What's a rational self-torturer to do?

DOUGLAS W. PORTMORE

ABSTRACT: This paper concerns Warren Quinn's "The Puzzle of the Self-Torturer." For the sake of argument, I accept Quinn's assumption that what the self-torturer ought to do is purely a function of how the relevant options compare in terms of satisfying his actual preferences. But I argue that, even so, we can explain why, in certain instances, advancing is not what the self-torturer ought to do without having to claim that, in those instances, his advancing would be no better than his not advancing in terms of satisfying his actual preferences. For we can admit that his advancing is, in every instance, better than his not advancing in terms of satisfying his actual preferences but argue that, in certain instances, advancing fails to meet some other necessary condition for being what he ought to do. One such necessary condition is that it be a genuine option for him. Another is that it be something that he can do without responding inappropriately to his reasons—or so I'll argue. I believe, then, that the solution to the puzzle lies in realizing that, in certain instances, advancing is something that the self-torturer can do only by responding inappropriately to his reasons.

Someone could 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 could 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 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, the government authority who is deciding how many citizens to order to reduce their carbon emissions. For every positive number n that's less than the total number of citizens, the authority prefers the world in which she orders only n citizens to reduce their

Figure 1

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

So, what should an agent with cyclical preferences do? To figure this out, it would helpful to consider an example that abstracts away from any irrelevant real-world complications. Fortunately, Warren Quinn (1990) has constructed 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. After all, what we want to know is how he should act *given the preferences that he actually has* (Quinn 1990, p. 80). What's more, he may be powerless to change his preferences. Indeed, let's assume so. Furthermore, let's accept Quinn's other assumptions. They are as follows.

(A1) Practical rationality is purely instrumental such that what a subject ought to do is solely 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 advance 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 advance all the way to 1,000.[5]

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

(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 advance to $n+1$ if she has, when at n, both (1) the option to advance 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).

advance 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 advance to and then stop at n+1 and (2) a preference for the way the world would be if she were to advance 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 advance to n+1, because advancing to and then stopping at n+1 is better than her remaining forever at n in terms of satisfying 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 advance to and then stop at n+1 and (2) a preference for the way the world would be if he were to advance 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 advance to n+1. [From P1–P2]

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

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

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

(C4) Therefore, if the self-torturer were to advance 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 advance all the way to 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 advance 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 advance all the way to 1,000, then she ought not, when at 0, to advance 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 advance 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 advance all the way to 1,000, then she ought not to advance from 0 to 1, because remaining forever at 0 is better than advancing all the way to 1,000 in terms of satisfying her actual preferences]

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

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

As this argument shows, 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, we should reject both. Given A1, we must assume that practical rationality is purely instrumental such that what a subject ought to do is purely a function of how the relevant options compare to each other in terms of satisfying her actual preferences. But, as I'll argue, neither the actualist view nor the possibilist view is correct about what we should be comparing. To illustrate, suppose that we're wondering whether someone attached to the self-torture device should advance from 0 to 1 the first week. According to the possibilist view, we should compare the worlds that she *could* actualize if she were to advance from 0 to 1 to the worlds that she *could* actualize if she were to refrain from advancing 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 only world that she could actualize if she were to refrain from advancing 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 advances to and then stops at 1 (a world that she could actualize if she were to advance from 0 to 1), the possibilist view implies that she should advance from 0 to 1.

By contrast, the actualist view holds that we should instead compare the world that she *would* actualize if she were to advance from 0 to 1 to the world that she *would* actualize if she were to refrain from advancing 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 advance 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 (that is, a person who

wouldn't stop until reaching 1,000 if she were to advance from 0 to 1) should not advance from 0 to 1.[6]

Both views 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 instances, advancing 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 can do only by responding inappropriately to her reasons. Consequently, we must, contrary to the possibilist view, exclude such options from the relevant comparison. For the only worlds that we should be comparing are those that she could actualize by responding appropriately to her reasons. And, so, what we should be comparing are the worlds that she could actualize by both responding appropriately to her reasons and advancing from *n* to *n*+1 to the worlds that she could actualize by both responding appropriately to her reasons and refraining from advancing from *n* to *n*+1. The key, then, to solving the puzzle of the self-torturer is to realize that someone attached to the self-torture device can't be required to advance from *n* to *n*+1 if that's something that she could do only by responding inappropriately to her reasons.

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

As noted, I believe that the key to solving the puzzle of the self-torturer 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 three 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-ing isn't even an option for her.

For instance, the reason that effecting world peace is not what I ought to do is that it isn't even an option for me.

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 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 I ought to do 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 action, φ, might not be what a subject ought to do. Here's another.

(R2) Although φ-ing is an option for her, it's incompatible with her performing an even better option.

[7] For some responses to worries about such a control condition (including both worries based on the thought that there's resultant moral luck and worries based on Frankfurt's examples involving counterfactual interveners), see Portmore Forthcoming (chap. 2).

To illustrate, assume that I have complete control over whether I raise both, neither, or just one of my two arms. And assume that, in raising just one, I have complete control over whether it's my left or my right arm that I raise. Lastly, assume that raising both my arms is better than every other option. In that case, refraining from raising my left arm is not what I ought to do, for it's incompatible with my performing an even better option: that of raising both my arms. And this result seems to generalize. So, I think that we should accept R2.

R1 and R2 are, I believe, fairly uncontroversial. So, I haven't actually argued for either of them. But I will argue that there is at least one other—admittedly, more controversial—reason why a possible action, φ, might not be what a subject ought to do. It's this.

(R3) Although φ-ing is an option for her that isn't incompatible with her performing an even better option, the only way for her to φ is by responding inappropriately to her reasons.

Why should we accept R3? Well, consider that what determines whether oughts, requirements, and responsibilities apply to a given subject is whether that subject has the capacity for responding appropriately to reasons. Indeed, it's the 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 who can have obligations and responsibilities from those who can't. And, given this, it would be nonsensical for a subject to be required to respond inappropriately to her reasons. For such a requirement would apply to her in virtue of her having the capacity to respond appropriately to reasons even though it's only by failing to properly exercise this capacity that she could come to fulfill this requirement. And that's implausible. For it's implausible to suppose that the very capacity in virtue of which a subject is responsible for her failures is the one that, when exercised fully and flawlessly, leads her to fail. And this holds whether we're talking about a failure to do what's required or a failure to do what ought to be done. It just can't be that what makes a subject responsible for her failures is the very thing that

causes her to fail. And, so, it can't be that a subject ought to do something that she could do only by responding inappropriately to her reasons.

In further support of R3, consider the following two examples. First, consider that I have been known to 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 causes people to live longer on the basis of my awareness of studies demonstrating a correlation between taking a daily multivitamin and living longer. So, if this is the only way for me to form this belief, then it seems that I ought not to form it. Yet, forming it could still be my best option (and, thus, not incompatible with my performing any better option), for suppose that this belief is both true and one that it would be beneficial for me to form. Nevertheless, it's not a belief that I should form if I can do so only by responding inappropriately to my reasons.[8] For if that were the case, this would constitute a directive that I could comply with only by doing something that I shouldn't do: inferring causation from correlation. And it's implausible to suppose that I should do something that I could do only by doing something that I shouldn't.

Second, suppose, contrary to fact, that I sometimes commit the following practical fallacy: that of forming an intention to φ on the basis of both an intention to ψ and the belief that my ψ-ing will cause φ to occur. Suppose, then, that I sometimes form the intention to sweat on the basis of both an intention to run and the belief that my running will cause me to sweat. Now, given that I know that I can't sweat merely by intending to sweat (at least, not directly), let's assume, as seems plausible, that I can't form the intention to sweat by responding appropriately to my reasons. Assume, then, that the only way for me to form this intention is by responding inappropriately to my reasons, forming it via the above practical fallacy. But, in that case, I don't see how anyone could plausibly claim that I ought or am obligated to form this

[8] Given that I'm stipulating that this belief is true and one that it would be beneficial for me to have, I concede that it would be good if I were to form this belief. I just deny that I ought to form it. For I can do so only by responding inappropriately to my reasons. So, this is a case where it would be rational for me to cause myself to be irrational by, say, taking a pill that would cause me to form this belief via a fallacious inference.

intention. For this just seems implausible given that this directive would apply to me only if I had the capacity to respond appropriately to my reasons, and, yet, I could comply with it only by failing to fully and flawless exercise this capacity. Thus, I think that we should accept R3.[9]

Now, if I'm right about R1–R3, then not every explanation for why advancing is not what the self-torturer ought to do must appeal to R2 and the thought that advancing would be incompatible with performing an option that's better in terms of satisfying his preferences. For, given R1, the explanation could instead appeal to the thought that advancing isn't even an option for him. Of course, Quinn assumes that the self-torturer always has the option of advancing (at least, until he reaches 1,000). So, the thought that advancing isn't an option for him is false, just as the thought that advancing is incompatible with his performing an even better option is. So, neither an appeal to R1 nor an appeal to R2 is going to help us solve the puzzle of the self-torturer. But, fortunately, there is, I've argued, one other reason why advancing might not be what the self-torturer ought to do—that is, R3. And, so, it could be that the reason that advancing to the next setting is, in certain instances, not what the self-torturer ought to do is that it's something that he could do only by responding inappropriately to his reasons. Indeed, I'll argue that this is 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 advance to n+1 if she has, when at n, both (1) the option to advance to and then stop at n+1 and (2) a preference for the way the world would be if she were to advance 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 a subject named Quintin (meaning 'the fifth', but I'll use it to bring to

[9] See Portmore Forthcoming (chap. 5) for an argument that we must accept R3 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.

mind the number 50) is now at 50, having resolved not to go beyond 50. Thus, he intends to resist the anticipated temptation to advance to 51 in light of his having both (1) the option to advance to and then stop at 51 and (2) a preference for the way the world would be if she were to advance 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 Quintin ought to advance to 51. But, as we'll soon see, advancing to 51 can't be something that Quintin ought to do. To see why, consider that, given that he's resolved not to go beyond 50, there are only two ways for him to advance 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 resolution not to go beyond 50, thereby changing his mind and deciding to advance by intentionally flicking the dial from 50 to 51. Call this *The Reconsidered Procession to 51*. And let's assume that, in both instances, Quintin will remain forever at 51 and not advance any further. Also, let's assume that 50 and 51 are both what I'll call *preferred stopping points*, where a setting counts as a preferred stopping point if and 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 advances all the way to 1,000.[10] And I'll call any stopping point that isn't a preferred stopping point a *dispreferred stopping point*.

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 his actual preferences. And, so, I admit that, 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 take it. Nevertheless, I deny that he ought to advance from 50 to 51. And I can deny this even if I accept Quinn's

[10] Given Quinn's assumptions, there is no perfect stopping point (that is, no stopping point that isn't dispreferred to some alternative), because, for every setting n ($0 \leq n \leq 1{,}000$), stopping at n will be dispreferred either to stopping at the next setting (i.e., $n+1$) or to stopping at the initial setting (i.e., 0).

assumption that practical rationality is purely instrumental. For, as I've shown, even if practical rationality is purely instrumental, there can be reasons why an act isn't one that ought to be performed that have nothing to do with how good or bad that act is in terms of satisfying the subject's preferences. Indeed, I've argued that there are at least two such reasons: R1 and R3. Moreover, as I've just shown, there are only two ways for Quintin to advance 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 R3. 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, as I'll show, one that he can perform only by responding inappropriately to his reasons. So, it can't be what he ought to do given R3. Therefore, advancing to 51 is not something Quintin 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 Quintin 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 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 have done. Likewise, Quintin 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 him. And, since it's not an option for him, it can't, according to R1, be something that he ought to do.

The only other way for Quintin to advance 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 made this resolution in the first place was to prevent the anticipated inclination to advance to 51 from thwarting his plan to stop at 50. And given that the function of a resolution is to prevent such anticipated inclinations from undermining one's ability to successfully follow through with one's 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 Quintin has resolved not to go beyond 50 and, so, intends to resist the anticipated inclination to advance to 51, he cannot now advance to 51 in response to that very inclination except by responding inappropriately to the decisive reason that he has for not doing so. 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 Quintin ought to advance to 51 via *The Reconsidered Procession to 51.*

Here, I'm relying on the following principle (call it *the no-reconsideration principle*): Assuming that a resolution was permissibly formed, a subject has decisive reason not to reconsider it in response to the very inclinations that it was intended to prevent from resulting in reconsideration—unless, that is, something relevant has changed between now and then, such as, perhaps, the subject's preference ranking. This principle implies that Quintin has decisive reason not to reconsider his resolution to stop at 50 in response to his inclination to get an additional $10,000, for let's assume, as seems plausible, that his resolution was permissibly formed. It implies this both because his resolution was intended to prevent him from reconsidering in response to this very inclination and because nothing relevant has changed. After all, he still ranks the end of ensuring that he doesn't end up at some dispreferred stopping point above the end of getting an additional $10,000.

Since this principle may not be entirely uncontroversial, I should explain why I accept it. The main reason for accepting 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 for accepting it is simply that it's intuitive. That is, it's intuitive to think that, if the point of your having formed a resolution was to prevent certain anticipated inclinations from leading you to reconsider your original plan, then it's irrational to reconsider that plan in response to those very inclinations—unless, that is, you were wrong to have formed the resolution in the first place or something relevant has changed since you formed it. This and the fact that the principle has plausible implications seems to constitute 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 deflect—at least, not when they were permissibly formed and nothing relevant has changed since they were formed. Also, note that this principle doesn't have the same problematic implications that a similar principle has. That principle (and call it *the problematic principle*) states that, "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)

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 this 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.[11]

[11] Unlike some others, I don't find that it's necessarily irrational to reconsider a resolution in light of a change in one's preference ranking even if that resolution was formed with the intention of keeping such a change from preventing one from following through with the original plan. As Andreou's case illustrates, such a change can make it rational to reconsider such a resolution. Moreover, the no-reconsideration principle explicitly allows for the possibility that such a change is a relevant one. Lastly,

Of course, if we want to test the implications of the no-reconsideration principle, we could just modify Andreou's original example to the point where the no-reconsideration principle would then imply that B should not reconsider A's resolution. We can, thereby, test whether this implication is intuitively plausible in such a modified case. To do so, let's assume that her earlier self (viz., A) made the resolution with the intention of preventing the anticipated future inclination to spend her spring break working on her dissertation from stopping her from following through on her plan to go to the yoga retreat. Let's assume that she made this resolution with the intention of ensuring that she doesn't get burnt out working 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 on her dissertation above the end of spending her spring break working on her dissertation. Now, if we make these modifications, 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. So, it seems that the no-reconsideration principle is not only intuitive in its own right, but also intuitive in its implications.

To sum up, then, we must reject P1 for it implies that Quintin ought to do something (viz., advance from 50 to 51) that he can do only either accidentally or by responding inappropriately to his reasons. For, given R1 and R3, 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 advance from 0 to 1, and if she prefers the way the world would be

---

given all this, it's important to note that, in the case of the self-torturer, his preference ranking never changes. At all times, he prefers 0 to 1,000 but 1 to 0, 2 to 1, 3 to 2, …, and 1,000 to 999.

if she were to stop at 0 to the way the world would be if she were to advance all the way to 1,000, then she ought not, when at 0, to advance to 1. To see that this premise is false, imagine that there is a subject named Imprudus (for being imprudent) 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 advancing all the way to 1,000 if he were to advance from 0 to 1. Now, P5 implies that Imprudus shouldn't advance from 0 to 1. But, as I'll show, this is false. And, so, we must reject P5.

Let's assume both that Imprudus is now at 0 and that 50 is, for some reason, the only preferred stopping point for him. Let's also assume that Imprudus (1) has the option of advancing to and then stopping at 50, (2) would advance 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—recall that 50 is the only preferred stopping point for him, and (4) prefers the way the world would be if he were to advance 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 Imprudus ought to advance from 0 to 50 while resolving to stop there. Yet, even so, P5 implies that Imprudus ought not to advance 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, advance all the way to 1,000 if he were to advance from 0 to 1.

The problem for the proponent of P5 is that she cannot accept both of the following two highly plausible claims: (claim$_1$) Imprudus ought to advance from 0 to 50 while resolving to stop there and (claim$_2$) *joint satisfiability*, which 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$_1$ (and it seems that we all should), she must deny claim$_2$ given P5's implication that Imprudus ought not to advance from 0 to 1. After all, Imprudus doesn't have the option of both (1) not advancing from 0 to 1 and (2) advancing from 0 to 50 while resolving to stop there. Thus, the proponent of P5 must reject joint satisfiability given that she should accept claim$_1$. But even if joint satisfiability

isn't entirely uncontroversial (see, e.g., White 2017a), we should accept it given the following argument.[12]

(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, 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) A subject who believes 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 has an irrational set of beliefs and intentions. [From the consistency requirement on beliefs and intentions[13]]

---

[12] This argument is, in part, inspired by one found in Kiesewetter Forthcoming, which is a response to White 2017a. It differs from his, though, in that it concerns the objective (i.e., fact-relative) ought.

[13] 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.

(C11) Thus, if this subject believes what's true and intends to do all that she ought to intend to do, then she'll have an irrational set of beliefs and intentions. [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, she'll have an irrational set of beliefs and intentions>. For it's implausible to suppose that a subject who believes what's true and intends to do all that she ought to intend to do has an irrational set of beliefs and intentions. [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 entailed on the view that practical oughts are normative for intentions. The intuitive idea behind this view is that, if you ought to perform an action, then, given that it couldn't be that you ought to perform it unintentionally, you ought to form the intention to perform it—or, at least, the intention to perform something that entails performing it.[14] Nevertheless, there may seem to be counterexamples to this view. For one, it may seem 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 could do intentionally. For, in that case, acting spontaneously doesn't seem to be a genuine option for me. Thus, we should hold, instead, only that I ought to perform (intentionally) some distinct act that would increase my chances of unintentionally acting spontaneously—e.g., the act of consuming a few stiff drinks. For another, it may seem that I ought to refrain from torturing 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—at least, in any typical context. So, if I need to form the intention to refrain from torturing children to prevent myself from doing so, then something has gone terribly

[14] For any subject S and any two of her options φ and ψ, S's φ-ing entails S's ψ-ing if and only if S doesn't have the option of φ-ing without ψ-ing. And, to save words, I will sometimes say 'φ-ing entails ψ-ing' instead of 'S's φ-ing entails S's ψ-ing'.

wrong. But although I admit that something has gone wrong if I need to form this specific intention, it still seems to me that I ought to form some (more general) intention that entails my refraining from torturing children—perhaps, the intention to avoid 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.[15] So, neither of these putative counterexamples turn out to be persuasive upon reflection. And, so, I believe that we must accept P7.

P8 is entailed by the view that any set of beliefs and/or intentions whose propositional contents are logically inconsistent is irrational. Yet, here too, there may seem to be some putative counterexamples to this, the consistency requirement on beliefs and intentions. 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 this set of beliefs consisting in her believing each one of these other claims is irrational or the consistency requirement is false. And it may seem that, given how carefully she's researched each one of her claims, there's nothing irrational about her believing each one. It may seem, therefore, that we should 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, it is irrational for her to believe all of them—at least, it is in the objective (fact-relative) senses of 'rational' and '(rationally) 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. That is, it's objectively irrational 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 all of them—specifically, she's wrong to believe the false ones.

[15] 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, it is objectively irrational for her to have all 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 in fact 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.[16] 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 is low enough where n equals 20, then just change the example so that n equals, say, twenty thousand or more. 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

[16] 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, in this case, what's up to her (and, thus, what's an option for her) is not whether she visits a given cathedral but only whether she *tries* to visit it.

To sum up, I've shown that P5 implies that it can't be that $claim_1$ and $claim_2$ are both true. Yet, $claim_1$ is incontrovertibly true. And, as I've argued, $claim_2$ is also true. So, we must reject P5.

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 preferred stopping point 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 preferred stopping point and to resolve to stop there.[17] And, ($conjunct_2$) for any setting n ($0 \leq n < 1{,}000$), she ought to advance from n to n+1 if and only if she has not violated the requirement stated in $conjunct_1$ and can intentionally advance 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—the views that give rise to P1 and P5, respectively. Like the possibilist view, the compromise view holds that such a subject should typically advance from n to n+1 if she has, when at n, both (1) the option to advance to and then stop at n+1 and (2) a preference for the way the world would be if she were to advance 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 in response to the very inclination that 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 advance to 1,000 or any other dispreferred stopping point. But, unlike the actualist view, it doesn't do this by forbidding her from advancing from n to n+1 if, in fact,

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

doing this would lead to her advancing to some dispreferred stopping point. Instead, it does this by requiring her to form, when necessary, a resolution that will ensure that she doesn't advance to any dispreferred stopping point. Thus, in the above example, Imprudus 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 dispreferred stopping point. Thus, on the compromise view, the mere fact that he's not going to form the resolution that he's required to form doesn't imply that it's impermissible for him to advance from 0 to 1. It just implies that it's impermissible for him to do so without forming, when necessary, the resolution that will prevent him from advancing to any dispreferred stopping point. After all, he's perfectly capable of advancing from 0 to 1 while resolving to stop at some particular preferred stopping point, and doing this will ensure that he doesn't advance to any dispreferred 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 advance each week until reaching the particular preferred stopping point that she has resolved not to go beyond. And it holds that she should ensure that she doesn't advance to any dispreferred stopping point by forming, when necessary, a resolution to stop at some particular preferred stopping point. Thus, it prohibits her from advancing to 1,000 or to any other dispreferred stopping point. Indeed, it prohibits her from advancing to even some preferred stopping point unless she either has already formed a resolution that will prevent her from advancing to any dispreferred stopping point or can do so without the risk of ending up at any dispreferred stopping point.

Of course, many will wonder why we should accept the compromise view's implication that such a subject shouldn't advance beyond the particular preferred stopping point that she's resolved to stop at. For instance, many will wonder why Quintin—who has resolved to stop at 50—shouldn't advance from 50 to 51 given that he has both (1) the option to advance to and then stop at 51 and (2) a preference for the way the world would be if she were to advance 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 Quintin 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 advancing 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 as relevant to 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 advancing 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 advance in response to the very inclinations that it was intended to overcome. Thus, I deny Stephen J. 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, revising that plan can't be what one ought to do if one doesn't even have the option of revising that plan—see R1. And that's true even if revising that plan would better serve one's interests. For another, revising that plan can't be what one ought to do if one can do so only by responding inappropriately to one's reasons—see R3. Thus, I think that we can meet the demand for a defense of the above implication by appealing to R3 and the fact that, given Quintin's resolution to stop at 50, he can revise that plan only by responding inappropriately to the decisive reason he has to refrain from doing so.

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)

[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.

have the self-torturer device set permanently to n and receive $n \times \$10{,}000$ or (B) have the device 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 this view holds that the only reason that it would be irrational 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 resolution that she was required to form. But since conjunct$_1$ of the compromise view doesn't require a subject facing such a one-off 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 (a name associated with 'impending doom' in Greek mythology) 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 advance. 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 he 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 advanced any of those weeks in which he had violated the requirement stated in conjunct$_1$—the requirement to form, when necessary, a resolution to stop

at some particular preferred stopping point. So, he went wrong in advancing each of those weeks in which he advanced having already violated the requirement stated in conjunct$_1$. And since he did end up at a dispreferred stopping point, it was clearly necessary for him (and, thus, a requirement for him) to have formed a resolution to prevent him from doing so.

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, according to the purely instrumental conception of practical rationality, should the self-torturer do this, the first, 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, or what he should do if he's powerless to change them. 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 (2012). Their view tells us that it would be permissible for the self-torturer to perform various series of actions over the next several weeks but not whether it is, a given week, permissible for him to advance to the next setting.[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$_1$) Imprudus ought to advance 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

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

course of action that would serve her concerns well.[20] So, on Andreou's interpretation of the standard view, Imprudus ought not to advance from 0 to 1 given that, as a matter of fact, he would end up advancing all the way to 1,000 if he were to do so. For given that he would end up advancing all the way to 1,000, his advancing from 0 to 1 would seem to be part of his advancing to 1,000, which is not a course of action that would serve his concerns well. But, given that Imprudus has the option of advancing 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 Imprudus ought to advance 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 advancing from 0 to 1 and that he doesn't have the option of both refraining from advancing from 0 to 1 and advancing 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 advancing

[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 assess 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. Of course, 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. This 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 Imprudus 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 Imprudus ought to advance from 0 to 1, 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 Imprudus advance from 0 to 1 the first week?" This is the sort of question that Quinn and I want an answer to.

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 advance 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]

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 the self-torturer's advancing 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 advancing. For what does this fact have to do with how the relevant options compare in terms of satisfying his actual preferences? For suppose, he is at 900 and is deciding whether to advance to 901. And suppose that he doesn't care whether advancing to 901 would figure as a step in a plan that he was rationally permitted to adopt at the outset. Why, then, would this fact have anything do with whether he's permitted, on a purely instrumental conception of practical rationality, to advance from 900 to 901? It seems that it doesn't.

Of course, White holds that which plans the self-torturer can, at the outset, permissibly adopt is entirely dependent on how the available plans compare to each other in terms of satisfying his preferences. But we're not asking which of the available plans the self-torturer is permitted to adopt at the outset. We're asking whether he's permitted to advance from 900 to 901. And given that he has no preference for acting in accord with a plan that he was rationally

[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 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.)

permitted to adopt at the outset, I don't see how this has anything to do with how his current set of options (that is, advance or don't advance from 900 to 901) compare in terms of satisfying his actual preferences. So, White fails to explain how we can take such a fact to be relevant if we're assuming A1—that is, if we're assuming that practical rationality is purely instrumental. Thus, White's view fails to respect one of Quinn's key assumptions: namely, A1.[22]

By contrast, my view respects A1 in that I accept that whether the self-torturer ought to advance a given week is solely a function of how the relevant options at present compare to each other in terms of satisfying his actual preferences. It's just that I deny that things like *The Accidental Procession to 51* and *The Reconsidered Procession to 51* are relevant options. I've argued that *The Accidental Procession to 51* isn't a relevant option, because it isn't even an option. And I've argued that *The Reconsidered Procession to 51* isn't a *relevant* option, because it isn't the sort option that could be what he ought to perform given that he can perform it only by responding inappropriately to his reasons.

Another view with F4 is Erik 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 advancing 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 advancing 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? Indeed, 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 advance from 0 to 1 without first selecting some particular setting and resolving not to go beyond it. Such views are subject to the

[22] Of course, White may think both (1) that, if we accept all of Quinn's assumptions, then there can be no satisfactory solution to the puzzle and (2) that, for this reason, we should deny A1. But, as I've argued here, we can accept all of Quinn's assumptions and still come up with a satisfactory solution.

following counterexample. Imagine that a subject named Prudus (for being prudent) agrees to have the self-torture device attached to him in return for the described conditions. And imagine that he advances from 0 to 1 with only a vague plan to stop somewhere well short of any dispreferred stopping point. And assume that he ends up stopping at some preferred stopping point—and, thus, he never violates the requirement stated in conjunct$_1$. Contrary to such views, it seems that he was right to advance from 0 to 1. But, on those plan-based views that insist that the self-torturer not advance even from 0 to 1 without first forming a resolution to stop at some particular preferred stopping point, this is false. And this is implausible given that, in some instances (such as Prudus's), forming such a resolution is unnecessary for ensuring that he stops at some preferred stopping point.[23]

Lastly, some rival views suffer from a sixth flaw ('F6' for short) in that they hold that there is never any option that the self-torturer is permitted 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 than if she φ-ed. On this theory, the self-torturer never has a permissible option. After all, it will, on the maximizing theory, be impermissible for him to stop at n ($0 \leq n < 1{,}000$) given that he could better satisfy his preferences by stopping at n+1. And it will, on the maximizing theory, also be impermissible for him to stop at 1,000 given that he could better satisfy his preferences by remaining forever at 0. So, every one of the self-torturer's options (stopping at 0, stopping at 1,000, or stopping somewhere in between) is suboptimal in terms of satisfying his preferences. And, hence, the maximizing theory implies that each of the self-torturer's options is impermissible. But that's implausible. For it seems that any plausible solution to the puzzle of the self-torturer must admit that the self-torturer is permitted to advance to some preferred stopping point and stop there.

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

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 something that she can intentionally do only by responding inappropriately to her reasons. This means that whether the self-torturer should advance from n to n+1 depends on how the worlds that he could actualize by both responding appropriately to his reasons and advancing from n to n+1 (call these *the advancing-appropriately worlds*) compare to the worlds that he could actualize by both responding appropriately to his reasons and refraining from advancing from n to n+1 (call these *the refraining-appropriately worlds*). And I've argued that the best advancing-appropriately world will be better than every refraining-appropriately world except where he has resolved not to advance beyond n. This is for two reasons. First, wherever he *has* resolved not to advance beyond n, there will be no world that he could actualize by both responding appropriately to his reasons and advancing from n to n+1—that is, there will be no advancing-appropriately worlds. And, clearly, a refraining-appropriately world is better than no world at all. Second, wherever he *hasn't* resolved not to advance beyond n, there will be some advancing-appropriately worlds that are better than any refraining-appropriately world. Thus, the self-torturer should not advance beyond any setting that he has resolved not to go beyond but should advance to some preferred stopping point since doing so is both 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 advance to some preferred stopping point and then remain there forever, forming, when necessary, the resolution not to advance beyond that preferred 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 puzzle

of the self-torturer. 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.[24] 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.[25]

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

[25] For very helpful comments and discussions on earlier drafts, I thank Chrisoula Andreou, Richard Yetter Chappell, Travis Timmerman, and Stephen J. White.

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.