CANADIAN JOURNAL OF PHILOSOPHY, 2016
VOL. 45, NO. 5–6, 562–575
http://dx.doi.org/10.1080/00455091.2015.1118232

Routledge
Taylor & Francis Group

The real puzzle of the self-torturer: uncovering a new dimension of instrumental rationality

Chrisoula Andreou

Department of Philosophy, University of Utah, Salt Lake City, UT, USA

ABSTRACT
The puzzle of the self-torturer raises intriguing questions concerning rationality, cyclic preferences, and resoluteness. Interestingly, what makes the case puzzling has not been clearly pinpointed. The puzzle, it seems, is that a series of rational choices foreseeably leads the self-torturer to an option that serves his preferences worse than the one with which he started. But this is a very misleading way of casting the puzzle. I pinpoint the real puzzle of the self-torturer and, in the process, reveal a neglected but crucial dimension of instrumental rationality.

ARTICLE HISTORY Received 30 October 2015; Accepted 5 November 2015

KEYWORDS Cyclic preferences; instrumental rationality; intransitivity; puzzle of the self-torturer; resoluteness; vagueness; Warren Quinn

1. Introduction

Warren Quinn's puzzle of the self-torturer raises intriguing questions concerning rationality, cyclic preferences, and resoluteness. The case of the self-torturer is supposed to illustrate that cyclic preferences can be rational and to suggest that, in cases where they are, rationality calls for some form of resoluteness. Criticisms of the case have largely focused on resisting the idea that the case of the self-torturer is a case of rational cyclic preferences.[1] My sense is that the responses to these criticisms by defenders of the puzzle are compelling and that the puzzle really does challenge some traditional assumptions about (instrumental) rationality.[2] But I also think that what makes the puzzle of the self-torturer puzzling has not been properly identified. The puzzle, it seems, is that a series of rational choices foreseeably leads the self-torturer to an option that serves his preferences worse than the one with which he started. But this is a very misleading way of casting the puzzle raised by the case of the self-torturer. My aim in this

CONTACT Chrisoula Andreou c.andreou@utah.edu

paper is to identify the real puzzle of the self-torturer and, in the process, reveal a neglected but crucial dimension of instrumental rationality. I will show that the subjective responses that instrumental rationality is responsive and accountable to are not just the agent's preferences, where preferences can be understood as *relational* appraisal responses, in a sense that will be discussed below. Our subjective responses include appraisals that do not qualify as relational in the relevant sense – appraisals associated with a rational requirement that can, in theory and in practice, justify an agent's sometimes purposely acting against his preference(s) regarding the options among which he must currently choose.[3]

2. Quinn's puzzle of the self-torturer and his proposed resolution

Quinn describes the situation of the self-torturer as follows:

> Suppose there is a medical device that enables doctors to apply electric current to the body in [extremely tiny] increments … . 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*.
>
> [T]he self-torturer cannot feel any difference in comfort between adjacent settings [or at least he cannot, with any confidence, determine whether he has moved up a setting just by the way he feels]…[but] there *are* noticeable differences in comfort between settings that are sufficiently far apart. Indeed, 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. (1993a, 198) [4]

Given the circumstances, the self-torturer finds himself with the following preferences: for every pair of settings n and n + 1, he prefers (the situation at) n + 1 over n; but he also prefers 0 to 1000. His preferences over the settings thus form a loop (as in Figure 1) and are, in this sense, cyclic.

According to Quinn, although it is tempting to dismiss the self-torturer's cyclic preferences as irrational, the preferences seem 'perfectly natural and appropriate given his circumstances';[5] and, given these preferences (and the

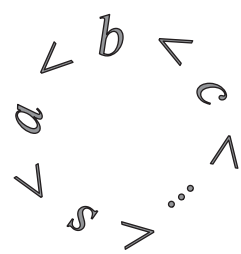

Figure 1. Read 'x < y' as 'y is preferred to x', and 'x < y < z' as 'x < y & y < z'.

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

possibility that the self-torturer might be stuck with them), the self-torturer has 'a real problem of rational choice: How to take reasonable advantage of what the device offers him without ending up the worse for it' (200).

In Quinn's view, it is clear that the self-torturer needs to pick an acceptable stopping point and then resolutely stick to his plan. But this approach is not supported by the prevailing theory of instrumental rationality, which prohibits an agent from 'forgo[ing] something that he would in fact prefer to get, all things considered' (205). Given that, for any setting n between 0 and 999, the self-torturer prefers to stop at setting n + 1 than to stop at setting n, and given that, when the self-torturer is at setting n, stopping at setting n + 1 is still an available option (in the sense that, were the self-torturer to decide that he should stop at setting n + 1 he could), stopping at setting n, is according to the prevailing theory of instrumental rationality, impermissible. Quinn thus rejects the prevailing theory of instrumental rationality in favor of a theory that requires some resoluteness. At the heart of the theory is

> the principle that a reasonable strategy that correctly anticipated all later facts (including facts about preferences) still binds. On such a theory of rationality some contexts of choice fall under the authority of past decisions….In these contexts… *[a]n agent is not rationally permitted to change course even if doing so would better serve his preferences*. (207)

Presented somewhat more formally, Quinn's reasoning in favor of resoluteness can be captured as follows:

P1: The self-torturer's cyclic preferences are rationally permissible.

P2: If the self-torturer's cyclic preferences are rationally permissible and rationality does not involve resoluteness, then rationally-governed choice will lead the self-torturer to an alternative that is worse than the alternative he began with (even if there are no unanticipated developments).

P3: Rationally-governed choice will not lead one to an alternative that is worse than the alternative one began with (at least if there are no unanticipated developments).

C: Rationality involves resoluteness.

3. A complication

As Quinn anticipated, many have had qualms about P1. But even if Quinn is right about P1 and the theorists who oppose P1 are, as Quinn suggests, making things 'too easy on [themselves]' and 'too hard on the self-torturer' (199), Quinn's reasoning seems problematic.

Notice first that, as Quinn makes explicit, his concern is with instrumental rationality. Moreover, putting aside complications he sees as irrelevant in relation to the puzzle of the self-torturer, he does not question, but instead endorses, the prevailing assumption that instrumental rationality 'is and ought to be the slave of the agent's preferences' (209).[6] But, in that case, it seems like 'worse' in

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

P3 must, for the sake of consistency, be interpreted as 'worse in terms of serving the agent's preferences.'[7] This, however, puts P3 in tension with Quinn's endorsement of resoluteness, since, as Quinn understands resoluteness, it sometimes requires an agent to choose an alternative that serves his preferences worse than another available alternative. And if it is sometimes permissible to end up with an alternative that serves one's preferences worse than another available alternative, why would it not be permissible to end up with an alternative that serves one's preferences worse than the alternative with which one started? Relatedly, if it is rationally permissible to make a series of choices that leads one to an alternative that is worse than another alternative that one could have opted for, why would it matter whether the other alternative is the alternative one began with, an alternative that was available after one took some further steps, or an alternative that would be available if one continued to proceed?

Consider, by way of illustration, the following case: Suppose, to borrow an example from (Andreou 2015a), one has access to five cups of tea. '[T]he leftmost tea (tea1) is very hot but not very flavorful, and the rest are such that each is more flavorful but not quite as hot as the one just to the left of it; tea5, on the far right, is very flavorful but also lukewarm' (1).[8] Suppose further that 'one's preferences over the cups of tea (taking into account both temperature and flavor) are cyclic, with tea2 preferred to tea1, tea3 preferred to tea2, tea4 preferred to tea3, tea5 preferred to tea4, but tea1 preferred to tea5' (1–2). Suppose also that these cyclic preferences are rationally permissible. Now assume that one does not possess any of the teas and that (as per Quinn's suggestion that, given rationally permissible cyclic preferences, it is rationally permissible to make a series of choices that leads one to an alternative that serves ones preferences worse than another alternative that one could have opted for) selecting teaN is rationally permissible even though teaN serves one's preferences worse than teaM. Why should the permissibility of selecting teaN change if, instead of it being the case that one never previously possessed any of the teas, the scenario is such that one was initially given teaM? In both scenarios, one selects teaN even though teaM is available. Why should the fact that one was initially given teaM change the permissibility of ending up with teaN?

My point, in short, is that, as soon as it is granted that, in cases like the case of the self-torturer, one is rationally permitted, indeed rationally required, to stick with an option even though it serves one's preferences worse than another available alternative, then it seems ad hoc to insist that rationality does not permit a series of choices that leads one to an option that serves one's preferences worse than the alternative one began with.[9]

4. The real puzzle of the self-torturer

What shall we say, then, about the case of the self-torturer? Well, if we grant, as I will, that it is rationally permissible for the self-torturer to end up with an

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

alternative that serves his preferences worse than some other alternative he could have opted for, then, given my reasoning in the previous section, we cannot just assume that rationality prohibits the self-torturer from making a series of choices that leads to an alternative that serves his preferences worse than the alternative he began with.[10] But then we've lost our apparent reason for thinking there is something irrational about the self-torturer's proceeding at each point and thus going to 1000.[11] And this is puzzling, since, intuitively, there is something irrational about the self-torturer's proceeding at each point and thus going to 1000. I turn now to identifying where the irrationality in this scenario lies. (As will become apparent, one can identify where the irrationality in this scenario lies, without pinpointing or even supposing that there is a specific setting at which the self-torturer rationally should stop. There may instead be a fuzzily bounded range of rationally acceptable stopping points, with no clear first rationally unacceptable stopping point. I say more concerning the presumed vagueness in the self-torturer's situation below.)

In a nutshell, the irrationality in the self-torturer's proceeding at each point and thus going to 1000 lies not in the fact that the self-torturer ends up with an alternative that serves his preferences worse than the one he started with, but in the fact that (though unimpaired by any lack of information about his situation) the self-torturer ends up with a terrible alternative when non-terrible alternatives are available. Now to explain.

My explanation relies on a distinction that is drawn from David Papineau's work on color perception – the distinction is between 'categorical responses' and 'relational responses.' I begin with a review of the relevant points, which borrows freely from my discussion of the distinction, and a variation on it, in my previous work.[12]

In 'Can We Really See a Million Colours?' Papineau (2015) argues that 'our conscious colour experience is the joint product of two different kinds of perceptual state' (277): via one state, we have *categorical* color responses, wherein we experience a surface as of a certain color, say cN, where cN is among the finite set of distinct conscious visual color experiences {c1, c2, c3, …cS} the perceiver can have; via the other state, we have *relational* color responses, wherein we experience adjacent color samples as either the same or as in some way different from one another.[13] As Papineau explains, his position has interesting implications concerning the interpretation of color discrimination data. Consider, for example, the view that 'human beings are capable of well over a million different conscious visual responses to coloured surfaces' (274). This view is based on (1) evidence that, when comparing pairs of color samples, humans can consciously register color differences between more than a million different samples, and (2) the assumption that 'our consciously registering a *difference in colour* must derive from our first having *one* colour response to the left-hand side surface, and *another* colour response to the right-hand [side] surface, and thence registering that there is a difference' (274). But if, as Papineau argues, 'the detection of colour differences between adjacent surfaces does not [always] derive from

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

prior [independent] responses to each surface, there is no need to posit a million such responses to account for the discrimination data' (275). For there is then room for the visual system to issue 'a relational judgement that two adjacent samples … *differ* even in cases where the two surfaces produce [the *same* conscious visual experience, and so] the *same* categorical colour response [when viewed each on its own]' (278); moreover, there is, as Papineau makes clear, room for the possibility that one's conscious visual experience when viewing one pair of color samples, say, sample 23 next to sample 24, can be the same as one's conscious visual experience when viewing a different pair of samples, say sample 27 next to sample 28.

As Papineau emphasizes, it is 'entirely consistent' with his view that categorical color responses can vary from person to person (276). Whereas I might have the same conscious visual experience when I view color sample 2 (by itself) and when I view color sample 3 (by itself), you might have a different conscious visual experiences when you view color sample 2 (by itself) than you have when you view color sample 3 (by itself). There is certainly room for

> variations in culture, training, and natural endowment [to] make a significant difference to the repertoire of [categorical] colour responses available to different individuals. Maybe some individuals are … only capable of a few dozen such responses, while others—painters or interior decorators, say—are capable of many hundreds. (276)

Whether Papineau's position concerning color perception is correct is not something we can or need to take up here. What is important for my purposes is that Papineau's distinction between categorical responses and relational responses, or rather the related distinction between *categorical appraisal responses* and *relational appraisal responses*, can be used to illuminate the nature of instrumental rationality and the puzzle of the self-torturer. Notice first that in appraising an alternative, I might respond *categorically* with something like 'X is terrible' or I might respond *relationally* with something like 'X is worse than Y.' The first sort of response is *categorical* in the sense that it indicates the appraisal category that I see X as falling in. The second sort of response provides no such category information. To appraise X as worse than Y leaves completely open the question of what category I place X in on the spectrum from, say, terrible to fantastic. It indicates only how I appraise X and Y in relation to each other.[14]

Note that, in the sense of interest here, to say that a response is a *categorical* appraisal response is not to say that it is or purports to be objective. My appraisal of the taste of vegemite as terrible counts as a categorical appraisal response even though my appraisal is, I grant, thoroughly subjective. Note also that, in the sense of interest here, to say that a response is a *categorical* appraisal response is not to say that there were no comparisons or contrasts in play when the response occurred. I might find a piece of chocolate terrible-tasting because I am used to very high-end chocolate. Still, 'this chocolate tastes terrible' says something about where I place (the taste of) this chocolate on the spectrum from, say, terrible to fantastic (and so about whether my culinary experience is

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

positive, negative, or neutral); the judgment 'this chocolate tastes worse than the chocolate I had yesterday' does not, in itself, provide any such information.

As in the color case, once one distinguishes between categorical and relational responses, there is room for scenarios such as the following: K is capable of both categorical responses and relational responses with respect to appraisals of a particular type in a particular domain or over a particular set of options; the number of distinct categorical responses K has along the most refined spectrum of categorical responses available to her for appraisals of the relevant type in the relevant domain or over the relevant options is finite; and, even when K uses the most refined spectrum of categorical responses available to her for appraisals of the relevant type in the relevant domain or over the relevant options, K sometimes has relational responses that prompt her to discriminate between alternatives that she has the same categorical response to when she considers each on its own. To take a concrete case, there is room for scenarios such as the following: K is capable of categorical and relational responses concerning the goodness (to K) of various samples of chocolate; the set {terrible, very bad, bad, fair, good, great, fantastic} figures as the most refined spectrum of K's categorical appraisal responses concerning the goodness (to K) of various samples of chocolate; in considering two chocolate samples, say A and B, K has the same categorical appraisal response when she considers each on its own, and yet she also has a relational response of the form 'A is worse than B.' As Papineau emphasizes, there is no guarantee that our categorical responses and our relational responses will prompt the same discriminations.[15]

Note that to say that the number of distinct categorical appraisal responses an agent has in a particular domain or over a particular set of options is finite is not to say that every alternative the agent considers will fall squarely into one appraisal category or another. The possibility of vagueness, understood as involving fuzzy boundaries, is by no means ruled out. Quinn casts the puzzle of the self-torturer as involving vagueness, and I will not here question the phenomenon or its role in generating the puzzle and supporting the possibility of rationally cyclic preferences.[16] My aim, recall, is to argue that, even if it is true that, given the self-torturer's situation, the self-torturer's subjective appraisals, as Quinn describes them, are rationally permissible, we need to rethink the puzzle of the self-torturer and the challenge it raises for the conception of instrumental rationality according to which an instrumentally rational agent always chooses in accordance with her preference(s) regarding the options among which she must currently choose.

It is important for my purposes that an agent with cyclic preferences (such as the self-torturer) can be led, over a *series* of steps guided by his preferences, from a certain option to one that is determinately in a lower (appraisal) category. Notice, however, that, given the possibility of vagueness, there is no need to suppose that, at some point along the way, the agent must have a preference that prompts him to swap his current option for one that is determinately in a lower category. For, roughly put, given fuzzy boundaries, the transition, over

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

a *series* of steps, from a certain option to one that is determinately in a lower category can occur without there being any single step in the agent's preference loop that takes him from his current option to one that is determinately in a lower category; it can thus occur without there being, at some point along the way, a preference favoring an option that is determinately in a lower category.

My discussion so far suggests that, if an agent's preferences are cyclic, one should not expect to find a thoroughly tidy relation between the agent's categorical responses to the alternatives she faces and the agent's relational responses to the alternatives she faces; or at least this is so if (i) her relational responses are understood as capturing her (pairwise) preferences between the available alternatives, and (ii) the categorical responses that are relevant in the case at hand, even if they are complicated by vagueness, involve categories that can be arranged from lowest to highest (as with the categories *terrible*, *very bad*, *bad*, *fair*, *good*, *great*, *fantastic*). However, it might be claimed that if an agent's preferences over a set of options are cyclic, then the only categorical responses she can have to the options will be such that the relevant categories cannot be arranged from lowest to highest, but instead form a loop. The case of the self-torturer speaks against this view. Although the self-torturer's preferences over his options are cyclic, the self-torturer can and does have categorical responses like 'that would be a terrible result' and 'that would be a fantastic result'; and he does *not* see the spectrum from *terrible* to *fantastic* as forming a loop so that talk of higher and lower appraisal categories is out of place – to the contrary, it is precisely because talk of higher and lower appraisal categories seems perfectly in order in the case of the self-torturer that it is plausible to suggest that the self-torturer should not end up with options in some of the available categories. (More on this below.) If the case of the self-torturer were such that talk of higher and lower appraisal categories were out of place, it is far from clear that we could substantiate the claim that some of the options in the case ought to be avoided. It is the combination of cyclic preferences and non-cyclic categories that makes the case particularly interesting.

It might be suggested that, insofar as the categories associated with a set of categorical responses can be arranged from lowest to highest, we can say that the agent has preferences over the categories, and that *these* preferences are not cyclic. For example, in the case of the self-torturer, we can say that the self-torturer has preferences over the categories in the spectrum from *terrible* to *fantastic*, and that *these* preferences are not cyclic. I will not here delve into this suggestion. I want only to emphasize that it in no way undermines the idea that the self-torturer's pairwise preferences between the options he actually faces are cyclic. Moreover, it does not support the idea that *all* of the self-torturer's subjective appraisal responses are preferences (understood as subjective *relational* appraisal responses). To say that the self-torturer prefers the category *fantastic* to the category *terrible* is to say that the self-torturer prefers options to which he has the subjective categorical appraisal response 'this is fantastic'

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

over options to which he has the subjective categorical appraisal response 'this is terrible'; subjective *categorical* appraisal responses (and the valences they convey) remain in play.

Now consider the following proposal, which is related to P3 (in my reconstruction of Quinn's reasoning).

P3*: Rationally-governed choice will not lead one to an alternative that is (determinately) in a lower appraisal category than another available alternative (at least if there are no unanticipated developments and the set of appraisal categories is finite).

Notice that P3* applies only when talk of higher and lower categories is in order, and so only when the categories in play do not form a loop.

P3* is, I think, quite plausible and it can accommodate the idea that it is rationally impermissible for the self-torturer to end up at 1000 without dismissing the self-torturer's cyclic preferences as rationally impermissible.[17] Some might see P3* as going further out on a limb than necessary relative to the case of the self-torturer, and favor instead the following more modest proposal:

P3': Rationally-governed choice will not lead one to a terrible alternative when an alternative that is (determinately) in a higher appraisal category is available.[18]

I should thus note that, while I will focus on P3*, the gist of my reasoning below holds even if P3* is replaced with P3' (and P2* in the argument below is altered accordingly).

Insofar as rational cyclic preferences are possible, P3* implies that instrumental rationality does not always endorse following one's preferences regarding the options among which one must currently choose, even if these preferences are rationally permissible. (Note that P3* is consistent with the possibility that rationality allows the agent to use her discretion in terms of deciding where exactly to deviate from her preferences, so long as the result conforms to P3*.) Preferences are relational responses. If the self-torturer had nothing but the relational responses that Quinn describes and these responses were rationally permissible, then there would be no way to show that it is irrational for the self-torturer to end up at 1000. But his subjective appraisal responses also include appraisal responses of the form 'alternative X is terrible,' and instrumental rationality is also accountable to these responses. From here, we can get to an internally consistent argument that fits with the spirit of Quinn's resolution of the puzzle of the self-torturer, though Quinn himself failed to properly identify the problem or its resolution:

P1: The self-torturer's cyclic preferences are rationally permissible.

P2*: If the self-torturer's cyclic preferences are rationally permissible and rationality invariably requires one to act on one's preferences/relational appraisal responses regarding the options among which one must currently choose, then rationally-governed choice will lead the self-torturer to an alternative that is in a lower appraisal category than another available alternative (even if there are no unanticipated developments).

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

P3*: Rationally-governed choice will not lead one to an alternative that is in a lower appraisal category than another available alternative (at least if there are no unanticipated developments and the set of appraisal categories is finite).

C*: Rationality does not invariably require one to act on one's preferences/relational appraisal responses regarding the options among which one must currently choose.

In the case of the self-torturer, the agent's categorical appraisal responses are such that some alternatives count as *terrible* and some do not – indeed, some may count as *good*, *great*, or even *fantastic*. As such, it is not rationally permissible for the self-torturer to end up with a terrible alternative. Notice that it need not be that, in all cases of cyclic preferences, the agent's categorical appraisal responses to the available alternatives fall in different categories. Since relational responses can prompt discriminations that are not prompted by the agent's categorical responses, it may be that an agent's relational responses to a set of alternatives reflect cyclic preferences even though her categorical responses place them all in the same category, say *fair*. Recall the tea case, wherein the leftmost tea (tea1) is very hot but not very flavorful, and the rest are such that each is more flavorful but not quite as hot as the one just to the left of it; tea5, on the far right, is very flavorful but also lukewarm. It may be that the agent's preferences over the teas are cyclic, even though she counts all the teas as *fair*. If so, and if the cyclic preferences are rationally permissible, then we can see why it can be rationally permissible for her to end up with any of the teas, and why it doesn't matter which one she started with. Given that the agent's preferences over the teas are rationally cyclic, instrumental rationality cannot forbid the agent from ending up with a tea that is dispreferred to another available tea; it can require that the agent not end up with a tea that falls in a lower category than another available tea, but when all the teas fall in the same category, this doesn't occur no matter which tea she ends up with.

Notice that my reasoning leaves room for the possibility that, when an agent's preferences over a set of options are not cyclic, rationality may, in that case, require that the agent act on her relational responses/preferences regarding the options among which she must currently choose, even if she has the *same* categorical response to all the options. As such, it does not follow from my reasoning that an agent need only attend to her categorical appraisal responses.

5. Conclusion: the moral regarding instrumental rationality

The traditional conception of instrumental rationality combines the idea that instrumental rationality is grounded in our subjective appraisal responses with the assumption that our preferences, understood as relational appraisal responses, exhaust our subjective appraisal responses; but, in addition to our relational appraisal responses, we have subjective categorical appraisal responses. It is precisely when the latter responses are in play that it can be irrational to end up with

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

some alternatives but not others even if one's preferences are rationally cyclic. Without categorical appraisal responses, any alternative in a preference loop would be just as rationally permissible as any other. With categorical appraisal responses, this need not be so. Relatedly, for some alternatives, ending up with that alternative can be rationally impermissible regardless of whether or not the alternative serves the agent's preferences worse than the one that the agent started with or whether or not the agent ended up there as a result of deviating from a prior plan; it can be impermissible because it is in a lower appraisal category than another alternative that the agent could have opted for. The moral, in short, is that, the subjective responses that instrumental rationality is responsive and accountable to are not just the agent's preferences. Our subjective responses also include appraisals that do not qualify as relational in the relevant sense – appraisals associated with a rational requirement (P3*) that can, in theory and in practice, justify an agent's sometimes purposely acting against his preference(s) regarding the options among which he must currently choose.

Notes

1. See, for example, (Voorhoeve and Binmore 2006) and (Arntzenius and McCarthy 1997).
2. For a recent discussion and defense of the puzzle, see (Tenenbaum and Raffman 2012).
3. As will become apparent, the justification can hold even if, as in the case of the self-torturer, there is no threat of the agent ending up with the alternative he started with minus repeated transaction costs. Otherwise put, the justification can hold even if the scenario is such that, once an alternative is passed up, it cannot be regained for a price, and so the problem at issue is not the problem of the agent being 'money-pumped.' For the original presentation of 'the money pump argument,' see (Davidson, McKinsey, and Suppes, 1955). For my critique of the view that the money pump argument establishes that cyclic preferences are irrational, see (Andreou 2007). There, I argue that what the money pump argument shows is that an agent should not always follow his preferences regarding the options among which he must currently choose, even if these preferences are basic and the agent finds that he is stuck with them even after he is fully informed. But because the problem at issue in the case of the self-torturer is not the problem of being money-pumped, we need a different justification for why the self-torturer should not always follow his preference(s) regarding the options among which he must currently choose.
4. Quinn himself does not add the qualification 'or at least he cannot, with any confidence, determine whether he has moved up a setting just by the way he feels,' but--for reasons that I will not get into, because they are complicated and tangential given my purposes in this paper--I think the qualification is helpful.
5. If the self-torturer's preferences are in order, then the case of the self-torturer qualifies as a 'spectrum case' supporting the intransitivity of '___ is rationally preferred to ___.' For extensive discussion of spectrum cases and intransitivity, see (Temkin 2012).
6. In 'Putting Rationality in its Place,' Quinn suggests that instrumental rationality is 'mere cleverness' and not a 'real virtue' of practical rationality if one's practical reasoning is not constrained by good ends (1993b, 234).

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

7. Interestingly, Quinn, at one point, maintains that 'better than... is transitive' (199). But, if 'better than' is understood as (something like) 'better in terms of serving the agent's preferences,' it is not clear that Quinn is entitled to maintain that 'better than' is transitive while also maintaining that the self-torturer's preferences are genuinely and rationally cyclic. And if 'better than' is not understood in terms of the agent's preferences, it is not clear that Quinn is entitled to assume that the relation is relevant to instrumental rationality, given his endorsement of the prevailing assumption that instrumental rationality 'is and ought to be the slave of the agent's preferences' (209). My aim of 'uncovering a new dimension of instrumental rationality' in this paper may ultimately be of help here, but the issue is complicated and so I am working it out in a separate manuscript on the 'better than' relation (in progress).
8. The Online First version of the article, which is the version currently available, is not officially paginated, but I have added page numbers for convenience.
9. In 'Intransitive Preferences, Vagueness, and the Structure of Procrastination,' Duncan MacIntosh argues that 'if the self-torturer really has intransitive preferences… he rationally should proceed to the maximum level' (2010, 73). Relatedly, he claims that, for an agent with intransitive preferences,

> each position he could have been in is such that if he does not move to a different position, he is pair-wise worse off. So, he would have been irrational to stay where he was. In moving, he has not made himself any worse off than he was before. (76)

I disagree with MacIntosh's reasoning, but my concerns about Quinn's take on the puzzle of the self-torturer have been influenced by MacIntosh's thought-provoking challenges concerning the assumed irrationality of the self-torturer's proceeding to 1000.
10. Quinn's suggestion that it is rationally permissible for the self-torturer to end up with an alternative that serves his preferences worse than some other alternative he could have opted for is, of course, controversial. Although defending the suggestion is beyond the scope of this paper, I here accept it as plausible enough to be worth taking on board, at least for the sake of argument.
11. Keep in mind that, since that the self-torturer's preferences are cyclic, we cannot say that his going to 1000 appears far lower in his ranking of his options (and is in this sense much less preferred) than the option of stopping at 0.
12. The discussion I am borrowing from in the next several paragraphs appears in (Andreou 2015b); there Papineau's distinction is used to illuminate the notion of parity.
13. Raffman (1994) raises this possibility and uses the distinction to argue that two color patches that are seen as belonging to different categories when judged singly can be seen as belonging to the same category when judged pairwise. This is in turn used to 'explain, in an *intuitively* compelling way, how a difference in kind can obtain between the endpoints (among others) of an effectively continuous series' and thus resolve the paradox in sorites cases (43). Quinn's puzzle incorporates the assumption that, whatever the explanation, a difference in kind can obtain between the endpoints of an effectively continuous series. (More specifically, Quinn assumes that someone can go from no pain to excruciating pain via a series of unnoticeable or barely noticeable differences.) I will make the same assumption without committing to any particular explanation (though I do find Raffman's explanation plausible).

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press

14. Note that, although in the case of the self-torturer, the focus is on the consequences of the available alternatives, there is nothing in the idea of an appraisal response that requires that appraisal responses to potential actions be consequence-oriented; relatedly, there is, for all I say here, room for appraising an action as terrible even if it does not have terrible consequences.
15. I say a great deal more about this and consider potential objections in (Andreou 2015b).
16. For some forceful argumentation suggesting that vagueness is not crucial for supporting the possibility of rationally cyclic preferences, see (Temkin, 2012, chapter 9).
17. Relatedly, P3* figures as a plausible initial response to the worry, raised by Temkin (1996), that, given the pervasiveness of intransitivity, there may be 'no rational basis for choosing between virtually any alternatives' (209). But Temkin seems more open to the possibility of rational dilemmas than Quinn, and so Temkin may not see P3* as supporting C*, but may instead cast P3* as ensuring, in coordination with the negation of C* (and assuming that one's preferences are rational), that we are in a rational bind. As indicated above, I have accepted, at least for the sake of argument, Quinn's view that it is rationally permissible for the self-torturer to end up with an alternative that serves his preferences worse than some other alternative he could have opted for.
18. Thanks to Sarah Stroud for pointing out that I could make do with this more modest proposal.

Acknowledgments

My thanks for helpful comments from Donald Bruckner, Matthew Frise, Preston Greene, Elijah Millgram, Michael Morreau, Doug Portmore, Theron Pummer, Andrew Reisner, Jonah Schupbach, Jacob Stegenga, Sarah Stroud, Christine Tappolet, Larry Temkin, Sergio Tenenbaum, Mariam Thalos, Ralph Wedgwood, Mike White, participants of the University of Wisconsin-Madison workshop for this special issue on *Belief, Action, and Rationality over Time*, participants at the 2015 work-in-progress workshop at the Philosophy Institute at the University of Saarlandes, students in my PHIL 4010 and PHIL 7500 courses, two anonymous referees, and audience members at my presentations at CRE at the University of Montreal, at the 2015 Pacific Division APA meeting, and at the 2014 Society for Applied Philosophy meeting. Thanks also to Arif Ahmed, Doug Portmore, and Sergio Tenenbaum for thought-provoking discussion (via PEA Soup) on interpreting Quinn's position. Finally, I am grateful to the College of Humanities at the University of Utah for a travel grant supporting my presentation of the ideas in this paper.

References

Andreou, Chrisoula. 2015a. "Cashing out the Money-pump Argument." *Philosophical Studies* doi:10.1007/s11098-015-0555-5.

Andreou, Chrisoula. 2015b. "Parity, Comparability, and Choice." *Journal of Philosophy* 112: 5–22.

Andreou, Chrisoula. 2007. "There Are Preferences and Then There Are Preferences." In *Economics and the Mind*, edited by Barbara Montero and Mark D. White, 115–126. London: Routledge.

Arntzenius, Frank, and David McCarthy. 1997. "Self Torture and Group Beneficence." *Erkenntnis* 47: 129–144.

Davidson, Donald, J. McKinsey, and Patrick Suppes. 1955. "Outlines of a Formal Theory of Value." *Philosophy of Science* 22: 140–160.
MacIntosh, Duncan. 2010. "Intransitive Preferences, Vagueness, and the Structure of Procrastination." In *The Thief of Time: Philosophical Essay on Procrastination*, edited by Chrisoula Andreou and Mark D. White, 68–86. New York: Oxford University Press.
Papineau, David. 2015. "Can We Really See a Million Colours?" In *Phenomenal Qualities*, edited by Paul Coates and Sam Coleman, 274–297. New York: Oxford University Press.
Quinn, Warren. 1993a. "The Puzzle of the Self-Torturer", In *Morality and Action*, edited by Philippa Foot. 198–209, Cambridge: Cambridge University Press.
Quinn, Warren. 1993b. "Putting Rationality in Its Place". In *Morality and Action*, edited by Philippa Foot. 228–255, Cambridge: Cambridge University Press.
Raffman, Diana. 1994. "Vagueness without Paradox." *The Philosophical Review* 103: 41–47.
Temkin, Larry S. 2012. *Rethinking the Good*. Oxford: Oxford University Press.
Temkin, Larry S. 1996. "A Continuum Argument for Intransitivity." *Philosophy & Public Affairs* 25: 175–210.
Tenenbaum, Sergio, and Diana Raffman. 2012. "Vague Projects and the Puzzle of the Self-torturer." *Ethics* 123: 86–112.
Voorhoeve, Alex, and Ken Binmore. 2006. "Transitivity, the Sorites Paradox, and Similarity-based Decision-making." *Erkenntnis* 64: 101–114.

https://doi.org/10.1080/00455091.2015.1118232 Published online by Cambridge University Press