versions of consequentialism is actually quite simple. According to BPV6, if a thing is genuinely valuable there is reason to adopt a positive attitude towards it. But according to at least many versions of consequentialism, whether there is reason to adopt a positive attitude towards a certain thing depends on what would be the consequences of so doing. To be more precise, it depends on whether doing so would help to bring about or preserve things that are genuinely valuable. This point also applies to the things that are genuinely valuable themselves. For instance, if pleasure is the only thing that is genuinely valuable, then – in the view of many consequentialists – there is reason to adopt a positive attitude towards pleasure only in so far as doing so would actually help to bring about or preserve pleasure. If my having a positive attitude towards pleasure would not help to bring about or preserve pleasure, then there does not seem to be any reason, from a consequentialist perspective, for having that attitude.

Attitudes are not central to most versions of consequentialism. According to these theories, things that are valuable are things that we have reason to bring about and preserve.

Whether we should also have positive attitudes towards these things depends on the outcome of having these attitudes. As Brunero points out, if having such an attitude would not help to bring about value, there is no reason why a consequentialist should insist that we have reason to have that attitude.[23]

Hence, any version of BPV which takes values to correspond to reasons for *attitudes* (indeed, any value analysis according to which values correspond to reasons for *attitudes*) is incompatible with at least many forms of consequentialism, and should therefore be rejected as a formal account of value. Consequently, both BPV6 and BPV11 fail as formal accounts of value. However, the feature which is the source of their failure is easy to get rid of. In fact, we need only return to Scanlon's original formulation of BPV in order to see that.

Scanlon does not formulate BPV in terms of positive attitudes, but in terms of positive responses.[24] As I read Scanlon, he uses 'response' in a wide sense, so that 'bringing about' and 'preserving' count as responses.[25] When discussing hedonism, Scanlon writes: '...the reasons generated by value would all be of the same simple form: reasons to bring about the most valuable states of affairs'.[26] He also states that '[o]ne could accept such an account [BPV] while still holding a purely teleological conception of value, since nothing in the argument just given rules out the possibility that the reasons associated with something's being valuable are all reasons to promote it'.[27] In some passages Scanlon does write about value in terms of reasons for attitudes, but that is mainly before he has explained his buck-passing account in detail.[28] In any case, as I use 'response' in this article, 'bringing about' and 'preserving' count as responses.

If we modify BPV11 by simply substituting *responses* for *attitudes*, and thereby make it more true to Scanlon's original formulation, consequentialism will no longer pose a problem. The reason is that consequentialists do believe that valuable things merit certain *responses*, namely such responses as bringing about and preserving. We get:

BPV12

X is good if and only if *X* has properties (other than its being good) that give us reason to respond in a positive way to *X*, just as long as those properties of *X* that give us reason to respond in that way to *X* would still be reason-giving in the absence of any consequences of having that particular response.[29]

While Brunero is right that BPV6 fails to comprise most versions of consequentialism, he does not make clear that the problem with BPV6 is its focus on attitudes. Some of his remarks give the impression that he thinks the problem lies elsewhere, namely in the particular kind of attitude that Lang focuses on, *admiration*:

Our reasons to admire something…seem to persist in the absence of the benefits to be had from admiring it. But this is not so with other attitudes, including those attitudes central to consequentialism. Our reasons to *aim to promote* something of value do not persist in the absence of the benefits to be had from so aiming. And, for the indirect consequentialist, our reasons to have certain other attitudes (such as *honoring* personal loyalty) do not persist in the absence of the benefits to be had from having those attitudes. Perhaps Lang's mistake is that he focuses on reasons to admire at the expense of *other* attitudes we may, according to some ethical theory, have reason to have.[30]

This is not Lang's mistake. Lang's mistake is that he focuses on attitudes at the expense of other responses that we may, according to some ethical theory, have reason to have.

Brunero uses an example of a consequentialist according to whom personal loyalties (or the enjoyment of personal loyalties) have genuine value, but according to whom we ought not to *aim to promote* personal loyalties, since having the attitude of aiming to promote personal loyalties is not – according to this consequentialist – an efficient means to bring about such loyalties (or to bring about any other valuable thing). Instead we ought to *honour* personal loyalty, because having this attitude is actually the best way to bring about such loyalties.

Now, since the reason for having this attitude would not persist in the absence of the benefits of having it, it is a reason of the wrong kind, according to BPV6. Brunero thinks this shows that BPV6 is incorrect, since personal loyalties are indeed genuinely valuable according to the consequentialist in our example. He thus seems to think that a correct version of BPV should yield the result that this reason is of the right kind.[31] But this is a mistake. This reason is indeed a reason of the wrong kind, just as BPV6 has it. The genuine value that our consequentialist ascribes to personal loyalties does not correspond to any reason for having some *attitude* towards personal loyalty; it simply corresponds to the reason for the *response* of *bringing about* such loyalties. And this latter reason has to persist in the absence of any (other) benefits of having this response (and hence be a reason of the right kind according to BPV6), or else personal loyalties are not – in themselves – genuinely valuable according to this consequentialist theory. If there are also indirect reasons to *honour* personal loyalty – because doing so helps to bring about such loyalties – these reasons do not bear on the *genuine* value of personal loyalty, from a consequentialist perspective. Lang's mistake is simply that he formulates BPV in terms of attitudes instead of responses.

IV. A COUNTER-EXAMPLE TO BPV12

Formulating BPV in terms of responses makes it compatible with consequentialism, but a different problem remains. It seems that there may be some reasons of the right kind that BPV12 incorrectly dismisses as reasons of the wrong kind, which is shown by the following counter-example.

Imagine a very peculiar *good* demon, one who *necessarily* benefits those who admire him. Suppose that this feature of the demon makes him genuinely valuable, that it provides a reason of the right kind to respond in a certain positive way towards him (this may sound odd, but there is nothing incoherent or inconceivable about it). Suppose further that the way we

have reason to respond to the demon because of this feature is to admire him. Now, this case seems to provide a counter-example to BPV12, because our reason to admire the demon does *not* obtain in the absence of a certain consequence of admiring him, namely that we are benefited. If we admire the demon but the demon does *not* benefit us, then the demon obviously does *not* possess the property that was supposed to make him valuable – the property of *necessarily* benefiting those who admire him – and then we cannot have *this* reason to admire the demon. Hence the reason is of the wrong kind for corresponding to a genuine value, according to BPV12. (For this counter-example to work, the demon must be designed so that he *necessarily* benefits those who admire him. If the demon's supposed good-making property were such that he did not have to respond to *every* instance of admiration by benefiting the admirer, we could have reason to admire the demon because of this property even if our particular instance of admiration would not result in our being benefited by him. The good-making property would then be merely a disposition.[32])

While this counter-example means trouble for BPV12, it does not mean trouble for the idea behind it. This idea, remember, is that our reason to respond in the appropriate way to a valuable thing does not *depend* on the consequences of so responding. By saying that a reason depends on something, *X*, I mean that *X* plays a part in the direct explanation of that reason. In order to illustrate this point – and to explain it – let me compare our reason to admire the good demon with our reason to admire the evil one (i.e. the one that threatens to inflict severe pain on us unless we admire him), but let us modify the evil demon so that he *necessarily* inflicts severe pain on those who fail to admire him.

The Evil Demon

Which is the direct explanation of our reason to admire the evil demon? In order to answer this question, let us consider which *fact* provides our reason to admire this demon. This is

clearly the fact that we will experience severe pain unless we do. This is also the direct explanation of this reason. An alternative explanation would be that the demon possesses the property of necessarily inflicting severe pain on those who do not admire him. But this is an indirect explanation of this reason: it serves as an explanation only because we can derive from the fact that the demon possesses this feature the further fact that we will experience severe pain unless we admire him. It is this latter fact that does the entire job: it is this fact that provides our reason to admire the demon. (Obviously, there is nothing admirable about the demon's property of necessarily inflicting severe pain on those who do not admire him.) Hence, the consequence of admiring the demon plays a part in the direct explanation of our reason to admire the demon. The reason *depends* on this consequence.

This is perhaps even easier to see in the case of the original demon who merely threatens to inflict severe pain on those who do not admire him. Our reason to admire this demon is that unless we do there is a great risk that we will experience severe pain. But if I learn that this demon will in fact not inflict severe pain on me if I fail to admire him today, his property of threatening to inflict severe pain on those who do not admire him gives me no reason at all to admire him today. Thus my reason clearly depends on the consequence of admiring the demon, and this, of course, is true also in the case of the demon who *necessarily* inflicts severe pain on those who do not admire him.

The Good Demon

Which is the direct explanation of our reason to admire the good demon? In order to answer this question, let us consider which *fact* provides our reason to admire this demon. This is clearly the fact that the demon possesses a certain property which makes him admirable (the example assumes), namely the property of necessarily benefiting those who admire him. This is also the direct explanation of this reason. It might be thought that an

alternative explanation is that the demon will benefit us if we admire him. But this fact does not explain this reason. It explains a different, instrumental reason that we have to admire the demon, but this is not the reason that corresponds to the (supposed) genuine value of the demon. The genuine value of the demon is based on his property of necessarily benefiting those who admire him; this is the feature that makes him admirable (his good-making property). Hence, the consequence of my particular instance of admiring the demon plays no part in the direct explanation of my reason to admire the demon. The reason does not *depend* on this consequence.

In the case of the evil demon, citing the demon's *property* works as an explanation of the reason to admire him only because we can derive from the fact that the demon possesses this property the *real* reason to admire him: that doing so will help us avoid a certain bad consequence. The case of the good demon is quite different. Although our reason to admire this demon depends on a property that involves in its description the reference to a certain consequence (which also plays a part in explaining why this property makes the demon valuable), it is not the case that our reason to admire him is explained by any consequence of our particular instance of admiration. If someone asks us why the demon is admirable, it is not correct to answer 'because we will be benefited if we admire him'. That is not the kind of fact that can make a thing genuinely admirable. In order to explain why (we think that) the demon is genuinely admirable we have to refer to his extraordinary property of necessarily benefiting those who admire him. Citing the additional fact that we will in fact be benefited if we admire him on this particular occasion does not add anything to this explanation.

The fact that BPV12 is susceptible to the good demon counter-example is due to its unfortunate formulation (in terms of 'absence') of the requirement that a reason of the right kind must not depend on any consequence of responding in accordance with it (which is a

remnant from Lang's BPV6). So we need to reformulate this requirement. This move gives us the last modification of BPV6 that I will consider, which I also take to provide the right version of 'the right kind of solution to the WKR problem' (i.e. the kind of solution to which I take Lang's suggestion to belong):

BPV13

X is good if and only if *X* has some property (other than its being good) that gives us reason to respond in a positive way to *X*, where that reason must not depend on any consequence of having that particular response.

This version of BPV is resistant to Olson's and Brunero's objections, as well as to Crisp's 'promise-keeping-case' and the good demon counter-example. It also captures what seems to me to be essential to the very idea of a genuine value, namely that our reason to respond to such a value (or, rather, to the property by virtue of which it arises) does not depend on any consequences of so responding. I therefore take BPV13 to be a promising candidate for solution to the WKR problem.

As a final clarification, I think the consequences on which reasons corresponding to values must not depend include both consequences of the relevant response of particular responding agents, and consequences of the aggregated relevant responses of the members of groups of responding agents. Consider again the example from section II concerning a version of rule-consequentialism. According to this theory we ought to adopt the strategy of following a certain set of rules, *S*, the general adoption of which is taken to have the best consequences. It could be argued that my reason to (choose to) adopt this strategy (the *S*-strategy) does not depend on any consequences of *my* particular adoption of it (because: even if it would turn out not to make any difference with respect to good consequences whether or not *I* adopt the *S*-strategy, I still ought to adopt it according to this theory, since *our* adoption of this strategy

has the best consequences). If so, it might be thought that BPV13 incorrectly (from the perspective of this version of rule-consequentialism) assigns genuine value to the *S*-strategy (since *adopting* or *choosing to adopt* plausibly counts as a positive response of the kind relevant to BPV). However, this result is avoided if the reasons deemed by BPV13 to be of the wrong kind include also such reasons that depend on consequences of our aggregated relevant responses. And to exclude such reasons from corresponding to genuine values seems just as plausible from the point of view of this value analysis as does the exclusion of reasons that depend on consequences of such responses of particular agents. The reason for exclusion is the same: if we take a thing to be valuable, we should not have to enter into any considerations regarding what will be the result of responding to it in the way that we take to correspond to its value in order to find out whether we have *some* reason to respond to it in that way.

V. SOME WORRIES CONCERNING THE BRENTANO-STYLE SOLUTION TO THE WKR PROBLEM

Of course, that BPV13 is a promising candidate for solution to the WKR problem does not mean that it is preferable to other suggestions for solution; there may be worthy competitors. However, most of the alternatives that have been offered so far have been exposed to serious objections of various kinds.[33] I will devote the remainder of this article to considering what I take to be the currently most interesting alternative to the BPV13-solution to the WKR problem, namely Danielsson and Olson's Brentano-style solution. As stated in the introduction, even this suggestion for solution has been put under critique, in particular by Lang.[34] However, since Olson has replied to this critique,[35] one may plausibly hold that it has yet to be decided whether or not Danielsson and Olson's suggestion is successful. I argue that, despite its merits, there are reasons to resist it.

The Brentano-style solution invokes Franz Brentano's notion of *correctness*, which – according to Danielsson and Olson – is supposed to be analogous to truth (as Brentano conceived of it). While Brentano took truth to be a property of cognitive attitudes, he took correctness to be a property of conative attitudes.[36] Now, Danielsson and Olson's suggestion for solution to the WKR problem relies on a distinction between *holding-reasons* and *content-reasons*, of which the latter are understood in terms of correctness: 'One thing is a *reason for having the attitude*, let us call this a *holding-reason*; quite another thing is a *reason for the correctness of the attitude*, let us call this a *content-reason*'.[37] Danielsson and Olson explain:

> To cite content-reasons for a belief is to present arguments to the effect that the belief is (would be) true. To cite holding-reasons for a belief is to present arguments to the effect that we ought to have the belief. To cite holding-reasons for a conative attitude is to present arguments to the effect that we ought to have that conative attitude. To cite content-reasons for a conative attitude is to present arguments to the effect that that conative attitude is (would be) correct.[38]

Moreover, holding-reasons are understood in terms of content-reasons: 'To say that there is a holding-reason to have some attitude is to say that there is a content-reason to favour the occurrence of this attitude, or possibly that there is a content-reason to disfavour the non-occurrence of this attitude'.[39]

On the Brentano-style solution, content-reasons are the right kind of reasons from the point of view of BPV. Thus, whereas there is indeed a holding-reason to favour the evil demon (that is what we ought to do), there is no content-reason for doing so (i.e. admiration would not be a *correct* attitude to have towards the demon).

Here I want to raise two concerns about the Brentano-style solution.[40] The first is actualized by Brunero's objection to Lang's proposal. Like Lang's favoured version of BPV, the Brentano-style solution has difficulties to account for the value ascriptions of many

consequentialists. Second, the Brentano-style solution should only attract those who, like Brentano, take correctness (or some analogous notion) to be a primitive normative notion in terms of which the other normative notions are understood. And even those, I suggest, should consider it a last resort because of its limited explanatory force. I begin with the failure to accommodate consequentialism.

The Brentano-style solution is put in terms of correctness of *attitudes*. The version of BPV that Danielsson and Olson end up with states that '*x is good* means that *x has properties that provide content-reasons to favour x*', where a content-reason is a reason for the correctness of a conative attitude.[41] However, as we saw in the section concerning Brunero's objection to Lang's proposal, consequentialists need not hold that we have reason to take any attitudes towards valuable things – instead they can maintain that values correspond to reasons for such responses as *binging about*, and *preserving* (and whether we have any reasons for adopting attitudes to different things depends on the consequences of so doing). But perhaps we can extend the Brentano-style analysis so that favouring also includes other kinds of responses than attitudes, thus making *bringing about* and *preserving* count as ways of favouring. This, of course, would be to violate Brentano's notion of correctness, since correctness, as we have seen, is a property of conative attitudes. But let us assume here that we are allowed to perform this move; let us introduce the notion of correctness*, which applies to responses such as bringing about as well as to conative attitudes. A content-reason for a response is then a reason for the correctness* of that response.

But now another problem appears. According to hedonistic act-consequentialism, *bringing about* is a correct* response to pleasure. That is to say, there is a content-reason to bring about pleasure. But there is also, of course, a holding-reason to bring about pleasure; bringing about pleasure is what we, according to this version of consequentialism, ought to do. As Danielsson and Olson point out: 'It is a plausible normative hypothesis that we ought

normally (but not always) to have correct attitudes. In other words, a content-reason for some attitude, *a*, implies a defeasible holding-reason for *a*'.[42] However, since holding-reasons are understood in terms of content-reasons (so that there being a holding-reason to favour *X* means that there is a content-reason to favour *favouring X*), the Brentano-style analysis implies that hedonistic act-consequentialism is committed to the claim that there is a reason to favour *bringing about pleasure*. But what sort of favouring could this possibly be? It cannot be to adopt a pro-attitude, since (again) whether we have reason to adopt any attitude depends on the consequences of so doing (according to the kind of consequentialism that we consider). Rather, it has to be – once again – the response of *bringing about*. So the act-consequentialism that we consider is thus taken to claim that we have reason both to bring about pleasure and to bring it about that we bring about pleasure.[43]

This claim is not implausible by itself. After all, a hedonistic consequentialist reasonably thinks that we have reason to become the kind of persons that bring about pleasure. The problem is that the above line of reasoning recurs: If there really is a content-reason to bring it about that we bring about pleasure, the normative hypothesis implies that we normally ought to bring it about that we bring about pleasure, i.e. that there is a holding-reason to bring it about that we bring about pleasure. But this, in turn, means that there is a content-reason to bring it about that we bring it about that we bring about pleasure. And so it continues. Apart from the facts that we risk to end up with an infinite regress of reasons, and that at some point in that regress these reasons stop to make sense, to claim that a hedonistic consequentialist's view that it is correct to bring about pleasure is to be analysed as the view that there is reason to bring it about that one brings about pleasure would clearly be to misconstrue this version of consequentialism completely.

I take it, then, that the Brentano-style solution has huge difficulties to account for the value ascriptions of many types of consequentialism. Unless it can overcome these difficulties

I believe it should not be accepted as a formal account of value. And I think there is at least another reason to resist it:

As we have seen, the point of departure for the Brentano-style solution is the notion of correctness, understood as a primitive normative notion in terms of which the other normative notions are understood. But there are of course many philosophers – even among the buck-passers – who do not accept Brentano's view on the normative notions and the relation between them (and Danielsson and Olson provide no arguments as to why we should accept this view), and these philosophers will hardly accept the Brentano-style solution either. But even those who accept Brentano's view have reason to resist the Brentano-style solution, or at least to consider it a last resort. The reason is that it lacks the kind of explanatory force that most other suggestions for solution to the WKR problem seek to achieve.[44]

The quite extensive literature on the WKR problem gives at hand that most philosophers who have thought about this problem think there is an explanation to be found in the foundations of our reasons (in the properties that provide them) of what distinguishes reasons of the right kind from reasons of the wrong kind. It seems plausible to assume that the different properties that give rise to reasons of the wrong kind have something in common that explains why these reasons are of the wrong kind. (According to BPV13, for instance, what they have in common is that they involve – in a certain way – the consequences of responding towards their bearers.) Relying, as it does, on the supposedly primitive notion of correctness – correctness being a property of attitudes – the Brentano-style solution evades such an explanation.[45] Strictly speaking, it is not incompatible with there being such an explanation: the same difference in reason-giving properties that explains why some of our reasons are of the wrong kind may also explain why some attitudes that we have reason to have are not correct. But such an explanation would make the Brentano-style solution superfluous. Once we acknowledge that there is a structural difference in the bases of our

reasons that explains the divide between reasons of the right kind and reasons of the wrong kind, we no longer need to appeal to the notion of correctness.

I take it that a version of BPV which convincingly explains the difference between reasons of the right kind and reasons of the wrong kind by reference to structural differences in the foundations of these reasons is preferable to an analysis which evades such an explanation, if only for the fact that there being such an explanation seems intuitively plausible.[46] The version of BPV that I suggest provides such an explanation: A reason of the right kind (or a reason for the correctness of an attitude) is a reason whose presence does not depend on the consequences of the response for which it is a reason.[47]

lars.samuelsson@philos.umu.se

NOTES

[1] T. M. Scanlon, *What We Owe to Each Other* (Cambridge, Mass., 1998), p. 97. Although it was Scanlon who introduced the term 'buck-passing' for this kind of account, he is not its founder. Some writers trace it back to Franz Brentano, and it is often ascribed to A. C. Ewing and Roderick Chisholm (see Sven Danielsson and Jonas Olson, 'Brentano and the Buck-Passers', *Mind* 116 (2007), pp. 511-22, at 511). Notice also that BPV is a purely *formal* account of value – it is supposed to cover any (fairly reasonable) substantive view about what is valuable. Thus the term 'thing' in the quotation from Scanlon should be interpreted widely, so as to cover any type of entity that one could take to be a bearer of value. For some notes on the alleged attractiveness of BPV, see Scanlon, *What We Owe to Each Other*, pp. 97-8, and Gerald Lang, 'The Right Kind of Solution to the Wrong Kind of Reason Problem', *Utilitas* 20 (2008), pp. 472-89, at 472-3.

[2] The phrase was coined by Wlodek Rabinowicz and Toni Rønnow-Rasmussen in 'The Strike of the Demon: On Fitting Pro-attitudes and Value', *Ethics* 114 (2004), pp. 391-423, at 393. I will not discuss any other problems for BPV in this article.

[3] Rabinowicz and Rønnow-Rasmussen, 'The Strike of the Demon', p. 407. This case – which is one of the most frequently discussed WKR cases in the buck-passing literature – is a modification of an example given by Roger Crisp (where the demon wants us to admire a saucer of mud), in his review of Joel Kupperman's *Value…and What Follows*, *Philosophy* 75 (2000), pp. 458-62, at 459. Rønnow-Rasmussen has recently reported that the version provided by him and Rabinowicz was originally suggested to them by Folke Tersman (Rønnow-Rasmussen, *Personal Value* (Oxford, 2011), p. 34n.).

[4] It may be reassuring to point out that we need not invoke evil demons in order to raise the WKR problem. Another example that is sometimes used is hedonism: In order to maximize pleasure we may need to value (adopt valuing attitudes towards) other things than pleasure, things that are not really valuable according to the hedonist (see e.g. Rabinowicz and Rønnow-Rasmussen, 'The Strike of the Demon', p. 403).

[5] Lang, 'The Right Kind of Solution', p. 484. A buck-passing account of negative value (badness) might be formulated in terms of negative attitudes (or negative responses).

[6] Cf. Toni Rønnow-Rasmussen, 'Instrumental Values – Strong and Weak', *Ethical Theory and Moral Practice* 5 (2002), pp. 23-43.

[7] In sect. IV I discuss a special case of genuine instrumental value ('The good demon') that may appear to be problematic for this account, and explain how it can be dealt with.

[8] Jonas Olson, 'The Wrong Kind of Solution to the Wrong Kind of Reason Problem', *Utilitas* 21 (2009), pp. 225-32; John Brunero, 'Consequentialism and the Wrong Kind of Reasons: A Reply to Lang', *Utilitas* 22 (2010), pp. 351-59.

[9] Danielsson and Olson, 'Brentano and the Buck-Passers', p. 513. The solution to the WKR problem that I will eventually suggest is actually quite close to this formulation (the important difference being that I focus on positive responses instead of positive attitudes). Another example of a solution of this kind is found in Philip Stratton-Lake, 'How to Deal with Evil Demons: Comment on Rabinowicz and Rønnow-Rasmussen', *Ethics* 115 (2005), pp. 788-98.

[10] Danielsson and Olson, 'Brentano and the Buck-Passers', p. 513.

[11] Wlodek Rabinowicz and Toni Rønnow-Rasmussen, 'Buck-Passing and the Right Kind of Reasons', *The Philosophical Quarterly* 56 (2006), pp. 114-20.

[12] Rabinowicz and Rønnow-Rasmussen present an argument against a particular version of this approach. In sect. IV I consider a somewhat similar argument in relation to the version that I defend.

[13] Olson, 'The Wrong Kind of Solution', pp. 229-32; Lang, 'The Right Kind of Solution', pp. 482-4.

[14] Olson, 'The Wrong Kind of Solution', pp. 225-32.

[15] Olson, 'The Wrong Kind of Solution', p. 226. One could argue that this first objection of Olson's is due to an unnecessarily narrow reading of Lang's proposal. The last instance of the word 'us' in BPV6 could simply be meant to refer to the members of the group of moral patients (however one wants to conceive of that group), to which the people on the other side of Earth belong.

[16] Olson, 'The Wrong Kind of Solution', p. 227.

[17] Olson, 'The Wrong Kind of Solution', p. 228.

[18] The addition 'genuine' is strictly speaking unnecessary: something is a value if and only if it is a genuine value. I include it only to emphasize that the account is not meant to cover instrumental 'values' in sense (1) discussed in the introduction above (i.e. the sense in which instrumental values are not really values at all, but *merely* means to something else that is valuable).

[19] All things considered it may be that we ought not to respond to a valuable thing *X* in the way corresponding to its value, if, for instance, an evil demon has threatened to punish us if we do, but we still have *reason* to respond to *X* in that way; it is just that that reason has been outweighed by some reason(s) not to respond to *X* in that way – e.g. the reason provided by the demon's threat.

[20] Plausibly, a particular agent's acting successfully in accordance with a rule-consequentialist rule is normally to some extent causally related to bringing about what that version of rule-consequentialism takes to be finally valuable, but exceptions are surely conceivable.

[21] Roger Crisp, 'Goodness and Reasons: Accentuating the Negative', *Mind* 117 (2008), pp. 257-65, at 260.

[22] Crisp, 'Goodness and Reasons', p. 262.

[23] Brunero, 'Consequentialism and the Wrong Kind of Reasons', p. 357.

[24] Scanlon, *What We Owe to Each Other*, p. 97. One may wonder why Lang has chosen to depart from Scanlon's formulation and formulate BPV in terms of attitudes. One reason may be that BPV is often combined (and sometimes perhaps conflated) with the so called 'fitting-attitudes analysis' of value, according to which (roughly) to be valuable is to be a fitting object of a positive attitude (the phrase 'fitting-attitudes analysis' is introduced in Rabinowicz and Rønnow-Rasmussen, 'The Strike of the Demon', p. 391). This kind of analysis is also prevalent among the early buck-passers.

[25] Cf. Crisp, 'Goodness and Reasons', p. 260.

[26] Scanlon, *What We Owe to Each Other*, p. 100.

[27] Scanlon, *What We Owe to Each Other*, p. 98 (see pp. 79-86 for Scanlon's understanding of teleological conceptions of value).

[28] E.g. Scanlon, *What We Owe to Each Other*, p. 95.

[29] Note that *refraining from doing something* may also – in some cases – count as a positive response. Perhaps the appropriate response to a certain valuable thing is to preserve it, where this is best achieved by doing nothing.

[30] Brunero, 'Consequentialism and the Wrong Kind of Reasons', p. 358.

[31] Brunero, 'Consequentialism and the Wrong Kind of Reasons', pp. 356-7.

[32] That is the case in the argument provided by Rabinowicz and Rønnow-Rasmussen ('Buck-Passing and the Right Kind of Reasons', p. 118) mentioned in footnote 12 above. Their argument draws on a person who is lovable because he has the disposition to respond with love to love. Hence that argument does not affect BPV12.

[33] See e.g. Lang, 'The Right Kind of Solution'; Rabinowicz and Rønnow-Rasmussen, 'The Strike of the Demon'; Rabinowicz and Rønnow-Rasmussen, 'Buck-Passing and the Right Kind of Reasons'; Danielsson and Olson, 'Brentano and the Buck-Passers'.

[34] Lang, 'The Right Kind of Solution', pp. 482-4.

[35] Olson, 'The Wrong Kind of Solution', pp. 229-32.

[36] Danielsson and Olson, 'Brentano and the Buck-Passers', pp. 516-17.

[37] Danielsson and Olson, 'Brentano and the Buck-Passers', pp. 514-15.

[38] Danielsson and Olson, 'Brentano and the Buck-Passers', pp. 516-17.

[39] Danielsson and Olson, 'Brentano and the Buck-Passers', pp. 518-19.

[40] For some other concerns, see Rønnow-Rasmussen, *Personal Value*, pp. 40-2.

[41] Danielsson and Olson, 'Brentano and the Buck-Passers', p. 520.

[42] Danielsson and Olson, 'Brentano and the Buck-Passers', p. 515.

[43] Of course, given BPV this would imply that not only is *pleasure* valuable according to this version of consequentialism, but also *bringing about pleasure*. While a hedonistic consequentialist reasonably holds the latter to be instrumentally valuable, this 'value' need not be conceived as a genuine value (see the discussion in section I above), which it would have to be on the Brentano-style BPV. I will not develop this point any further, however.

[44] Cf. Rønnow-Rasmussen, *Personal Value*, p. 39; Lang, 'The Right Kind of Solution', pp. 482-3.

[45] As Danielsson and Olson write: 'the distinction between content-reasons and holding-reasons is not drawn in terms of the properties that provide reasons' (Danielsson and Olson, 'Brentano and the Buck-Passers', p. 515n.)

[46] Note again that also Danielsson and Olson find such an explanation intuitively plausible: 'Intuitively, one is inclined to think that this is the heart of the WKR problem: reasons of the right kind are not provided by the consequences of taking up the relevant attitude' (Danielsson and Olson, 'Brentano and the Buck-Passers', p. 513).

[47] Some of the points raised in this article were presented at the Seventh European Congress of Analytic Philosophy (ECAP7) in Milan, September 2011. I am grateful for the comments I received on that occasion. I am also grateful to the participants in a seminar at my department where I presented an earlier draft of the paper. Special thanks are due to Bertil Strömberg with whom I have discussed this work on several occasions, and whose comments have been particularly valuable.