What's a rational self-torturer to do?

DOUGLAS W. PORTMORE

ABSTRACT: This paper concerns Warren Quinn's famous "The Puzzle of the Self-Torturer." I argue that even if we accept his assumption that practical rationality is purely instrumental such that what the self-torturer ought to do is simply a function of how the relevant options compare to each other in terms of satisfying his actual preferences that doesn't mean that every explanation as to why he shouldn't advance to the next setting must appeal to the idea that so advancing would be suboptimal in terms of the satisfaction of his actual preferences. Rather, we can admit that his advancing would always be optimal, but argue that advancing isn't always what he ought to do given that it sometimes fails to meet some necessary condition for being what he ought to do. For instance, something can be what he ought to do only if it's an option for him. What's more, something can be what he ought to do only if it's something that he can do without responding inappropriately to his reasons—or, so, I'll argue. Thus, the solution to the puzzle lies in realizing that, in certain circumstances, advancing is not what the self-torturer ought to do given that he can do so only by responding inappropriately to his reasons.

Someone may prefer A to Z even though she prefers B to A, C to B, D to C, …, and Z to Y. That is, her preferences may be as depicted in Figure 1, where '$\varphi > \psi$' stands for 'she prefers φ to ψ'. Such preferences are cyclical and, consequently, intransitive. And even though standard decision theory considers them to be irrational, they turn out to be quite common. Take, for instance, the person who plans to quit smoking. For each possible one-last-cigarette that she could smoke, she prefers quitting after smoking it to quitting before smoking it, and, yet, she prefers quitting smoking before having any more cigarettes to never quitting. Or take the person who needs to go on a diet. For each possible one-last-bite of her favorite non-diet food, she prefers going on her diet after having that bite to going on her diet before having that bite, and, yet, she prefers going on her diet without having any more bites of non-diet food to never going on her diet. And it's not just self-interested preferences that can be cyclical. Moral preferences can be as well. Take, for instance, a government authority who is deciding how many citizens to order to curb their carbon emissions. For every positive number *n* that's less than the total number of

Figure 1

citizens, the authority prefers the world in which she orders only n citizens to curb their carbon emissions to the world in which she orders n+1 citizens to curb their carbon emissions. And, yet, she prefers the world in which she orders all citizens to curb their carbon emissions to the world in which she orders none of them to do so.

So, what should an agent with such cyclical preferences do? To figure this out, it would be best to consider an artificial example that abstracts away from any irrelevant real-world complications. Fortunately, Warren Quinn (1990) has constructed just such an example.

> There is a medical device that enables doctors to apply electric current to the body in increments so tiny that the patient cannot feel them. The device has 1001 settings: 0 (off) and 1…1000. Suppose someone (call him the self-torturer) agrees to have the device, in some conveniently portable form, attached to him in return for the following conditions: The device is initially set at 0. At the start of each week he is allowed a period of free experimentation in which he may try out and compare different settings, after which the dial is returned to its previous position. At any other time, he has only two options—to stay put or to advance the dial one setting. But he may advance only one step each week, and he may *never* retreat. *At each advance he gets $10,000*. Since the self-torturer cannot feel any difference in comfort between adjacent settings, he appears to have a clear and repeatable reason to increase the voltage each week. The trouble is that there *are* noticeable differences in comfort between settings that are sufficiently far apart. [Consequently,] …the self-torturer's *step-wise* preferences are [cyclical and] intransitive. All things considered, he prefers 1 to 0, 2 to 1, 3 to 2, etc. . . . but certainly not 1000 to 1. [Moreover,] …his preferences…exhibit various kinds of *indeterminacy*. Not only is there no empirically determinable *first* setting that he disprefers to 0, there is no empirically determinable *first* setting at which these preferences become indeterminate. (1990, pp. 79–82)

As Quinn notes, many theorists condemn the self-torturer's cyclical preferences as irrational, because such preferences make him susceptible to being used as a money pump (see, for instance, Arntzenius & McCarthy 1997, p. 131). But, to insist that the self-torturer get new, more "rational" preferences invites bad faith (Quinn 1990, p. 80). After all, what we want to know is how he should act *given his preferences*, for he may be powerless to change his preferences. Indeed, let's assume so. Furthermore, let's embrace all of Quinn's assumptions. There are five of them in total.

(A1) Practical rationality is purely instrumental such that what a subject ought to do is simply a function of how the relevant options compare in terms of satisfying her actual preferences.[1]

(A2) For any setting n ($0 \leq n < 1{,}000$), the self-torturer has, when at n, the option to proceed to and then stop at $n+1$.[2]

(A3) For any setting n ($0 \leq n < 1{,}000$), the self-torturer prefers the way the world would be if he were to stop at $n+1$ to the way the world would be if he were to stop at n.[3]

(A4) The self-torturer is and will forever remain both practically rational and fully informed such that he will, each week, perform the option that he ought to perform.[4]

(A5) The self-torturer prefers the way the world would be if he were to remain forever at 0 to the way the world would be if he were to proceed to and then stop at 1,000.[5]

The question, then, is: what, according to the instrumental conception of rationality, is such a person to do the first week and each subsequent week? And let's start with the question of whether he should, the first week, proceed from 0 to 1. Unfortunately, there are seemingly plausible arguments for both a 'yes' answer and a 'no' answer. To illustrate, consider the following argument.

(P1) For any setting n ($0 \leq n < 1{,}000$) and any subject S who has agreed to have the self-torture device attached to her in return for the described conditions, S ought, when at n, to proceed to $n+1$ if she has, when at n, both (1) the option to proceed to and then stop at $n+1$ and (2) a preference for the way the world would be if she were to

[1] Quinn says: "I am thinking of rationality (as I have been throughout) as instrumental—as something that is and ought to be the slave of the agent's preferences" (1990, p. 90).

[2] Quinn says of the self-torturer: "No inability stands in his way. It isn't that he lacks the will-power to stop at some reasonable initial goal" (1990, p. 88).

[3] Quinn says: "The self-torturer's step-wise preferences are intransitive. All things considered, he prefers 1 to 0, 2 to 1, 3 to 2, etc. …but certainly not 1000 to 1" (1990, p. 79).

[4] Quinn rejects one proposed solution to the puzzle, because "it cannot work for a self-torturer who assumes that he will always act rationally" (1990, p. 84). I take it, then, that we are to assume that the self-torturer will always act rationally and knows this about himself.

[5] Quinn says of the self-torturer: "if he keeps advancing, he can see that he will eventually reach settings that will be so painful that he would then gladly relinquish his fortune and return to 0" (1990, p. 79).

proceed to and then stop at n+1 over the way the world would be if she were to remain forever at n. [From A1 and *the possibilist view*, which holds that, if such a subject has, when at n, both (1) the option to proceed to and then stop at n+1 and (2) a preference for the way the world would be if she were to proceed to and then stop at n+1 over the way the world would be if she were to remain forever at n, then she ought to proceed to n+1 because proceeding to and then stopping at n+1 is better than her remaining forever at n in terms of the satisfaction of her actual preferences]

(P2) For any setting n ($0 \leq n < 1{,}000$), the self-torturer has, when at n, both (1) the option to proceed to and then stop at n+1 and (2) a preference for the way the world would be if he were to proceed to and then stop at n+1 over the way the world would be if he were to remain forever at n. [From A2 and A3]

(C1) Therefore, for any setting n ($0 \leq n < 1{,}000$), the self-torturer ought, when at n, to proceed to n+1. [From P1–P2]

(C2) Therefore, the self-torturer ought, when at 0, to proceed to 1. [From C1]

(P3) For any setting n, if the self-torturer is at n and ought, when at n, to proceed to n+1, he'll proceed to n+1. [From A4]

(C3) Therefore, when at 0, the self-torturer will proceed to 1. When at 1, he'll proceed to 2. When at 2, he'll proceed to 3. …And, when at 999, he'll proceed to 1,000. [From C1 and P3]

(C4) Therefore, if the self-torturer were to proceed from 0 to 1, he wouldn't stop until reaching 1,000. [From C3]

(P4) The self-torturer prefers the way the world would be if he were to remain forever at 0 to the way the world would be if he were to proceed to and then stop at 1,000. [From A5]

(P5) For any subject S who has agreed to have the self-torture device attached to her in return for the described conditions, if S wouldn't stop until reaching 1,000 if she were to proceed from 0 to 1, and if she prefers the way the world would be if she were to remain forever at 0 to the way the world would be if she were to stop at 1,000, then she ought not, when at 0, to proceed to 1. [From A1 and *the actualist view*, which holds that, if such a subject wouldn't stop until reaching 1,000 if she were to proceed from 0 to 1, and if she prefers the way the world would be if she were to remain forever at 0 to the way the world would be if she were to stop at 1,000, then she ought not to proceed from 0 to 1 because remaining forever at 0 is better than proceeding to and then stopping at 1,000 in terms of the satisfaction of her actual preferences]

(C5) Therefore, the self-torturer ought not, when at 0, to proceed to 1. [From C4, P4, and P5]

(C6) Therefore, P and not-P (a logical contradiction). [From C2 and C5]

As this argument demonstrates, we must reject at least one of P1 and P5 if we are to accept Quinn's assumptions, for this argument entails a logical contradiction and these two premises are the only two that are not simply entailed by his assumptions A1–A5. But which one should we reject? Surprisingly, I think that we should reject both. Given A1, we must assume that practical rationality is purely instrumental such that what a subject ought to do is simply a function of how the relevant options compare to each other in terms of satisfying her actual preferences. But I'll argue that neither the actualist view nor the possibilist view is correct about what we should be comparing when considering whether a subject should proceed from one setting to the next. To illustrate, suppose we're wondering whether a subject attached to the self-torture device should proceed from 0 to 1. According to the possibilist view, we should compare the best world that she *could* actualize if she were to proceed from 0 to 1 to the best world that she *could* actualize if she were to refrain from proceeding from 0 to 1, where which worlds she could actualize if she were to φ depends on what it would be *possible* for her to simultaneously and subsequently do if she were to φ. And the best (and, indeed, the only) world that she could actualize if she were to refrain from proceeding from 0 to 1 is the world in which she remains forever at 0. And since this world isn't as good as the world in which she proceeds to and then stops at 1 (a world that she could actualize if she were to proceed from 0 to 1), the possibilist view implies that she should proceed from 0 to 1.

By contrast, the actualist view holds that we should compare the world that she *would* actualize if she were to proceed from 0 to 1 to the world that she *would* actualize if she were to refrain from proceeding from 0 to 1, where which world she would actualize if she were to φ depends on what she would, *in actual fact*, simultaneously and subsequently do if she were to φ. So, if, in fact, she wouldn't stop until reaching 1,000 if she were to proceed from 0 to 1, then we are, on the actualist view, to compare the world in which she doesn't stop until reaching 1,000

to the world in which she remains forever at 0. And it's stipulated that she prefers the latter. So, the actualist view implies that such a person should not proceed from 0 to 1.[6]

Both are, I believe, incorrect. The possibilist view is, however, the closest to being correct. The only problem with it is that it overlooks the fact that, in some cases, proceeding from n to n+1 is something that the subject could do only by responding inappropriately to her reasons. And, as I'll argue, it can't be that a subject ought to do what she could do only by responding inappropriately to her reasons. Consequently, we must, contrary to the possibilist view, exclude such options from the relevant comparison. What we should be comparing, then, is the best world that she could actualize if she were both to respond appropriately to her reasons and to proceed from n to n+1 to the best world that she could actualize if she were both to respond appropriately to her reasons and to refrain from proceeding from n to n+1. So, the key to solving the self-torturer puzzle is, I believe, to realize that a subject attached to the self-torture device can't be required to advance from one setting to the next if doing so is something that she could do only by responding inappropriately to her reasons.

1. Four Reasons Why φ Might Not Be What a Subject Ought to Do

As I just noted, I believe that the key to solving the self-torturer puzzle is to realize that one reason why an action (such as advancing to the next setting) might not be what the self-torturer ought to do is that it's something that he could do only by responding inappropriately to his reasons. Indeed, there are, in general, at least four reasons why some possible action φ might not be what a given subject ought to do. Perhaps, the most obvious of these is the following.

---

[6] These views are named after the related views from Jackson & Pargetter 1986: "By *Actualism* we will mean the view that the values that should figure in determining which option is the best and so ought to be done out of a set of options are the values of what *would* be the case were the agent to adopt or carry out the option, where what would be the case includes of course what the agent would simultaneously or subsequently in fact do: the (relevant) value of an option is the value of what would in fact be the case were the agent to perform it. We will call the alternative view that it is only necessary to attend to what is possible for the agent, *Possibilism*" (1986, p. 233).

(R1) φ isn't even an option for her.

For instance, the reason it's not the case that I ought to effect world peace is that I don't even have the option of effecting world peace.

Note that I'm using 'option', here, as a technical term. Everyone agrees that it's not just any possible event that can be what a subject ought to do. Take, for instance, the possible event of my walking on water. Or take the possible event of *your* going for a run. Or take the possible event of Halley's comet colliding with Earth. None of these are eligible for the status of being what I ought to do. Halley's comet's colliding with Earth isn't even something that someone can opt for. *Your* going for a run isn't something that *I* can opt for. And even though my walking on water is something that it's possible, in some sense, for me to opt for, the sense of 'possibility' that's relevant in determining whether an event is eligible for being what ought to be done seems to be much narrower than this. For it seems that for some possible event to be something that I ought to perform, it must be that whether I perform it is, in the relevant sense, under my control.[7] So, let me stipulate that φ is an option for a subject if and only if whether she performs it is, in the relevant sense, under her control—the relevant sense being the one that guarantees that 'ought to φ' implies 'has the option to φ'. So, even if it's open to debate whether 'ought' implies 'can', there can be no doubt that 'ought' implies 'option'. Thus, one obvious reason for a possible action's failing to be what ought to be done is that it's not even an option.

But, of course, not all options ought to be performed. So, there must be other reasons why some possible event, φ, might not be what a subject ought to do. Here's another.

(R2) Although φ is an option for her, it's not her best option.

[7] For some responses to worries about such a control condition both on the basis of the thought that there's resultant moral luck and on the basis of Frankfurt's putative counterexamples involving counterfactual interveners, see Portmore Forthcoming.

This is, for instance, the reason why burning all my cash is not something I ought to do. Although I have the option of burning all my cash, there are better things that I could do with my cash. Thus, it's not my best option and, consequently, not something I ought to do. It seems, then, that I ought to φ only if φ is my best option.

R1 and R2 should, I believe, be fairly uncontroversial. But I'll argue that there are at least two other—admittedly, more controversial—reasons why a possible event, φ, might not be what a subject ought to do. The first of these is the following.

(R3) Although φ is her best option, her φ-ing is incompatible with her performing an even better option.

Of course, you may wonder how an option could be her best option and yet be incompatible with her performing an even better option. This is possible because what makes an option best is that it is better than every *alternative* option, not that it is better than every *distinct* option. Let me explain.

Two options φ and ψ are distinct if and only if it's not the case that each entails the other, where one option entails another if and only if doing the one without doing the other is a non-option. For instance, running and running fast are distinct options, because although there's the option of running without running fast, there isn't the option of running fast without running. That is, although running doesn't entail running fast, running fast does entail running. And since it's not the case that each entails the other, the two are distinct. But, of course, even if distinct, they are not alternatives. For two options φ and ψ are alternatives if and only if performing both φ and ψ is a non-option. Thus, running is an alternative to walking, because both running and walking is a non-option. But running is not an alternative to running fast given that both running and running fast is an option.

Given that they're distinct, running fast can be better than running even though they're not alternatives. To illustrate, suppose, for the sake of argument, that an option φ is better than an option ψ if and only if the possible world that *would* be actualized if one were to φ is better

than the possible world that *would* be actualized if one were to ψ—where, again, the possible world that *would* be actualized if one were to *x* depends on what one would, in actual fact, simultaneously and subsequently do if one were to *x*. And let's imagine that I'm in the vicinity of a not-very-fast predatory animal whose killer instinct is triggered only by the sight of its prey running. Thus, assume that the situation is as follows. If I refrain from running, I'll be slightly mauled but live after the animal loses interest. If I run but don't run fast, the predatory animal will catch me and kill me. If I run fast, I will outrun the animal and escape unscathed. Now, assume that my intent is to commit suicide by running slow. Thus, as a matter of fact, if I were to run, the possible world that would be actualized is the one in which I run slow and am both caught and killed by the animal.

So, refraining from running is my best option. For φ counts as my best option if and only if it is better than every alternative option.[8] And refraining from running is better than every alternative option. After all, the *only* alternative to my refraining from running is my running. And whereas my running would, it's stipulated, result in my being killed. My refraining from running would, it's stipulated, result in my being only slightly mauled. And being slightly mauled is, we're assuming, better than being killed. So, refraining from running is, indeed, my best option—that is, an option that's better than every alternative.

Nevertheless, besides the options of running and refraining from running, I have several distinct options, including running fast. And this option is better than both the distinct option of running and the distinct option of refraining from running. Thus, it seems that of all my distinct options (that is: running, refraining from running, running fast, refraining from running fast, running non-fast, refraining from running non-fast, etc.), the one that I have most reason to perform—and, thus, the one that I ought to perform—is the option of running fast. Moreover,

[8] This comports with ordinary usage. To illustrate, note that we would use the phrase 'my best option' to describe my watching football on TV this Sunday morning if and only if we thought that there was no alternative (such as my watching something other than football on TV this Sunday morning or my doing something other than watching TV this Sunday morning) that was at least as good. Thus, we wouldn't deny that this was my best option just because we thought that I had an even better distinct option, such as the option of watching football on TV both this Sunday morning and next Sunday morning.

since my running fast is incompatible with my refraining from running such that I don't have the option of both running fast and refraining from running, it is, I believe, a mistake to think that, additionally, I ought to refrain from running. After all, as we've just established, I ought to run fast and I don't have the option of both running fast and refraining from running. So, if we were, in addition, to claim that I ought to refrain from running, we would have to deny what's known as *joint satisfiability:* if a subject both ought to φ (e.g., run fast) and ought to ψ (e.g., refrain from running), then she has the option of both φ-ing and ψ-ing (e.g., both running fast and refraining from running).[9] And although joint satisfiability is not uncontroversial (see, e.g., White 2017a), we should accept it given the following argument.[10]

(P6) It is not the case that <if a subject both ought to φ and ought to ψ, then she has the option of both φ-ing and ψ-ing>. In other words, there is a subject who both ought to φ and ought to ψ but doesn't have the option of both φ-ing and ψ-ing. [Assumption for *reductio*]

(C8) Thus, if this subject believes what's true, she'll believe that she doesn't have the option of both φ-ing and ψ-ing. [From P6]

(P7) For all actions *x* and all subjects S, if S ought to do *x*, then she ought to intend to do something that entails her doing *x*. [From the fact that practical oughts are normative for intentions—more on this below]

(C9) Thus, if this subject intends to do all that she ought to intend to do, she will both intend to do something that entails her φ-ing and intend to do something that entails her ψ-ing. [From P6 and P7]

(C10) Thus, if this subject believes what's true and intends to do all that she ought to intend to do, then, as a result, she'll believe that she doesn't have the option of both φ-ing and ψ-ing while both intending to do something that entails her φ-ing and intending to do something that entails her ψ-ing. [From C8 and C9]

(P8) It's irrational for a subject to believe that she doesn't have the option of both φ-ing and ψ-ing while both intending to do something that entails her φ-ing and

[9] See, for instance, Kiesewetter Forthcoming.

[10] This argument is, in part, inspired by one found in Kiesewetter Forthcoming, which is a response to White 2017a.

intending to do something that entails her ψ-ing. [From the consistency requirement on beliefs and intentions[11]]

(C11) Thus, if this subject believes what's true and intends to do all that she ought to intend to do, then, as a result, she'll be irrational. [From C10 and P8]

(P9) It's not the case that <if this subject believes what's true and intends to do all that she ought to intend to do, then, as a result, she'll be irrational>. For it's implausible to suppose that a subject could be irrational as a result of her believing what's true and intending to do all that she ought to intend to do. [Assumption]

(C12) Therefore, if a subject both ought to φ and ought to ψ, then she has the option of both φ-ing and ψ-ing. [From P6–P9 by *reductio*]

The two most controversial premises are, I believe, P7 and P8. I'll take each in turn. P7 is, I believe, intuitively plausible. If you ought to perform an action, then, given that it couldn't be that you ought to perform it unintentionally, it must be that you ought to form the intention to perform it—or, at least, the intention to perform something that entails it. Nevertheless, there may seem to be counterexamples to P7—the view that practical oughts are normative for intentions. For one, it may seem that it could be that I ought to act spontaneously even if it's not the case that I ought to form the intention to do so given that such an intention would only be self-defeating. But we should deny that I ought to act spontaneously if this isn't something that I can do intentionally. We should hold, instead, only that I ought to perform (intentionally) some distinct act that would increase my chances of acting spontaneously—e.g., the act of consuming a few stiff drinks. For another, it may seem that I ought to refrain from torturing

[11] According to this requirement, a subject must be such that the set consisting of the propositional contents of all her beliefs and all her intentions is logically consistent, where the propositional content of one's belief that *p* is '*p*' and the propositional content of one's intention to φ is 'one will φ'. Thus, the propositional content of one's intention to do something that entails one's φ-ing entails the proposition 'one will φ'. Proponents of this requirement include Michael Bratman (1987 & 2009), Jacob Ross (2009), and Ralph Wedgwood (2007, p. 109). And note that accepting this requirement doesn't entail accepting the metaphysical view that intending to φ involves, or consists in, believing that one will φ. Rather, it entails accepting only the normative view that intending to φ rationally requires one to believe that one will φ.

children even though it's not the case that I ought to form the intention to refrain from torturing children. After all, the thought of torturing children should be unthinkable. So, if I need to form the intention to refrain from torturing children to prevent myself from doing so, then something has gone terribly wrong. But although I admit that something has gone wrong if I need to form this intention, it still seems that I should form some intention that entails my refraining from torturing children—perhaps, the general intention to refrain from hurting others. In any case, if I do find myself tempted to torture children, I should certainly resolve (and thereby intend) to refrain from doing so.[12] So, neither of these putative counterexamples turn out to be persuasive upon reflection.

P8 is also intuitively compelling, but, like P7, there may seem to be some putative counterexamples to it. One putative counterexample is the preface paradox. Imagine that a non-fiction writer apologizes in the preface of her book for the extremely likely fact (given her inductive evidence) that at least one of the other claims in the book is false, and false despite her best efforts to carefully research each one. So, either she ought not to believe all these other claims or the consistency requirement (that is, P8) is false. And it may seem that, since it's subjectively rational (that is, rational in the evidence-relative sense) for her to believe all these other claims given how carefully she's researched each of them, we must reject the consistency requirement.

But this is, in fact, no counterexample to the consistency requirement. For although it may be subjectively rational for her to believe all these other claims, she ought not to believe all of them—at least, not in the objective (fact-relative) sense of 'ought' that I'm employing throughout this paper. For we're assuming that some of these other claims are false, and one (objectively) ought not to believe a false claim. So, although she's right to believe that not all the other claims in the book are true (assuming that this is so), she's wrong to believe every single one of them—specifically, she's wrong to believe the false ones.

[12] A resolution to φ consists in both a first-order intention to φ and a second-order intention not to let that first-order intention be deflected by the anticipated temptation not to φ—see Holton 2009, pp. 11–12.

Nevertheless, there are other putative counterexamples to the consistency requirement. Consider, for instance, the Cathedral Paradox.

> Susan is planning her trip to Europe. There are 20 cathedrals she would like to visit. Each one has a fee. She really would like to see each one…. [But] she can only afford 19 cathedrals…. Let the cathedrals number 1 through 20. And let φ_n be the action of visiting cathedral n. Susan intends φ_1, intends φ_2, ..., and intends φ_{20}. However, Susan knows that it is impossible for her to perform the conjunctive action φ_1, ..., and φ_{20}. She simply doesn't have the cash. And so Susan plans to skip at least one cathedral, intending the action: not φ_1 or ... or not φ_{20}. (Goldstein 2016, p. 2)

Either Susan ought not to have all these intentions or the consistency requirement is false. And since there seems to be nothing subjectively irrational about her having all these intentions, it may seem that we must reject the consistency requirement. But this too is no counterexample to the consistency requirement. For although it is subjectively rational for her to have all these intentions, she (objectively) ought not have every single one of them. After all, she will not, in fact, visit all twenty cathedrals. And, so, she objectively ought not to intend to visit whichever cathedral it is that she'll skip.

But what if we suppose that there is no fact of matter as to which cathedral she'll skip—perhaps, which cathedral she'll skip depends on some quantum indeterminacy. Suppose, then, that there is, for each cathedral, a 1-in-n objective chance that she'll have to skip it.[13] In that case, it may seem that given the low objective probability that she'll have to skip any given cathedral, she should intend to visit each one—and if you deny that the objective probability that she'll skip a given cathedral is low enough to make her intending to visit it rational, then just change the example so that n equals, say, twenty billion or more. It would, then, certainly be low enough. And, perhaps, doing this makes for a more plausible counterexample. But I doubt it. In this case, we have to acknowledge that it's not entirely up to Susan which cathedral she skips or which cathedrals she visits. Consequently, we should deny that she objectively ought to intend to visit each one. Instead, we should hold only that she objectively ought, for each cathedral, to

[13] The objective probability that some event will (or would) occur is the percentage of the time that it will (or would) occur under identical causal circumstances—circumstances where the causal laws and histories are exactly the same.

intend only to *try* to visit it. For what's up to her is not whether she visits it but only whether she tries to visit it.

So, if all this is correct, then one reason why φ (e.g., my refraining from running) might not be what I ought to do is that it's incompatible with my performing some even better option ψ (e.g., my running fast). For it may be, as in the above case, that I ought to perform this even better option ψ. And given this, joint satisfiability, and the fact that I don't have the option of both φ-ing and ψ-ing, it follows that it can't be that I ought to φ. Thus, one reason for thinking that φ is not something that a subject ought to do is that R3 obtains. And, so, it seems that we should accept R3.[14]

There is, I think, a fourth possible reason why φ might not be what ought to be done. It's this.

(R4) Although φ is her best option, the only way for her to φ is by responding inappropriately to her reasons.

Here are some examples. It seems that I can infer causation on the basis of correlation. After all, this is a fairly common fallacy and not one to which I'm immune. Nevertheless, it seems that I should not form the belief that taking a daily multivitamin will cause me to live longer on the basis of my awareness of studies demonstrating a correlation between taking a daily multivitamin and living longer. Yet, forming this belief could be my best option, for suppose that this belief is both true and beneficial for me to possess. Nevertheless, I don't see how it could be that I ought to form this belief if I could do so only by responding inappropriately to my reasons—that is, by inferring causation from correlation. To take another example, consider wishful thinking. This too is a fallacy to which I'm not completely immune. Nevertheless, it can't be that I ought to form the belief that I have an immortal soul on the basis of wishful

[14] For a further defense of R3 see Portmore Forthcoming.

thinking given that doing so would necessarily involve my responding inappropriately to my reasons. Or, so, I'll now argue.

What makes me someone to whom oughts, requirements, and responsibilities apply is that I have the capacity for responding appropriately to reasons. Indeed, it's this capacity for responding appropriately to reasons, which normal adult humans typically possess and which primitive animals, very young children, and the severely mentally impaired typically lack, that distinguishes those beings who can have obligations and responsibilities from those beings who can't. But, given this, it would be nonsensical for me to be required to respond inappropriately to my reasons. For such a requirement would apply to me in virtue of my having the capacity to respond appropriately to reasons even though it's only by failing to fully exercise this capacity that I could come to fulfill it. And that's just implausible. For it's implausible to suppose that the very capacity in virtue of which I'm responsible for my failures is the one that, when exercised fully and flawlessly, leads me to fail. And this holds whether we're talking about my failure to do what's required of me or my failure to do what I ought to do. It just can't be that what makes me responsible for my failures is the very thing that causes me to fail. And, so, it can't be that I ought to do something that I could do only by responding inappropriately to my reasons. Thus, we should accept R4.[15]

Now, if I'm right about R1–R4, then Quinn's assumption that rationality is purely instrumental (viz., A1) allows that not every explanation as to why the self-torturer ought not to advance must appeal to the fact that so advancing would be worse (or, at least, not as good) in terms of fulfilling his actual preferences. For, given R1, it could instead appeal to the fact that advancing isn't an option for him. Of course, Quinn assumes that the self-torturer always has the option of advancing. So, we won't be able to appeal to R1 in explaining why the self-torturer's advancing is not always what he ought to do. But, as I've argued above, there are other reasons why an act might not be one that the self-torturer ought to perform that don't

[15] See Portmore Forthcoming (chap. 5) for an argument that we must accept R4 in order to accommodate the idea that a moral theory ought to be morally harmonious—that is, ought to be such that the agents who satisfy it, whoever and however numerous they may be, are guaranteed to produce the morally best world that they could, in the relevant sense, together produce.

appeal to any facts about what would best fulfill his actual preferences. For, as R4 allows, it could be that the self-torturer ought not to advance because this is something that he could do only be responding inappropriately to his reasons. Indeed, I'll argue below that this is sometimes the case.

2. Why We Should Reject P1

Recall that P1 says: For any setting n ($0 \leq n < 1{,}000$) and any subject S who has agreed to have the self-torture device attached to her in return for the described conditions, S ought, when at n, to proceed to n+1 if she has, when at n, both (1) the option to proceed to and then stop at n+1 and (2) a preference for the way the world would be if she were to proceed to and then stop at n+1 over the way the world would be if she were to remain forever at n. This premise is false. To see why, imagine that that a subject named Faizal is now at 50, having resolved not to go beyond 50. Thus, he intends to resist the anticipated temptation to proceed to 51 in light of his having both (1) the option to proceed to and then stop at 51 and (2) a preference for the way the world would be if she were to proceed to and then stop at 51 over the way the world would be if she were to remain forever at 50.

Now, P1 implies that Faizal ought to proceed to 51. But, as we'll soon see, proceeding to 51 couldn't be something that Faizal ought to do. To see why, consider that, given that he's resolved not to go beyond 50, there are only two ways for Faizal to proceed from 50 to 51. One way is for him to do so unintentionally by accidentally flicking the dial from 50 to 51. Call this *The Accidental Procession to 51*. The other way, of course, is for him to do so intentionally by reconsidering his previous intention not to go beyond 50, thereby changing his mind and deciding to proceed to 51 by intentionally advancing the dial from 50 to 51. Call this *The Reconsidered Procession to 51*. And let's assume that, in both instances, Faizal will remain forever at 51 and not proceed any further. Also, let's assume that 50 and 51 are both reasonable stopping points, where a setting counts as a reasonable stopping point only if the given subject

(determinately) prefers the world in which she stops at it both to the world in which she remains forever at 0 and to the world in which she proceeds to and then stops at 1,000.[16]

Now, I concede that the self-torturer prefers both the world in which *The Accidental Procession to 51* occurs and the world in which *The Reconsidered Procession to 51* occurs to the world in which he remains forever at 50. And, so, I concede that both these worlds are better than the world in which he remains forever at 50 in terms of satisfying of his actual preferences. Thus, if he has the option of taking a pill that would cause either *The Accidental Procession to 51* or *The Reconsidered Procession to 51* to occur, he should certainly take that pill. Nevertheless, I deny that he ought to proceed from 50 to 51. The thought that he ought to proceed from 50 to 51 just doesn't follow from the above. And this is so even if we assume, as Quinn does, that rationality is purely instrumental. For, as I've shown, even if rationality is purely instrumental, there can still be reasons why an act isn't the one that ought to be performed that have nothing to do with how good or bad that act is. Indeed, I've argued that there are at least three such reasons: R1, R3, and R4. Moreover, as I've just shown, there are only two ways for Faizal to proceed from 50 to 51: *The Accidental Procession to 51* and *The Reconsidered Procession to 51*. And neither can be what he ought to do given R1 and R4. For, as I'll show, *The Accidental Procession to 51* isn't an option for him and, so, can't be what he ought to do given R1. And although *The Reconsidered Procession to 51* is an option for him, it is, I'll show, one that he can perform only by responding inappropriately to his reasons, and, so, it can't be what he ought to do given R4. Thus, proceeding to 51 is not something Faizal ought to do, contrary to what P1 implies.

Recall that an option for a subject is something she controls whether or not she performs. But Faizal can't control whether he accidentally flicks the dial from 50 to 51. Such an accidental act may count as one that he can perform, but it doesn't count as an option for him. And, thus, it can't be something that he ought to do. To illustrate, consider that although accidentally knocking over a glass of wine is something that I can do (indeed, it's something that I have done

[16] Given Quinn's assumptions, there is no perfect stopping point, for every possible stopping point is either inferior to the next or inferior to 0 in terms of the satisfaction of the subject's actual preferences. Also, note that I remain neutral as to whether the above necessary condition is also a sufficient condition.

several times), it's not (and never has been) an option for me. For I've never had the relevant sort of control over whether I was to accidentally knock over a glass of wine—that is, I've never had the sort of control that would make it eligible for being something that I ought to do. Likewise, Faizal doesn't have the relevant sort of control over whether he accidentally flicks the dial. So, *The Accidental Procession to 51* is not an option for Faizal. And, since it's not an option, it can't, given R1, be something that he ought to do.

The only other way for Faizal to proceed from 50 to 51 is to do so intentionally by reconsidering his previous resolution not to go beyond 50. But the whole point of his having formed this resolution in the first place was to prevent the anticipated inclination to proceed to 51 from causing him to be unsuccessful in following through with his plan to stop at 50. And given that the whole function of a resolution is to prevent such anticipated inclinations from undermining one's ability to successfully follow through with one's previous plans, it is, as Richard Holton has argued, contrary to reason to revise such "a contrary inclination defeating intention (a resolution) in response to the presence of those very inclinations" (2009, p. 78). So, given that Faizal resolved not to go beyond 50 and, so, intended to resist the anticipated inclination to proceed to 51, he cannot proceed to 51 in response to that inclination except by responding inappropriately to the decisive reason he has for not reconsidering his resolution in response to that very inclination. And, since as R4 states, no act that can be performed only by responding inappropriately to one's reasons can be something that one ought to do, it can't be that Faizal ought to proceed to 50 via *The Reconsidered Procession to 51*.

Here, I'm relying on the following principle (call it *the no-reconsideration principle*): A subject has decisive reason not to reconsider a resolution in response to the very inclinations that it was intended to prevent her from reconsidering in response to—unless, of course, something relevant has changed between now and then, such as how she ranks her various ends. This principle implies that Faizal has decisive reason not to reconsider his resolution to stop at 50 in response to his inclination to get an additional $10,000. It implies this both because this resolution was intended to prevent him from reconsidering in response to this very inclination and because nothing relevant has changed. For instance, he still ranks the end of

ensuring that he doesn't end up at some unreasonable stopping point above the end of getting an additional $10,000.

Since this principle may not be entirely uncontroversial, I should explain my reasons for accepting it. My main reason for accepting it is that it's intuitive in the same way that it's intuitive to think that a subject has decisive reason not to intend to do what there is no good reason for her to do. Thus, the main reason to accept it has nothing to do with either the thought that good consequences will come from being generally disposed to abide it or the thought that a subject will likely regret how things turn out if she doesn't abide by it. Rather, the main reason to accept it is simply that it's intuitive. This and the fact that it has plausible implications seems to be sufficient reason for accepting it.

With respect to its plausible implications, note that it doesn't imply that resolutions should never be reconsidered, but only that they should never be reconsidered in response to the very inclinations that they were intended to prevent from resulting in reconsideration—and only if nothing relevant has changed since the subject originally formed the resolution. Also, note that this principle doesn't have the same problematic implications that a similar principle has. That principle (call it *the problematic principle*) is as follows: "other things being equal, one should not reconsider a resolution, since this defeats the point of having formed the resolution" (2014, p. 283). As Chrisoula Andreou explains, this principle is problematic.

> Suppose A wants B to go to a yoga retreat during spring break, but A also knows that B will want to use the time to work on her dissertation. Suppose further that A goes ahead and makes all the necessary reservations. But B, finding herself inclined to work on her dissertation, insists that the reservations be cancelled. The question then arises as to whether B is being irrational. …It is proposed [on the problematic principle] that B can be charged with being irrational because cancelling the reservations would defeat the purpose of A's having made them, which is for B to go to a yoga retreat during spring break. [Yet, this]…charge of irrationality seems particularly difficult to defend if it is granted that, were B to consider the matter carefully rather than just going along with A's plan, we can expect that rational deliberation would prompt B to side with her inclination and rank working on her dissertation above going to the yoga retreat. (2014, p. 284)

Now, Andreou believes that, even if we take A and B to be time-slices of the same person, we should not accept, as the problematic principle implies, that B should not reconsider

her resolution just because her doing so would defeat the point A had in forming the resolution. And I agree. But note that my no-reconsideration principle doesn't have this implausible implication, for something relevant has changed since A made her resolution. For whereas her earlier self (viz., A) ranked the end of B's going to the yoga retreat above B's working on her dissertation, her current self (viz., B) has the opposite ranking.

Now, if we want the no-reconsideration principle to imply that B should not reconsider A's resolution, we'll need to modify the case. Let's assume, then, that her earlier self (viz., A) made the resolution with the intention of preventing her inclination to spend her spring break working on her dissertation from causing her later self (viz., B) to reconsider her resolution to go to the yoga retreat during spring break—a resolution that she made with, say, the purpose of ensuring that she doesn't get burnt out working too much on her dissertation. And let's assume that both her earlier and later selves rank the end of ensuring that she doesn't get burnt out working too much on her dissertation above the end of spending her spring break working on her dissertation. Now, if we make these revisions, we find that the no-reconsideration principle now implies that B ought not to reconsider A's resolution to go to the yoga retreat in response to the inclination that she has to spend her spring break working on the dissertation. But, in this case, the implication seems entirely plausible, as does the principle's implications in the original case. So, it seems that the no-reconsideration principle has plausible implications in a range of cases.

To sum up, then, we must reject P1 given that it implies that Faisal ought to do something (viz., proceed from 50 to 51) that he can do only either accidentally or by responding inappropriately to his reasons. For, given R1 and R4, neither can be something that he ought to do.

3. Why We Should Reject P5

We should also reject P5. Recall that P5 says: For any subject S who has agreed to have the self-torture device attached to her in return for the described conditions, if S wouldn't stop until reaching 1,000 if she were to proceed from 0 to 1, and if she prefers the way the world would be

if she were to stop at 0 to the way the world would be if she were to stop at 1,000, then she ought not, when at 0, to proceed to 1. To see that this premise is false, imagine that there is an irrational subject named Erasmus who (1) has agreed to have this device attached to him in return for the described conditions, (2) possesses all the preferences that the original self-torturer possesses, and (3) would end up proceeding all the way to 1,000 if he were to proceed from 0 to 1. Now, P5 implies that Erasmus shouldn't proceed from 0 to 1. But, as I'll now show, this is false. And, so, we must reject P5.

Let's assume both that Erasmus is now at 0 and that 50 would, for some reason, be the only reasonable stopping point for him. Let's also assume that Erasmus (1) has the option of proceeding to and then stopping at 50, (2) would proceed from 0 to 50 and then stop there so long as he now resolves to stop there, (3) would now resolve to stop at 50 if he were now to respond appropriately to his reasons, and (4) prefers the way the world would be if he were to proceed from 0 to 50 and then stop there to the way the world be if he were to just remain forever at 0. Given 1–4, it seems clear that Erasmus ought to proceed from 0 to 50 while resolving to stop there. Yet, even so, P5 implies that Erasmus ought not to proceed from 0 to 1 given that he's not going to respond appropriately to his reasons and, so, is not going to resolve to stop at 50. Consequently, he would, as a matter of fact, proceed all the way to 1,000 if he were to proceed from 0 to 1.

So, the proponent of P5 must deny at least one of the following two highly plausible claims: (claim_1) Erasmus ought to proceed from 0 to 50 while resolving to stop there or (claim_2) joint satisfiability, which, you'll recall, holds that if a subject both ought to φ and ought to ψ, then she has the option of both φ-ing and ψ-ing. For if she accepts claim_1, she must deny claim_2. After all, Erasmus doesn't have the option of both (1) not proceeding from 0 to 1 and (2) proceeding from 0 to 50 while resolving to stop there. So, the proponent of P5 must reject joint satisfiability if she accepts claim_1. But I don't see how anyone could reject claim_1. Clearly, of all the distinct options that Erasmus has, this is the very best one. Moreover, I've argued that we must accept joint satisfiability. So, since P5 implies that it can't be that both claim_1 and claim_2 are true, we should reject P5 rather than one of these two highly plausible claims.

4. The Compromise View and Why Its Superior to Its Rivals

We should, I believe, replace P1 and P5 with the following conjunctive view: (conjunct$_1$) For any reasonable stopping point RSP and any subject S who has agreed to have the self-torture device attached to her in return for the described conditions, S is required to select, when necessary, some particular RSP (call it P-RSP) and to resolve to stop there.[17] And, (conjunct$_2$) for any setting n ($0 \leq n < 1{,}000$), she ought to proceed from n to n+1 if and only if she has already complied with the requirement stated in conjunct$_1$ and can intentionally proceed from n to n+1 without violating the no-reconsideration principle. I will call this *the compromise view,* because it's a compromise between the possibilist view and the actualist view—that is, the views that give rise to P1 and P5, respectively. Like the possibilist view, the compromise view holds that such a subject should typically proceed from n to n+1 if she has, when at n, both (1) the option to proceed to and then stop at n+1 and (2) a preference for the way the world would be if she were to proceed to and then stop at n+1 over the way the world would be if she were to remain forever at n. But, unlike the possibilist view, it makes an exception when doing so intentionally would require reconsidering a previous resolution not to so proceed in response to the very inclinations it was intended to overcome. And, like the actualist view, it requires her to respond to her situation in a way that ensures that she won't proceed to 1,000 or any other unreasonable stopping point. But, unlike the actualist view, it doesn't do this by forbidding her from proceeding from n to n+1 if, in fact, doing this would lead to her proceeding to some unreasonable stopping point. Instead, it does this by requiring her to form, when necessary, a resolution that will ensure that she doesn't proceed to any unreasonable stopping point. Thus, in the above example, Erasmus ought to advance from 0 to 1 even though he would, as a matter of fact, advance all the way to 1,000 if he were to do so. But, on the compromise view, he's required to form, when necessary, a resolution that will prevent him from advancing to any

[17] It becomes necessary to select some particular reasonable stopping point and to resolve to stop there when proceeding to the next setting without doing so risks one's proceeding beyond any reasonable stopping point. And note that I take no stand on whether there is any non-arbitrarily way to select a particular reasonable stopping point.

such unreasonable stopping point. Thus, on the compromise view, the mere fact that he's not going to form this resolution doesn't imply that it's impermissible for him to advance from 0 to 1. It's just implies that it's impermissible for him to do so without forming, when necessary, the resolution that will prevent him from proceeding to any unreasonable stopping point. After all, he's perfectly capable of advancing from 0 to 1 while resolving to stop at some reasonable stopping point, and doing this will ensure that he doesn't proceed to any unreasonable stopping point.

The compromise view is, I believe, intuitively plausible. It holds that any subject who has agreed to have the self-torture device attached to her in return for the described conditions should proceed each week until reaching the particular reasonable stopping point that she has resolved not to go beyond. And it holds that she should ensure that she doesn't proceed to any unreasonable stopping point by forming, when necessary, a resolution to stop at some particular reasonable stopping point. Thus, it prohibits her from proceeding to 1,000 or to any other unreasonable stopping point. Indeed, it prohibits her from proceeding to even some reasonable stopping point unless she either has already formed a resolution that will prevent her from proceeding to an unreasonable stopping point or can do so without the risk of ending up at some unreasonable stopping point.

Of course, many will wonder why we should accept the compromise view's implication that such a subject shouldn't proceed beyond the particular reasonable stopping point that she's resolved to stop at. For instance, many will wonder why Faizal—who has resolved to stop at 50—shouldn't proceed from 50 to 51 given that he has both (1) the option to proceed to and then stop at 51 and (2) a preference for the way the world would be if she were to proceed to and then stop at 51 over the way the world would be if she were to remain forever at 50. For, as some philosophers have pointed out, this implication "stands in need of defense" (Tenenbaum & Raffman 2012, p. 108) given that we're assuming that what Faizal ought to do is simply a function of how the relevant options compare in terms of satisfying his actual preferences. But we can provide the required defense by appealing to the fact that we shouldn't be comparing the option of stopping at 50 with the option of reneging on his resolution and proceeding to 51.

We shouldn't be comparing these two, because the latter isn't a relevant option. For I've argued that only those actions that can be performed intentionally without responding inappropriately to one's reasons count are relevant in making such a comparison. So, although his stopping at 51 is better than his stopping at 50 in terms of satisfying his actual preferences, his proceeding to 51 can't be what he ought to do given that it's something he can do intentionally only by responding inappropriately to the decisive reason that he has to refrain from reconsidering his resolution not to proceed beyond 50 in response to the very inclinations that it was intended to overcome. Thus, I deny Stephen White's claim that "when the time comes for one to carry out a prior plan, if it's obvious that one's interests would be better served by revising that plan, then that's what one should do" (2015, p. 5).[18] Indeed, this is clearly false. For one, changing plans can't be what one ought to do if one doesn't even have the option of changing plans—see R1. And, for another, changing plans can't be what one ought to do if one can do so only by responding inappropriately to one's reasons—see R4. Thus, I think that we can meet the demand for a defense of the above implication by appealing to R4 and the fact that, given Faizal's resolution to stop at 50, he can change plans only by responding inappropriately to the decisive reason he has to refrain from reconsidering his previous plans.

Another merit of the compromise view is that it conforms to what Sergio Tenenbaum and Diana Raffman (2012) call *non-segmentation*. Non-segmentation is a claim about the following sort of one-off case. Suppose that a subject is offered only a single choice: either (A) have the self-torturer device set permanently to n and receive $n \times \$10{,}000$ or (B) have the devise set permanently to n+1 and receive \$10,000 in addition to that sum (that is, $\$10{,}000 + [n \times \$10{,}000]$). Non-segmentation is the claim that for no setting n ($0 \leq n < 1{,}000$) would it be irrational for such a subject to choose B. Now, the compromise view conforms to this claim, because the only reason that it would be irrational on this view for a subject to act so as to end up at n+1 instead of n is that she could do so only by reneging on some previous resolution not

[18] Michael Huemer makes a similar assumption: "If one has a choice between A and B, and one rationally prefers A to B, it is rational to choose A" (2018, p. 94). Yet, it's clearly irrational to respond inappropriately to one's reasons, and sometimes one can choose the preferred option only by responding inappropriately to one's reasons.

to do so. But since conjunct$_1$ of the compromise view doesn't require a subject facing such a single choice to form any resolution, choosing B will never involve reneging on a resolution that she was required to form.

The compromise view also respects Stephen White's generality constraint. According to this constraint, a satisfactory solution to the self-torturer puzzle must "explain why going all the way to 1000 is irrational in all cases where the self-torturer has the relevant preferences and is fully informed about the relevant facts" (2017b, p. 588). And, in particular, such a solution must be able to explain what has gone wrong in the following sort of case. Imagine that a man named Moros agrees to have the self-torture device attached to him in return for the described conditions and sets off advancing each week without ever coming up with any plan about how to proceed. Assume that "he just figures he'll stop advancing the dial at some point before the pain gets too bad. Suppose he's wrong about this, though. Every week he decides to take the money and he finally ends up at the last setting, in horrible pain and wishing he'd never agreed to play this twisted game" (2017b, p. 589). As White points out, Moros hasn't violated any principle of rational intention-revision—such as my no-reconsideration principle. And, given this, many plan-based views—that is, views that hold that "the self-torturer should (a) adopt a reasonable plan at the outset about when to stop, and (b) stick to that plan" (2017b, p. 586)—will be at a loss to explain which week he went wrong in advancing. Yet, clearly, there must have been some week where shouldn't have advanced if he ends up wishing he had never agreed to play the game in the first place. Fortunately, the compromise view can account for this. For, according to conjunct$_2$, Moros ought not to have proceeded any of those weeks without having first complied with the requirement stated in conjunct$_1$—the requirement to form any necessary resolutions. So, he went wrong in proceeding each of those weeks in which he advanced without having first complied with the requirement stated in conjunct$_1$.

The compromise view is not only intuitively plausible in its own right, but it's also superior to its rivals from the literature. For each of its rivals are, I'll argue, subject to one or more significant flaws. Perhaps, the worst flaw of them all is to fail to even address the central issue at hand, which, as you'll recall, is: What ought the self-torturer to do this week and each

subsequent week given his actual preferences? I call this 'Flaw 1' or 'F1' for short. Views with F1 include the Arntzenius-McCarthy view (1997). For although their view tells us that the self-torturer's preferences are irrational and, so, must be changed, it doesn't tell us what he should do given his actual preferences. It doesn't, for instance, tell us whether he should advance from 0 to 1 the first week. Another view with F1 is the Raffman-Tenenbaum view. Their view tells us that it would be permissible for the self-torturer to perform various series of actions but not whether it is, a given week, permissible for him to advance.[19] I take such views, although interesting in their own right, to be inadequate in that they fail to address Quinn's central question.

Other rival views suffer the same flaw that P5 suffers from: the flaw of denying that both of the following highly plausible claims are true: (claim$_1$) Erasmus ought to proceed from 0 to 50 while resolving to stop there and (claim$_2$) joint satisfiability, which, you'll recall, holds that if a subject both ought to φ and ought to ψ, then she has the option of both φ-ing and ψ-ing. And I call this 'Flaw 2' or 'F2' for short. Views with F2 include Chrisoula Andreou's interpretation of the standard view (2006). On the standard view, an agent ought to φ if and only if her φ-ing would serve her concerns well. And, on Andreou's interpretation of this view, her φ-ing would serve her concerns well if and only if her φ-ing would be part of her performing an action or a course of action that would serve her concerns well.[20] So, on Andreou's interpretation of the

[19] White makes this criticism in his 2017b.

[20] Technically, what she says is: "whether an action serves the agent's concerns well [or, more precisely, at least as well as the alternative available moves] depends on what action(s) or course(s) of action it is part of" (2006, p. 594)—and note that the brackets are in the original. But this doesn't allow us to evaluate prospective actions. For a prospective action isn't one that's been performed and, so, we can't say what courses of action it *is* a part of but only what courses of action it *could* or *would* be a part of if it were performed. And, admittedly, Andreou may just want to deny the possibility of assessing an act such as advancing from 0 to 1 independent of its being a part of some larger course of action. That this may be what she wants is suggested by the following passage: "Without any information about what larger action or course of action, if any, Tanya's picking up the tempting shot of tequila is part of, we can safely say that Tanya's picking up the shot, considered *in and of itself*, does not serve her concerns at all" (2006, p. 598). Perhaps, then, she thinks that we can say only that Erasmus ought not to advance from 0 to 1 as part of going all the way to 1,000 but ought to advance from 0 to 1 as part of going to, and then stopping at, 50. Perhaps, then, she would deny that we can say whether Erasmus ought to advance from 0 to 1,

standard view, Erasmus ought not to proceed from 0 to 1 given that, as a matter of fact, he would end up proceeding all the way to 1,000 if he were to do so. For given that he would end up proceeding all the way to 1,000, his proceeding from 0 to 1 would be part of his proceeding to 1,000, which is not a course of action that would serve his concerns well. But, given that Erasmus has the option of proceeding from 0 to 50 while resolving to stop there and would end up stopping at 50 if he takes this option, it seems that everyone should admit that Erasmus ought to proceed from 0 to 50 while resolving to stop there. And if the proponent of this interpretation of the standard view admits this, then, given that she must also hold both that he ought to refrain from proceeding from 0 to 1 and that he doesn't have the option of both refraining from proceeding from 0 to 1 and proceeding from 0 to 50 while resolving to stop there, she'll be forced to deny joint satisfiability, which, I've argued, she can't plausibly deny.

Still other rival views suffer the same flaw that P1 suffers from: the flaw of allowing that the self-torturer could be permitted to do something that he can do only by responding inappropriately to his reasons. I call this 'Flaw 3' or 'F3' for short. Views with F3 include Stephen White's view (White 2017b). He holds that "in deciding whether to advance from his current setting (n) to the next one (n+1), the self-torturer may do so if and only if proceeding from n to n+1 would have figured as a step in a plan he would have been rationally permitted to adopt at the outset" (2017b, p. 595). On this view, the self-torturer is, for any setting n ($0 \leq n < 1{,}000$), permitted to proceed from n to n+1 even if the only way for him to do so intentionally is by responding inappropriately to the decisive reason that he has to refrain from reconsidering his previous resolution not to go beyond n. Thus, White's view fails to acknowledge that a subject can't be permitted to do what she can do only by responding inappropriately to her reasons.[21]

---

considered in and of itself. But, in that case, I think that her view suffers from F1. It fails to answer questions such as the following: "Should Erasmus advance from 0 to 1 the first week?" This is the sort of question Quinn wanted an answer to.

[21] Admittedly, the claim that a subject cannot be permitted to do what she can do only by responding inappropriately to her reasons is a stronger claim than R4. But I think that it's equally plausible. For if I were permitted to do what I could do only by responding inappropriately to my reasons, then, given that I'm required to φ just when φ is my only permissible option, I would be required to do something that I

White's view has a further flaw. It fails to take seriously Quinn's assumption that practical rationality is purely instrumental. I call this 'Flaw 4' or 'F4' for short. White's view has this flaw, for it fails to explain why the fact that his proceeding from *n* to *n*+1 doesn't figure as a step in any plan he was rationally permitted to adopt at the outset is relevant to the permissibility of his so proceeding. For what does this fact have to do with how the relevant options compare in terms of satisfying his actual preferences? Whether proceeding from *n* to *n*+1 figures in such a plan seems to have nothing to do with how the relevant options compare in terms of satisfying his actual preferences. So, unlike myself, White fails to explain how we can take such a fact to be relevant if we're assuming A1—i.e., if we're assuming that practical rationality is purely instrumental. So, White's view fails to respect one of Quinn's key assumptions: namely, A1.

Another view with F4 is Eric Carlson's view (Carlson 1996). His is a complicated view, but, for our purposes, what matters is only that it holds that, for any setting *n* ($0 \leq n < 1{,}000$), the self-torturer's proceeding from *n* to *n*+1 is permissible only if *n*+1 is not greater than *k* (where *k* is less than 1,000). The problem is that Carlson fails to explain why the fact that *n*+1 is not greater than *k* is relevant to the permissibility of his proceeding from *n* to *n*+1. For what does this fact have to do with how the relevant options compare in terms of satisfying his actual preferences? Whether *n*+1 is greater than *k* seems to have nothing to do with how the relevant options compare in terms of satisfying his actual preferences. So, Carlson's view also fails to respect A1.

Other rival views suffer from a fifth flaw ('F5' for short) in that they hold, implausibly, that it could never be permissible for the self-torturer to proceed from 0 to 1 without first selecting some particular setting and resolving not to go beyond it. Such views are subject to the following counterexample. Imagine that a subject named Prudus agrees to have the self-torture devise attached to him in return for the described conditions. And imagine that he proceeds

---

could do only by responding inappropriately to my reasons when this option is my only permissible option. But, as R4 implies, I can't be required to do what I can do only by responding inappropriately to my reasons. (I'm assuming, here, that if I'm required to φ, then I ought to φ.)

from 0 to 1 with only a vague plan to stop somewhere well short of 1,000. And assume that he ends up stopping at some reasonable stopping point. Contrary to such views, it seems that he was right to proceed from 0 to 1. But, on those plan-based views that insist that the self-torturer not proceed even from 0 to 1 without first forming a resolution to stop at some particular reasonable stopping point, this is false. And, as we've just seen, this is implausible.[22]

Lastly, some rival views suffer from a sixth flaw ('F6' for short) in that they hold that there is no option that the self-torturer ought to take. As Carlson (1996, p. 147) has pointed out, this is true of the following sort of maximizing theory: for any option φ and any subject S, S is permitted to φ if and only if there is no alternative option ψ such that S's preferences would be better satisfied if she ψ-ed rather than φ-ed. On this theory, the self-torturer has no permissible option. After all, for any setting n ($0 \leq n < 1{,}000$), S could better satisfy his preferences by stopping at n+1 than by stopping at n. And the self-torturer could better satisfy his preferences by remaining at 0 than by stopping at 1,000. So, every one of the self-torturer's options is suboptimal in terms of preference satisfaction. And, hence, the maximizing theory implies that each of the self-torturer's options is impermissible. But this is implausible. Any plausible solution to the self-torturer puzzle must admit that the self-torturer ought to proceed to some reasonable stopping point and stop there.

So, not only is the compromise view intuitive in its own right, but, unlike its rivals, it is subject to none of F1–F6.

5. Conclusion

I've argued that no subject can be required to do what she can do intentionally only by responding inappropriately to her reasons. This means that whether the self-torturer should proceed from n to n+1 depends on how (BW_P) the best world that he could actualize if he were both to respond appropriately to his reasons and to proceed from n to n+1 compares to (BW_R) the best world that he could actualize if he were both to respond appropriately to his reasons

[22] Tenenbaum & Raffman make this criticism in their 2012 (p. 108).

and to refrain from proceeding from n to n+1. And I've argued that BW_P will compare favorably to BW_R except where he has resolved not to proceed beyond n. For where he has resolved not to proceed beyond n, there will be no world that he could actualize while both responding appropriately to his reasons and proceeding from n to n+1. But where he hasn't so resolved, there will be such a world and one that compares favorably BW_R. Thus, the self-torturer should not proceed beyond any setting that he has resolved not to go beyond. But he should definitely proceed beyond 0, because proceeding to some reasonable setting is both something that is better than remaining forever at 0 and something that he can do without responding inappropriately to his reasons. Thus, I've argued that he should proceed to some reasonable stopping point and then remain there forever, forming, when necessary, the resolution not to proceed beyond some particular reasonable stopping point. And I've argued that this view—the compromise view—is superior to its rivals, which, as I've shown, all have one or more significant flaws.

The fact that no subject can be required to do what she can do intentionally only by responding inappropriately to her reasons is, I think, an important and often overlooked fact. As I've argued, here, it can help us solve a puzzle that otherwise seems intractable: the self-torturer puzzle. But, as I've argued elsewhere (Portmore Forthcoming, chap. 5), it also helps us to solve other puzzles: e.g., how to best understand the common thought that a moral theory ought to be such that the agents who satisfy it, whoever and however numerous they may be, are guaranteed to produce the morally best world that they could (in the relevant sense) together produce.[23] And, as I've argued elsewhere (Portmore 2011, chap. 2), it has important implications with respect to whether we should accept moral rationalism: the view that a subject can be morally obligated to do only what she has decisive reason to do, all things considered.[24]

[23] Proponents of this thought include Baier 1958, Casteñeda 1974, Parfit 1984 (p. 94), Pinkert 2015, Regan 1980, and Zimmerman 1996.

[24] For very helpful comments and discussions on an earlier draft, I thank Chrisoula Andreou.

References

Andreou, C. (2014). "Temptation, Resolutions, and Regret." *Inquiry* 57: 275–292.

———. (2006). "Temptation and Deliberation." *Philosophical Studies* 131: 583–606.

Arntzenius, F. and D. McCarthy (1997). "Self Torture and Group Beneficence." *Erkenntnis* 47 129–44.

Baier, K. (1958). *The Moral Point of View*. Ithaca, NY: Cornell University Press.

Bratman, M. E. (2009). "Intention, Belief, Practical, Theoretical." In S. Robertson (ed.), *Spheres of Reason*, pp. 29–62. Oxford: Oxford University Press.

———. (1987). *Intention, Plans, and Practical Reason*. Cambridge, Mass.: Harvard University Press.

Carlson, E. (1996). "Cyclical Preferences and Rational Choice." *Theoria* 62: 144–160.

Castañeda, H.-N. (1974). *The Structure of Morality*. (Springfield, Ill.: Charles Thomas Publisher).

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

Huemer, M. (2018). *Paradox Lost*. Palgrave Macmillan.

Jackson, F. and R. Pargetter (1986). "Oughts, Actions, and Actualism." *Philosophical Review* 95: 233–255.

Kiesewetter, B. (Forthcoming). "Contrary-to-Duty Scenarios, Deontic Dilemmas and Transmission Principles." *Ethics*.

Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Pinkert, F. (2015). "What If I Cannot Make a Difference (and Know It)." *Ethics* 125: 971-998.

Portmore, D. W. (Forthcoming). *Opting for the Best: Oughts and Options*. New York: Oxford University Press.

———. (2011). *Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York: Oxford University Press.

Quinn, W. (1990). "The Puzzle of the Self-Torturer." *Philosophical Studies* 59: 70–90.

Regan, D. (1980). *Utilitarianism and Co-operation*. New York: Oxford University Press.

Ross, J. (2009). "How to Be a Cognitivist about Practical Reason." In R. Shafer-Landau (ed.) *Oxford Studies in Metaethics:* Volume 4, pp. 243–282. Oxford: Oxford University Press.

Tenenbaum, S. and D. Raffman (2012). "Vague Projects and the Puzzle of the Self-Torturer." Ethics 123: 86–112.

Wedgwood, R. (2007). *The Nature of Normativity*. Oxford: Oxford University Press.

White, S. J. (2017a). "Transmission Failures." *Ethics* 127: 719–732.

———. (2017b). "The Problem of Self-Torture: What's Being Done?" *Philosophy and Phenomenological Research* 94: 584–605.

Zimmerman, M. J. (1996). *The Concept of Moral Obligation*. Cambridge: Cambridge University Press.