Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
doi: 10.1111/phpr.12261

The Problem of Self-Torture: What's Being Done?

STEPHEN J. WHITE
Northwestern University

We commonly face circumstances in which the cumulative negative effects of repeatedly acting in a certain way over time will be significant, although the negative effects of any one such act, taken on its own, are insubstantial. Warren Quinn's puzzle of the self-torturer presents an especially clear example of this type of predicament. This paper considers three different approaches to understanding the rational response to such situations. The first focuses on the conditions under which it is rational to revise one's prior intentions. The second raises the possibility of a fundamental disconnect between the rational assessment of an extended pattern of choices and the assessment of the individual choices that make up that pattern. I show that neither adequately addresses the underlying issues. I propose a third approach, according to which the rational assessment of the "self-torturer's" choices is guided, not by any plan or intention the he has actually adopted, but by the plan or plans it would have been reasonable for him to adopt from the outset. The larger significance of this conclusion is brought out through the identification of conditions under which one's past choices can non-derivatively constrain the rational response to one's present circumstances.

There is a doughnut shop in my neighborhood that I pass on my way to the train. When I walk by it, I'm sometimes tempted by the following thought: I could stop in for a doughnut, which I would enjoy very much, and doing so would make practically no difference to my weight, or my health, or my bank account. Nothing I care about, I think to myself, would be affected negatively in any but the most negligible of ways were I to eat a doughnut right now. Moreover, this thought is true no matter how many doughnuts I've eaten in the past, or how many I will go on to eat in the future. There is apparently no downside to having one now.

Most of us are familiar with this type of predicament. Smoking one more cigarette is unlikely, by itself, to make any difference to your life expectancy. Reading one more page of your book won't prevent you from getting a good night's sleep—if it's not too late already, another 30 seconds won't make the difference. So why not take advantage of these facts and enjoy the pleasures that one more cigarette or one more doughnut will bring? Why not find out what happens next in the novel?

When I'm able to resist this peculiarly rational-sounding rendition of the doughnut shop's Siren song, it's usually by attending to the repeatable nature of this reasoning, along with the fact that, were I to eat a doughnut *every* time I passed by the shop, the results would not be pretty—not worth the momentary pleasures each doughnut would provide.

And yet, it's not clear just what rational bearing this second thought has on my decision as to whether I should have a doughnut *today*. It doesn't, for instance, seem to contradict my earlier premise that having a doughnut today will have a definite upside and no non-negligible downside. So, what gives?

This puzzle receives especially clear and vivid expression in Warren Quinn's case of the self-torturer. Here is how Quinn presents the case:

> Suppose 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.*[1]

The problem for the self-torturer is this: he can't (Quinn assumes) feel the difference between any two adjacent settings of the device, but he can tell the difference between settings that are far enough apart. Certainly, by the time he gets to 1000 on the dial, he will be in quite a lot of pain. He will be in enough pain that he would be willing to return all the money he had received ($10,000,000) in order to turn the device off.[2] Since he can't tell the difference between adjacent settings, however, when it comes to any two settings, n and n+1, he prefers the higher setting. At n+1, he feels no worse than he does at n, and he'll have an extra $10,000 dollars to spend on whatever he wants. His preferences are thus intransitive. Yet they seem

[1] Warren Quinn, "The Puzzle of the Self-Torturer," rep. in Quinn, *Morality and Action* (Cambridge University Press, 1993), p. 198.

[2] Is it plausible that he would be in this much pain at setting 1000, given how tiny each increase in voltage is from one setting to the next? I think so, once we remember that the device will cause him constant pain for the rest of his life. Assuming he has a long time still to live, the pain felt at any particular moment would not, I think need to be particularly intense in order for it to ruin the rest of his life.

reasonable. Considering that there is no experiential difference between adjacent settings while there is a great financial difference, it seems that, all else equal, for any n the self-torturer has good reason to prefer n+1 to n. And given a plausible view of the value of avoiding terrible, unending pain as compared to being very rich, the self-torturer seems also to have good reason to prefer setting 0 (no pain, no financial gain) to 1000 (lots of pain, lots of money).

In what follows, I will consider three different approaches to resolving the puzzle presented by Quinn's example. The first, and most common, is the plan-based approach, according to which the self-torturer should (a) adopt a reasonable plan at the outset about when to stop, and (b) stick to that plan. Different versions of this solution offer different theories, first, of what constitutes a reasonable plan, and second, of why the self-torturer should not abandon or revise it as he proceeds. I then consider a very different account recently put forward by Sergio Tenenbaum and Diana Raffman. Both approaches, I argue, are in important ways incomplete. I will then propose a different type of solution—one which combines the virtues of the other two approaches while avoiding their pitfalls. According to the solution I offer, the rational assessment of the self-torturer's choices is guided, not by any plan or intention the he has in fact adopted, but by the plan or plans it would have been reasonable for him to adopt, whether or not he has done so. In the final sections of the paper, I describe the general conditions under which this "hypothetical-plan" mode of reasoning is called for and answer the charge that this solution to the puzzle is problematically *ad hoc*.

1. Three Aspects of the Puzzle

First, we need to get clearer on what exactly the puzzle is that Quinn's case raises.

Broadly speaking, it's clear enough what the self-torturer should do. On normal background assumptions, what he should do is advance the dial at least a few times, and thereby make a lot of money, but stop before the pain gets to be too bad. Certainly, it seems he should stop before reaching a point at which he would definitely prefer to return the money and remove the device, if only he could. One question is how we could be more specific in our advice to the self-torturer. What principles or strategies can we offer to help the self-torturer select a stopping point or at least to avoid ending up in an unacceptable amount of pain? We might call this the "practical" aspect of the puzzle.

It's important, however, to distinguish this practical question from the deeper theoretical issues raised by the self-torturer's predicament. For even given the obvious and vague description of what the self-torturer should do—namely, that he should stop advancing the dial at some acceptable combination of pain and money—the problem is to understand the rationality of this response. More specifically, the puzzle raises the following questions.

First, what explains the irrationality of the self-torturer's proceeding all the way to setting 1000? In virtue of what, exactly, would this be irrational? After all, at each point the opportunity is presented, it looks like it makes sense for the self-torturer to increase the voltage. Supposing the self-torturer were to take advantage of every such opportunity, what would be his mistake?

Second, for any stopping point that is acceptable—does not involve too much pain, but secures him a reasonable sum of cash—what explains why, despite appearances, the self-torturer does not have decisive reason move on to the next setting? More generally, how can it be rational for the self-torturer to stop at any setting earlier than 1000?

There are then, two further aspects to the puzzle beyond the practical question of how best to achieve a desirable result. One is to understand why it would be irrational for the self-torturer to proceed all the way to the final setting. The other is to understand why it would not be irrational to stop advancing the dial prior to reaching the final setting.

2. Constraints on a Solution

I now want to note two constraints on an adequate solution to the theoretical problems just described.

First, an adequate solution to Quinn's puzzle will need to explain why it makes a difference that the self-torturer is (a) faced with a series of choices of a given type and (b) knows that choosing in the same way over and over again will have unacceptable consequences. This first constraint is captured by what Tenenbaum and Raffman call "non-segmentation."[3] Suppose the self-torturer were merely offered a single choice—he could either have the device set to one position, and receive the amount of money corresponding to that position, or he could have the device set to the subsequent position and receive an additional $10,000. Non-segmentation is the claim that, no matter which two adjacent settings we consider in such a one-off case, the self-torturer would be rationally permitted to opt for the higher of the two in return for the extra money.

A satisfactory solution must conform to non-segmentation if we accept the intuitive assumptions that generate the puzzle in the first place. It's essential that, for every setting n, it makes sense for the self-torturer to

[3] Sergio Tenenbaum and Diana Raffman, "Vague Projects and the Puzzle of the Self-Torturer," *Ethics*, vol. 123, no. 1 (2012), p. 98.

prefer n+1 to n. To deny this is simply to deny that there is any puzzle here to solve.[4]

The second constraint is this. The account 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. That is to say, the account must be suitably general. This may seem too obvious a constraint to bother stating explicitly. The reason I do so is that, as I will argue below, an important class of solutions to the puzzle all fail precisely because they violate this constraint.

3. The Plan-Based Approach

The standard approach to Quinn's puzzle is plan-based.[5] Plan-based accounts proceed in two stages. The first stage is to settle on a principle or method for deciding on a setting at which to stop. The second stage is to explain how, once the self-torturer reaches that setting, it can be rational for him to stick to his plan and stop.

For advocates of the plan approach, much of the philosophical action happens at the second stage. This is because, although it will be hard to say exactly where the self-torturer should plan to stop—maybe the best we can do is advise him to pick, more or less at random, a setting that falls within some acceptable range—nevertheless, it *will* be clear that he should not, for instance, intend from the beginning to go all the way to 1000.[6] That would obviously be a bad plan. The really difficult question, therefore, is why, supposing he's adopted a plan to stop at, say, 300, he shouldn't abandon that plan once he gets there and move on to 301. After all, we would normally suppose 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

[4] Those who wish to defend orthodox rational choice theory in the face of Quinn's puzzle must deny Non-segmentation, and so in effect deny that the case raises any genuine puzzle—or at any rate, the sort of puzzle that Quinn believes it raises. For rebuttals of various attempts to reconcile the self-torturer case with standard rational choice theory, see Quinn, "Puzzle," and Tenenbaum and Raffman, "Vague Projects." While I am persuaded that the case presents a genuine counterexample to the orthodox view, I do not have anything original to add to this debate, and so will not focus on it.

[5] For proponents of plan solutions, see Quinn, "Puzzle;" Michael Bratman, "Toxin, Temptation, and the Stability of Intention," in Bratman, *Faces of Intention, (*Cambridge University Press, 1999); Chrisoula Andreou, "Temptation and Deliberation." *Philosophical Studies* 131, 3 (2006); Erik Carlson, "Cyclical Preferences and Rational Choice" *Theoria* 62, 1–2 (1996). In fact, Carlson does not fully develop a plan-based solution, though he clearly endorses this approach. This is because he explicitly restricts himself to the first stage of the problem—determining the setting at which it's rational for the self-torturer to plan to stop—leaving it open why exactly the self-torturer should stick to his plan.

[6] Of course, there are more sophisticated strategies we might recommend. See section 8 below.

plan, then that's what one should do. Thus, if we take this approach, what we need to do is spell out certain conditions on the rational revision of prior intentions that make it clear why the self-torturer should stick with his original intention in this case.[7]

Even without getting into the details, it's clear why this approach is attractive. The puzzle depends on the thought that, when you look at every choice the self-torturer faces as he proceeds from week to week, the relevant factors are always the same and they always seem to favor the higher setting. But this is why having a plan can help: it introduces another factor that will at some point be relevant—viz., that this is the setting at which he intended to stop. And *this* factor will count *against* moving to the higher setting.

Nevertheless, this whole approach must be rejected, for a simple reason: plan-based solutions all violate the generality constraint on a satisfactory account. They can't 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. They can only explain the irrationality in cases where, in addition, the self-torturer has formed an intention to stop at a particular point. But imagine that the self-torturer does not come up with any plan about how to proceed. 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. He hasn't violated any principle of rational intention-revision. But surely he's gone wrong somewhere. The puzzle remains.

It will no doubt be suggested that plan solutions do offer an explanation of this sort of case. The self-torturer has gone wrong in not coming up with any plan in the first place. But this, I think, is the wrong kind of explanation. Even if it is foolish of the self-torturer not to plan ahead there seems to be some additional irrationality involved in his proceeding all the way to 1000. This is clear if we imagine the case as before, in which the self-torturer forms no plan about where to stop, but suppose that this time he's *correct* in thinking that he'll stop at some reasonable point before the pain gets too bad. It's possible that the self-torturer still deserves some criticism

[7] Different theorists have different accounts of the conditions on rational revision of one's intentions or plans. Quinn, for example, suggests that "a reasonable strategy that correctly anticipated all facts (including facts about preferences) still binds" (p. 207). If one has formed a reasonable intention about how to act at a later time, it's rational to revise that intention only if new information comes to light. And Bratman argues that that it is irrational to abandon one's plan if one can foresee that one will later regret doing so. He claims that the self-torturer has good reason to expect he will regret abandoning his plan since, according to Bratman, he should think that sticking to his initial plan is his best shot at avoiding the slippery slope to an unacceptable amount of pain. See Bratman "Toxin, Temptation and the Stability of Intention."

for not having decided ahead of time where to stop (though it's hard to see what the criticism would be exactly). But he has managed to avoid the central form of irrationality exhibited in repeatedly taking the money and thereby ending up in terrible pain. The plan-based approach is therefore no help when it comes to understanding and articulating *that* form of irrationality.[8,9]

4. Vague Projects and Top-Down Irrationality

Sergio Tenenbaum and Diana Raffman have recently proposed a novel approach that avoids this problem and easily satisfies the generality constraint. In their view, we can resolve the puzzle through a fairly modest and independently plausible extension of our ordinary conception of instrumental rationality. The self-torturer behaves irrationally in failing to take the necessary means to his presumed end of living a relatively pain-free life. But because living a relatively pain-free life is what Tenenbaum and Raffman call a "vague" end or project, the irrationality exhibited by the self-torturer takes a "top-down" form. That is, what is irrational is the extended pattern of choice and action taken as a whole, where this does not imply that any of the individual choices or acts that make up that extended pattern are themselves irrational.

A vague project is one such that what counts as success in realizing it is vague. Your project may be to write an interesting and illuminating book on ethics. But we should not expect a very precise account of what would make such a book sufficiently interesting and illuminating to constitute a successful execution of this project. Moreover (and crucially, for present purposes), a vague project is one whose completion requires the performance of certain types of action over time (typing, editing, researching, etc.) but is such that no particular token act is necessary for its completion. To succeed in writing an interesting book on ethics, you will need to do some typing, some editing, some researching, and so forth, but there will not be any particular occasion on which you must type in order to complete the book successfully.

[8] Although plan solutions do arguably offer reasonable advice for avoiding that form of irrationality. They thus offer a viable answer to what I earlier called the "practical" question—they suggest a means for the self-torturer to effectively take advantage of his situation. This perhaps explains part of their appeal, despite their lack of generality.

[9] Tenenbaum and Raffman offer a different objection to plan-based solutions. They imagine the following: someone adopts a plan to stop at, say, setting 25. But once she gets there, she changes her mind and instead decides to stop at 26, which she prefers to 25. Supposing she does stop at 26, they wonder what grounds we could have for convicting this person of irrationality. This is a good question, but it doesn't by itself show that the plan-based approach is wrongheaded. It merely points out what such an account needs to do in order to succeed. See "Vague Projects," p110.

The key to resolving the puzzle of the self-torturer, according to Tenenbaum and Raffman, is to ask what instrumental rationality requires when it comes to vague ends. For given the above properties of vague ends, it may be that one fails to achieve one's end without ever failing to perform some action that was necessary for achieving that end. Thus, one may be instrumentally irrational—in that one fails to (intend to) take means adequate to one's ends—without that irrationality being localized in any of the particular actions one performed or choices one made. That is, the irrationality may be top-down.[10]

Tenenbaum and Raffman argue that, given the plausibility of interpreting the self-torturer as pursuing a vague end, we can view him as exhibiting this top-down form of irrationality. Since, presumably, what will count for him as adequate in terms of physical comfort over the course of his life is vague, there is no particular point at which it is necessary for him to stop advancing the dial. Thus, there it is no point at which his choice to advance the dial is instrumentally irrational, though if he never chooses to stop, he will certainly fail to achieve his end.

Tenenbaum and Raffman's account appears, then, to provide the resources for understanding the irrationality of the self-torturer's advancing to the final setting of the device. In particular, it is well suited to explain how this can be irrational despite the fact that at each decision point the balance of reasons appears to favor accepting the money and increasing the voltage by one setting. For on their view, we need not take any of these individual decisions to be irrational in order to convict the self-torturer of irrationality in advancing to the final setting. This addresses the first aspect of the puzzle I distinguished in section 1.

The other aspect of the puzzle was this: How could it be rational at any point for the self-torturer to refuse the money offered to him in exchange for an imperceptible (or barely perceptible) increase in voltage? Can Tenenbaum and Raffman's solution help us here?

At first blush, it seems it can. They point out that, if it's ever rational to pursue a vague project, then it must sometimes be rationally permissible to choose among one's immediately available options in a way that fails to maximize expected utility when considering that choice on its own.[11] If one were *not* permitted to sometimes act in the service of one's vague projects instead of doing what would at that moment maximize expected utility, one would frequently be unable to avoid (top-down) irrationality with respect to one's vague projects. We have, then, a kind of transcendental deduction of the permission to deviate from the aim of maximizing expected utility at each moment.

[10] Ibid., p. 101

[11] Ibid., p. 102

In Tenenbaum and Raffman's terminology, it may be permissible for an agent to perform an action that fails to maximize expected utility if that act is "generally implicated" in one of the agent's vague projects. An act is generally implicated in a vague project if the act is of a type such that it is necessary to perform some acts of that type if the agent is to succeed in her project, although there is no particular occasion on which the agent's performing an act of that type is necessary for success.[12] The fact that a given option, ϕ, is generally implicated in one of the agent's projects can thus be set against the fact that some other option, ψ, has a greater expected utility, making it permissible to choose ϕ over ψ.

For the self-torturer, although advancing the dial and accepting the money has, at each point, greater expected utility than the immediately available alternative, at some point he will need to refuse the money if he is to lead a relatively pain free life. Refusing the money is therefore generally implicated in that project. It's this that is meant to explain why it's rationally permissible for the self-torturer to refuse the money and stop advancing the dial.

5. The Incompleteness of Tenenbaum and Raffman's Account

The account proposed by Tenenbaum and Raffman is not adequate as it stands, however. Although an option's being generally implicated in an agent's project may in some circumstances help to make it permissible for the agent to go for it, it's clearly not sufficient. There obviously will be circumstances in which one has decisive reason not to choose a generally implicated option.[13] But Tenenbaum and Raffman say little about the conditions under which general implication in a vague project can help to make an option rationally permissible. Without knowing more about these conditions, we don't yet have the resources to resolve Quinn's puzzle.

Imagine, for example, that you've just learned that the nuclear power plant down the road from your office has malfunctioned and is emitting higher than normal levels of radiation into the surrounding area. The authorities have recommended evacuating the area for a few weeks until they can fix the problem. Now, let's assume there are two vague projects at issue for you here. One is to live a long and healthy life. The other is to write a decent article on practical rationality. The levels of radiation aren't that high. Staying in your office to work on the article won't necessarily entirely undermine your ability to live a reasonably long life. But you do have reason to believe that staying behind will reduce the length of your life considerably. On the other hand, evacuating will *not at all* jeopardize your ability

[12] Ibid., p. 104.

[13] As Tenenbaum and Raffman acknowledge. See ibid., p. 105.

to write your article (the deadline is a long ways off; you can resume writing at home with ease, etc.). It seems absurd to allow that it would be reasonable to remain in your office typing away.

Here we have one option—working on a draft of your article—which is generally implicated in one of your projects, but which is such that forgoing this option will not at all affect your success in this project. And we have a second option—relocating to your home office—which you can expect will considerably enhance the extent to which you will successfully realize another of your ends, viz., living a reasonably long and healthy life. It's one thing to claim that one is not required at every moment to maximize expected utility; it's quite another to insist that in this sort of case, you are not required to forgo the option of continuing to work on your draft, despite its being generally implicated in one of your projects.

This, however, raises a crucial question for Tenenbaum and Raffman's account: What is the relevant difference between the options the self-torturer faces at any one time and the options you face as you contemplate whether to flee the radiation or continue to work on your draft? After all, the above example might tempt us to accept a principle like the following—a much more modest principle than one that requires one always to select the option that maximizes expected utility. Assume, first, that one is deciding between two immediately available options that are each implicated in one's projects, and second, that the relevance to one's projects is all that matters to one's choice in the circumstances. It seems plausible that, given these assumptions, if taking the first option will itself substantially increase the level of success one is likely to enjoy with respect to a significant and worthwhile project, whereas, neither performing nor forgoing the second option would make any appreciable difference to the success of any project (and one knows all this), then one should take the first option—one is not rationally permitted to choose the second option.

Now, clearly, some such constraint is needed, as the above example of needlessly exposing yourself to radiation shows. And the modest principle proposed in the previous paragraph is consistent with Tenenbaum and Raffman's argument. The problem is that it seems to direct the self-torturer to accept the money and advance the dial every chance he gets, thereby winding up at the final setting. So, we need a different account of when and how an option's being generally implicated in one's projects bears on the permissibility of performing that option. Tenenbaum and Raffman do not provide us with such an account.

6. Possible Plans and Relevant Alternatives

There is, I believe, something right in the attempt to locate the self-torturer's rational success or failure in the extended series of choices he makes. But Tenenbaum and Raffman leave it mysterious how the self-torturer is supposed to bring the operative principle to bear on any particular decision he faces. This is one respect in which the plan-based approach has an advantage.

Insisting that the self-torturer formulate a plan ahead of time also makes it clear that the self-torturer is to deliberate in such a way that he conceives of his options as including not just the alternatives he faces at any particular point in time but the whole range of settings. Planning ahead requires that he consider which, of this expanded set of options, it would make sense to choose.

The availability of this broader perspective makes it possible to ask, not just whether it would be worth it to increase the voltage by an imperceptible amount in return for $10,000, but whether the additional money that would come with *any* increase in voltage would be worth it, given how much more pain the self-torturer would be in as a result of the increase.[14] Indeed, what generates the puzzle is the discrepancy between our comparative assessments of adjacent settings and our comparisons of settings that are farther apart. It would appear, therefore, to make sense for someone who was rationally deliberating about what to do in this situation to take up such a broader perspective in comparing possible stopping points. And it would, correspondingly, be a mistake for the self-torturer to assess the choiceworthiness of each setting, solely in relation to the shifting baseline of its immediate predecessor, so that the question is always only whether an extra $10,000 is worth the (at most) minute increase in discomfort caused by moving the dial one setting.

It's a feature of the plan-based approach that it offers a means of resisting this sort of mistake. Plan solutions refer the self-torturer back to his original decision concerning where to stop and thereby import the original baseline of 0 and its associated evaluations into the self-torturer's later deliberations. But the plan approach goes wrong in arguing that it's the prior *decision* that's binding, rather than the associated evaluations of relative choiceworthiness. For we've already seen that the reference to a plan previously adopted provides too limited a solution to the puzzle.

So let's ask: why not insist merely that the self-torturer govern himself in light of these underlying evaluations (i.e., evaluations of choiceworthiness relative to all the other possible stopping points), regardless of whether or not he has previously formed any intentions about where to stop? What we're looking for, in effect, is a plan-based solution minus the plan.

[14] That is, for any two settings i and j, i<j, we can ask whether (pain_j – pain_i) would, for the self-torturer, be worth $10,000(j) – $10,000(i).

7. A Solution to the Puzzle

The suggestion, in effect, is that the self-torturer choose *as if* he'd had a plan all along. That is, at a first approximation, the self-torturer should choose to advance the dial from his current setting (whatever it is) to the next only if doing so could have been a step in a plan it would have been rational for him to adopt at the outset.

Before going on to refine this suggestion, let me flag what I take to be the main question about this general idea. What we will need to worry about is whether this solution to the puzzle is unacceptably *ad hoc*. It seems obvious that this type of decision procedure is not normally required of us. It's not clear that it would even make sense in many cases. When deciding whether to take the dog for a walk now, or to make coffee first, I don't need to consult any hypothetical plan I might have adopted at some point in the past. (Which point in the past would that even be?) Rather, it seems, I look to current considerations. I will come back to this concern in the next section, where I will attempt to explain why the special features of the self-torturer's choice situation in particular call for this mode of reasoning.

First, though, I want to get on the table a specific proposal about how the self-torturer ought to approach his choices.[15] The central claim is this: 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.

Under what conditions, then, would it be rationally permissible for the self-torturer to adopt an initial plan that included as a step proceeding from n to n+1? I offer the following proposal, due to Erik Carlson, as a plausible account.[16] The guiding question is whether the sum of accumulated money would be worth the discomfort the self-torturer would feel at a given setting as compared with every previous setting. We are, after all, asking whether it would make sense to plan to move from his initial position to that later position. If there is an intermediate position that is clearly preferable, he should not pass it up.

Carlson argues, first, that if we can assume the self-torturer's preferences between any two settings are determinate (i.e., for any two settings, he

[15] I put this forward somewhat tentatively and primarily as a way to illustrate and flesh out the general idea that there are conditions—which I spell out below—under which we should be guided, in our momentary choices, by consideration of what would have been a reasonable plan to adopt given the available alternatives.

[16] Carlson, "Cyclical Preferences." As noted, Carlson sees this account as, in effect, a contribution to the first stage of a plan-based solution to the puzzle (see section 3, above). Cf. the related proposal Wlodek Rabinowicz makes in discussing the so-called "lawn-crossing" problem in Wlodek Rabinowicz, "Act-Utilitarian Prisoner's Dilemmas," *Theoria* 55, 1 (1989).

either prefers one to the other or is indifferent between them), then he should plan to stop at the highest setting that is preferred to every lower setting. Suppose, for example, that this is setting 300. It doesn't make sense for the self-torturer to plan to go beyond 300, since it would be preferable to stop at some earlier point rather than to proceed to 301. The extra money will not be worth the added discomfort as compared to that earlier setting (whatever it is). The exception to this rule is where the self-torturer is strictly indifferent between 300 and some higher setting. In that case, we may allow that it's reasonable for him to pick either position as his eventual stopping point.

So far, the account assumes that the self-torturer's preferences among the different settings are determinate. It's natural to expect, however, that the self-torturer will be unable to determine whether he prefers or is indifferent between some settings (whether because there is no fact of the matter as to which is preferable, or simply because he is unable to tell). How should we take this into account?

To deal with indeterminacy, Carlson proposes that the self-torturer consider a "filtered" series of settings spaced out at regular intervals, where the intervals are large enough so that at some point there is a determinate preference-reversal.[17] For example, it may be that, if the self-torturer considers the series {0, 50, 100, 150, . . . 950, 1000} he will be able to determine that his situation at, say, 300, is preferable to every lower setting in that filtered series, but that his situation at 350 is definitely dispreferred to some previous setting in the series. Taking the most fine-grained filtered series in which there is a determinate preference reversal, he should then plan to stop at the highest setting which is preferred to every previous setting in that series.

This strikes me as a plausible account. But my aim is not to defend the details. My interest here is in explaining how some such account of what would be a reasonable plan can serve as a basis for the rational assessment and guidance of his choices over time, whether or not he adopted such a plan at the outset.

Let's provisionally take Carlson's proposal on board, then, and see how it applies to the momentary choices the self-torturer makes each week to advance the dial or not. For any n, the question of whether he is rationally permitted to advance to n+1 will then be answered as follows.

i. If there is some setting higher than n, which is such that he would either prefer or be indifferent between the combination of money and discomfort at that setting as compared to the combinations

[17] Ibid., pp. 153–154. The idea of a filtered series comes from Quinn, "Puzzle," pp. 206–207.

associated with each previous setting, then he may proceed from n to n+1.

ii. If he is unable to determine whether there is a setting higher than n that meets condition (i), then he should consider the filtered series that includes n and contains the smallest interval over which he can determine a definite preference-reversal relative to lower settings; if there is in that series a setting higher than n such that he would either prefer or be indifferent between it and each previous setting in the series, then he may proceed to n+1.

iii. Otherwise, he is not rationally permitted to advance the dial from n to n+1.

The proposal, then, is that in choosing between two settings, n and n+1, the self-torturer is to govern himself in accordance with (i)–(iii), regardless of whether he has in fact decided on any plan ahead of time. This offers us a way of making precise the idea that he is to evaluate his options on the basis of their choiceworthiness relative to the whole range of alternative settings, from 0 up through his current position, rather than in relation to his current position alone.

In order to justify this solution, I'll show, first, how it answers to the central theoretical questions raised by the case. I'll then identify the special features of the self-torturer's situation and argue these features do indeed ground the application of the sort of hypothetical-plan-based reasoning I have just appealed to.

8. Evaluating the Proposed Solution

Recall the two main questions that Quinn's puzzle presents us with: What explains the irrationality of always opting for the money, thereby ending up the last setting? And why doesn't the self-torturer have decisive reason, at each point, to proceed to the next setting? If it's the case that the self-torturer should deliberate along the lines suggested, then we can give the following explanation. Were the self-torturer to choose the money at every point, then at some point he must have opted for a setting where the money was not worth the pain as compared to some previous state, which he could have opted for instead. This answers the first question. The irrationality here is explained by his failure to govern himself in accordance with his evaluations (preferences) for overall pain+wealth combinations relative to the alternatives he had when he began.

It follows that it is a mistake for the self-torturer to evaluate the option of proceeding to the next setting solely in terms of its choiceworthiness relative to his current state and then act on that basis. That this is a mistake implies, moreover, that the fact that a given setting is preferable to the previous one does *not* in and of itself provide a decisive reason to choose that higher setting. We thus have an answer to our second question, as well.

It's also clear that the current proposal satisfies the two constraints on an adequate solution to the puzzle I laid out in section 2. First, it is consistent with Non-segmentation. What matters, on this account, is whether it would be rationally permissible for the self-torturer to plan, at the outset, to proceed to a particular setting—say, 800. Obviously, the answer to this question depends on what his alternatives were. If his only other option was to set the dial to 799, then it seems sensible for him to choose instead to advance to 800 and thereby win an extra $10,000. But if, on the other hand, he is free to stop at 300, and he clearly prefers the pain/money combination at 300 to that at 800, it would be irrational for him to plan to proceed all the way to 800. We can therefore explain why it makes a difference that the self-torturer faces a series of choices, rather than just a one-off choice between two adjacent settings.

Second, this hypothetical-plan solution, unlike *actual*-plan-based solutions, is suitably general. Even if the self-torturer fails to form any specific intention at the start about when to stop increasing the voltage, the relevant norm still applies and he should conform his actions to it.

The proposed account thus seems more adequate to addressing the puzzle raised by Quinn's example (and its real-life analogs) than the alternatives. But further justification is clearly needed. In particular, as noted above, the solution implies that the self-torturer's evaluation of an option as choiceworthy relative to earlier states should take priority over his evaluation of that option when compared just with his current state (whatever it is). Why should we accept this requirement? As I've pointed out, it is not always, or even typically, a mistake to evaluate one's options in relation to one's current circumstances alone and without regard to whether they could figure in plans it would have been rational to adopt at some earlier time.

In response, I begin with a conjecture regarding the general rule under which the specific proposal for the self-torturer falls—a conjecture, that is, about the conditions that call for the consideration of whether one's present actions could fit into plans it would have been rational to adopt at some earlier time. I will then argue that the explanation of why these conditions require this mode of reasoning or evaluation is that they are sufficient conditions for holding the agent responsible for the temporally extended course of action covering the relevant stretch of time (at least on the assumption the agent is responsible for anything she does).

9. The Hypothetical-Plan Rule

First, some terminology. Let's say that an agent, S, is in an *epistemic position* at t_1 to adopt a plan about what to do at some point in the future, t_2, if

S can be sufficiently confident about what her circumstances and opportunities for action will be over the period t_1–t_2 in order to rationally form a plan about what to do at t_2. I won't try to say exactly when this is the case. Obviously, how confident one needs to be about what the future will be like, and what specifically one needs to know or believe for purposes of planning, will depend on the content of the plan and its level of detail. I will not need to be very confident that my circumstances and opportunities will be precisely as I now expect them to be in order to rationally adopt a relatively generic plan—for instance, to take a vacation next summer. It seems likely, though, that it would not be rational to adopt a plan if I have no good reason to think that I will have the opportunity to execute it or if I think it's very likely that my circumstances will require me to revise or abandon my current plan. In this case, I won't be in an epistemic position to adopt a plan at the requisite level of specificity.

Here, then, is what I will call the *Hypothetical-Plan Rule:*

If, at time t, the following are true of an agent, S:

a. S is in an epistemic position to make a plan about what to do over the time period t_1–t_n (where t is earlier than, or simultaneous with t_1),

b. among the plans S is in an epistemic position to adopt at t for the period t_1–t_n, there are some that it would be unreasonable to adopt, given the alternatives available,

c. S knows or should know (a) and (b),

AND if, at some later time, t_m ($1<m\leq n$), the following are true of some action, Φ, which S has the option of performing:

d. of the plans mentioned in (b) that are still accessible to S at $t_{m,}$ given what S has done since t, it would have been unreasonable to adopt at t any such plan that included S's Φing at t_m, whereas it would have been reasonable to adopt some still-accessible plan requiring S not to Φ at $t_{m,}$[18]

e. no circumstances that would have been relevant to S's planning at t have changed in ways S could not have reasonably foreseen

THEN (provisionally): S should not Φ at t_m.

This conclusion is only provisional, since there may be conflicts between the requirements of hypothetical plans meeting conditions (a)–(c)

[18] I am grateful to an anonymous reviewer for pointing out the need to make explicit the assumption that there is still at least one reasonable plan accessible to S at $t_{m.}$

for different time periods. Where there is such a conflict, S should give priority to those covering the longest time period for which (a)–(c) hold.

To summarize, then, the idea is that if, at some earlier time, one had enough information to plan for one's current circumstances (and nothing unexpected has come up in the meantime), and if it would at that time have been stupid to plan on performing an action one is now considering, then one should not perform it.[19]

I can now explain why it is only in comparatively rare circumstances that we need to explicitly consider which plans it would have been reasonable to adopt at some earlier time. First, it's often the case that one cannot be expected to know that (a) and (b) obtain with respect to a given time period (often because they don't), at least for plans with a certain degree of specificity. Thus, we are frequently not required to act in accordance with hypothetical plans it would be reasonable to adopt at a particular time in the past. I was not, last week, in an epistemic position to adopt a very definite plan about, for instance, how much time to spend answering email today. So given the choices I face today—for example, whether to put off answering some of the items in my inbox in order to prep for class—the plans I would have made last week will not be relevant. The only such plans I

[19] What if conditions (a)–(c) are met for some time t, and yet, given how S has acted since then, she finds that none of the plans still accessible to her at t_m are ones that would have been reasonable for her to adopt at t. Can the Hypothetical-Plan Rule be extended to cover such a case? I think there is a fairly straightforward extension for cases where it is possible to compare the relevant hypothetical plans as more or less unreasonable. For such cases, we should modify condition (d) and hold that S should not Φ at t_m if it would have been *more unreasonable* for S to adopt at t any of the plans that involve phi-ing at t_m than it would have been to adopt one of the remaining accessible plans that involve S's not Φing at t_m, even if it would also have been unreasonable to adopt that plan. To illustrate, consider the following example (which I owe to an anonymous reviewer). Suppose I am told on Friday that I will be given a large reward if and only if I perform a boring and pointless task on both Saturday and Sunday, and yet I fail to do it on Saturday. Suppose the reward is large enough that the only plan it would have been reasonable to adopt on Friday would be to perform the task on both days. Neither the plan to forgo the task on Saturday but perform it on Sunday, nor the plan to forgo the task on both days would have been reasonable. However, it's plausible that it would be *less* unreasonable to adopt the latter plan on Friday than the former, given the pointlessness of performing the task on Sunday alone. And it does seem clear that, having failed to perform the task on Saturday, one should not bother with it on Sunday.

However, there also seem to be cases where it is not possible to compare the remaining plans still accessible to S at t_m as being more or less unreasonable than one another. This may be true for the self-torturer, for example, should he proceed beyond a reasonable stopping point. (It is not clear to me, for instance, whether it would be more or less unreasonable for the self-torturer to plan at the beginning to proceed to setting 998 as opposed to 999.) In such cases I think we can safely say that the question of which plans it would have been reasonable for S to adopt at t is no longer relevant to what S should do at t_m. Even if this is so, however, we should leave it open that the Hypothetical-Plan rule might apply with respect to some time later than t but earlier than t_m. Thanks to A.J. Julius for a helpful conversation about this issue.

would have been in an epistemic position to make will be too vague to bear on the alternatives I am currently deciding between.

Secondly, in many other cases, although there are hypothetical plans that meet the above conditions (and so bear on my present choices), the considerations that would have been relevant to my earlier planning are exactly the same as the considerations that are relevant to my current situation. So whether I take into account what would have been a reasonable plan or not, my decision will be the same. Consider my earlier example: I have to decide whether to make coffee before or after walking the dog. I'm sure I was in a position last night to form an intention about what to do this morning, and nothing unexpected has occurred. But (unlike in the self-torturer's case) the reasons I now have to make coffee first (it will steel me for the elements, etc.) are just the same as the reasons I would have had last night to plan on making coffee first. In other words, I have no reason to expect that the outcome of applying the Hypothetical-Plan Rule will differ from the outcome of merely attending to the relevant features of my present circumstances and alternatives. Thus, even where conditions (a)–(e) are met with respect to certain options one now faces, one can often be confident that one will conform to the above rule simply by considering one's present circumstances. This, too, helps explain the sense that we're not normally required to consider, explicitly or independently, what decision it would have made sense to make at this or that time in the past.

Still, there will be some cases where the conclusions will conflict with ordinary "from-now-on" reasoning—the self-torturer's situation is a case in point. And it may seem hard to see how, in such cases, reasoning in accordance with the Hypothetical-Plan Rule could make sense. In the self-torturer's case, the proposal that what he should treat as decisive is *not*, as one would expect, his evaluation of his current options in relation to *one another,* but rather his evaluation of those options in relation to alternatives he has already passed by and which are no longer available to him. And this implies that he should treat as still relevant to what he should do now, going forward, gains that are already secure, and costs that are already sunk. Doesn't this amount to some sort of fallacy?

Because the Hypothetical-Plan Rule is thus opposed to a fairly well-entrenched "forward-looking" conception of rational choice, I do not want to rest the entire case for it on its promise for resolving Quinn's puzzle. So, to bolster my case, I will first show that my account of the conditions under which the rule applies also helps make sense of a different type of case of interest to moral philosophy. I will then conclude with a suggestion about how we might ultimately explain the warrant for this form of reasoning.

10. Testing the Rule in Other Contexts

The first thing to note is that the orthodox view that, for instance, it is a mistake to consider sunk costs in making one's decisions is most obvious in cases where there would be no conflict with the Hypothetical-Plan Rule. For example, the advice, "Don't throw good money after bad," will in most cases be consistent with the rule, since presumably it would not have been rational to adopt ahead of time a plan that included sinking additional funds into a failed venture. It's also likely that the circumstances have not turned out as the investor expected they would. In that case, condition (e) would fail to hold in the first place.

Let's turn, then, to a case where applying the Hypothetical-Plan Rule does make a difference and also gets the intuitively correct result. You have made me a promise—say, to pick up some milk on the way home from work. It's a fairly minor promise (the milk is not urgent). Hence, if it became relatively inconvenient to pick it up tonight, this would be sufficient to justify not fulfilling the promise. For instance, if the store that's on your way home happened to close early, this would be excuse enough; you wouldn't have to go out of your way to get the milk.

There is, however, an important exception to this basis for excuse. Suppose you yourself have acted in a way you should have known would make it substantially less convenient to fulfill the promise. In that case the added inconvenience will not so easily get you off the hook. If you decide to go for a beer with friends after work and therefore do not make it to the closest grocery store before it closes, you now have to go to a market that is less convenient for you, but open later. In this case, you ought to go out of your way to keep your promise, though this would not have been necessary had the inconvenience not been your own doing.

As this example shows, what matters morally here is not just your present circumstances vis-à-vis your promise and the costs of fulfilling it. Yet it's not the case that you have *already* done something wrong, which you now need to make up for. You violated no obligation in going out with your friends after work. Other things equal, there is nothing wrong with spending an hour at the pub and then walking the extra distance to the late-night market to get milk, thus keeping your promise. What you may not do is act in a way that you should realize will make it inconvenient to keep your promise, and then appeal to the inconvenience as justification for not fulfilling your promise.

This case appears to confirm the Hypothetical-Plan Rule. First, we can assume that you were in an epistemic position to make a plan before leaving work about what to do that evening, and in particular, about how to keep the promise to get milk. Second, although it would have been reasonable, other things equal, to plan on either (a) skipping the pub and buying milk at the store that is on your way home, or (b) stopping off at the pub and then going out of your way to buy milk at the late-night market, it would not be permissible for you to plan on (c) going to the pub, and then going straight home without the milk. That last plan would simply be a plan to break your promise. Third, it's reasonable to expect you to know all of this. Finally, since you have decided stop at the pub with your friends, you should not now go straight home without the milk. And this is so even if the inconvenience of having to go out of your way to buy milk would have otherwise been sufficient grounds for not fulfilling your promise.

So, we have here another type of predicament (again, not one that is particularly unusual), which appears to call for the hypothetical-plan mode of reasoning and evaluation under the specified conditions. This should go some way toward answering the charge that the solution in the self-torturer's case is problematically *ad hoc.*

11. The Significance of What One is in the Process of Doing

This last example is also helpful for understanding *why* this type of hypothetical-plan reasoning is warranted under the conditions listed above. The reason you cannot appeal to the inconvenience of having to go all the way to the late-night market to excuse your not buying the milk (though in other contexts this would be just the type of consideration that would get you off the hook) is that you are responsible for the fact that it is now less convenient to buy the milk. What this means, here, is that what you need to answer for is not just the decision to go straight home rather than go the extra mile to the late-night market. Rather, you have to answer for the series of choices you made, first, to do something that would prevent you from shopping at the more convenient grocery store and then to go home without the milk. It's this whole package that you are accountable for—i.e., that you need to defend as a justifiable course of action. If you cannot defend this sequence as a whole, then you are not justified in breaking your promise. You should therefore keep it if you still can.

These considerations suggest that the conditions of application for the Hypothetical-Plan rule can be interpreted as (sufficient) conditions for imputing to an agent a temporally extended course of action—or unified sequence of actions—as something he or she is appropriately held responsible for. I'll elaborate this suggestion with reference once again to the self-torturer case.

The suggestion, as it applies to the self-torturer, is that we should think of each decision the self-torturer makes to proceed from one setting to the next as merely a phase in a larger course of action that consists in proceeding from 0 to that later setting. If he advances to 400, we can legitimately hold him responsible for what he has done under the description "advancing the dial from 0 to 400." It's that whole course of action that we can criticize him for, or ask him to defend.

It is because of this that the self-torturer must, for instance, treat his original situation (at setting 0) as relevant in his deliberation, even though he no longer has the option of returning to that situation. If what he is doing is just part of proceeding from 0 to, say, 400—if we can expect him to take responsibility for proceeding from 0 to 400, under that description—then if he is rational, he will only move to 400 if there is good reason for him to prefer the combination of money and discomfort at 400 to his initial impoverished, though physically comfortable, state. If, on the other hand, setting 0 is preferable to setting 400, then he should not advance the dial from 0 to 400. And since, I am suggesting, that is exactly what we should see him as doing (and what he should see himself as doing) in advancing the dial from 399 to 400, he should not in that case proceed from 399 to 400.[20]

The crucial premise is that we can legitimately hold the self-torturer responsible for—or expect him to take responsibility for—the series of choices he makes considered as a unified course of action extended over time, whether or not he intends it under such a description. What makes this legitimate is that he knew (or at least was in a position to know) at the outset that given his predicament there were ways of proceeding it would certainly have been crazy to undertake— and therefore certain plans it would have been irrational for him to adopt. For example, it would not be reasonable to intend to proceed all the way to 1000. And of course this is because it would not be reasonable to *do* that—that is, to proceed from 0 to 1000. But if there is something such that a person knows she should not do it, and also knows or should know that she is doing it (or is about to do it), then, other things equal,[21] she can be held responsible for doing it. And this is so whether or not she specifically intends to be doing it or has it as her aim or end.

Part of what is required of the self-torturer, then, is that he conceive of what he is doing, and of what he might do, in a certain way. And really this shouldn't be surprising. Thinking well about what to do, both morally and prudentially, requires doing a good job of specifying one's options. The fact that a particular course of action one is considering can be brought under a certain description is obviously not enough to show that it is the description most relevant to one's decision. What is more surprising, perhaps, is that in some cases the description of a possible course of action that *is* relevant may make essential reference to what one has done in the past (and not merely because what one has done in the past can make a difference to consequences one's action may have going forward). Grasping this fact about

[20] Cf. Chrisoula Andreou, "The Good, the Bad, and the Trivial," *Philosophical Studies* 169 (2014) for a related suggestion about how certain paradoxes in this area might be resolved by attending to the different levels of description that can apply to what an agent is doing at a particular moment in time.

[21] That is, assuming she is not acting under duress, or hypnosis, or some other responsibility-mitigating influence.

the continuity of our past and present actions is necessary for resolving Quinn's puzzle. What the puzzle helps us see is that we cannot always regard our possibilities for action as if we have just arrived on the scene, our only true concern being what will happen from now on. Rather, at least sometimes, what we have *been* doing matters to how we should understand what we *will be* doing in making the choices we make. And it can therefore matter to which choices we *should* make.[22]

[22] I would like to thank Mark Alznauer, Fabrizio Cariani, Kyla Ebels-Duggan, A.J. Julius, Joshua Knobe, Wlodek Rabinowicz, and an anonymous reviewer for helpful discussion and feedback on earlier versions of this paper.