Forthcoming in *Ethics*

Vague Projects and the Puzzle of the Self-Torturer[*]

Sergio Tenenbaum and Diana Raffman

In this paper we advance a new solution to Quinn's puzzle of the self-torturer. The solution falls directly out of an application of the principle of instrumental reasoning to what we call "vague projects", i.e., projects whose completion does not occur at any particular or definite point or moment. The resulting treatment of the puzzle extends our understanding of instrumental rationality to projects and ends that cannot be accommodated by orthodox theories of rational choice.

In Warren Quinn's notorious puzzle of the self-torturer,[1] a person has agreed to wear a device that delivers a constant but imperceptible electric shock. She is then offered the following tradeoff: she will receive a large sum of money—say, $100,000—if she agrees to raise the voltage on the device by a marginal, i.e., imperceptible or just barely perceptible, amount. She knows that she will be offered this same tradeoff again each time she agrees to raise the voltage. It seems that, at each step of the way, the agent should and would raise the voltage; after all, each rise in voltage makes at most a marginal difference in pain, well worth a gain of $100,000, so it would be irrational to do otherwise. But of course in so doing, she would eventually find herself in unbearable pain, and would gladly return all of the money, even pay some in addition, to be restored to the initial setting, at which she was poor but pain free. Thus the self-torturer appears to face a dilemma: no matter which choice she makes—continue indefinitely or stop at some point—her action seems irrational, or leads quickly to a state of affairs that no rational agent would accept.

A number of solutions to the puzzle of the self-torturer have been proposed. We cannot canvass all of them, but we will examine some representative ones and find that they either reject very intuitive assumptions about acceptable preferences, or else impose rational requirements that cannot be otherwise justified. We believe that a different approach is needed. Instead of focusing narrowly on the puzzle, we look more broadly at the nature of what we call "vague projects" or "vague ends". We show that a solution to the puzzle falls directly out of an application of the principle of instrumental rationality to vague projects or ends. This solution preserves all of the intuitive assumptions that seemed to generate the puzzle, yet it employs only a minimal set of normative principles. We believe that this solution to the puzzle is of wider interest; in particular, it extends our understanding of instrumental rationality to ends and projects that cannot be accommodated by orthodox theories of rational choice.

1. The Puzzle, in Greater Detail

Suppose that the shocking device is worn on a strap around the wrist, like a watch.[2] Instead of a watch face, it has a dial with a large number of ordered settings $a_0 \ldots a_n$, on which higher settings correspond to higher voltages; and we assume that severity of pain varies with voltage level. The self-torturer (hereafter, "ST") can raise the voltage by turning the dial. The difference in voltage between adjacent settings of the dial is very small. In fact, Quinn assumes it is so small that ST cannot discriminate between the pain levels she experiences at adjacent settings. Nevertheless, she experiences significant differences in pain between some settings that are farther apart. In particular, she feels no pain at the initial setting a_0, when the device is turned off, but feels unbearable pain at some higher setting a_n.

Initially the device is set at a_0 and ST is presented with the following choice:

(1) Stay at a_0 and be paid nothing.

(2) Raise the voltage to the next setting (a_1) and receive $100,000.

If she decides to raise the voltage, ST is then presented with the same choice *mutatis mutandis*. These are the only choices available to her; in particular, ST never has the option to return to a previous, lower setting. Moreover, the device will remain on her wrist forever, or at least for a very long time. She knows this, and knows also that each choice situation in which she turns the dial will be followed by another just like it.

We assume the following about ST's preferences:

(a) ST prefers more money over less money and less pain over more pain.

(b) There is some setting of the voltage level a_j such that for any setting a_k where $k \geq j$, and for any monetary reward, ST prefers a_0 over the conjunction of (i) remaining for a long period of time at setting a_k, and (ii) getting the associated monetary reward.

(c) No other preferences of ST are relevant for these choice situations.

Now we can generate the puzzle. Given that adjacent voltage levels feel the same to ST, the only relevant preference in each case is a preference for more money. Hence, since she is paid $100,000 to move to the next setting, she should do so in each case. It follows that insofar as ST is rational, she will accept the monetary payoffs all the way up to the last, highest setting. At the same time, according to (b), she will eventually reach a stage so painful that she would prefer to return to a_0. ST's preferences are thus nontransitive (and cyclical), even though they seem entirely ordinary: most people prefer more money over less, and most people would not accept *any* amount of money to be tortured for a long period of time. The result is that seemingly ordinary preferences fail to meet a seemingly fundamental constraint on rational choice—*viz.*, that one's preferences be transitively ordered.[3]

2(a). Self-Torture as a Challenge to Orthodox Rational Choice Theory[4]

We can interpret the puzzle of the self-torturer in either of two ways: (1) as a challenge to orthodox rational choice theory,[5] or (2) as arising simply from certain intuitive judgments. Let's begin by considering these two interpretations in turn.

It is important to understand exactly how the case of ST is supposed to challenge rational choice theory (RCT). The puzzle is not meant to expose any internal incoherence in RCT. Orthodox RCT says only that the *self-torturer* is irrational: since her preferences are nontransitive, we cannot match them to a utility function. Rather, the trouble is supposed to be that such a charge of irrationality is counterintuitive: the self-torturer does not seem irrational. What could be irrational about forming preferences so that one accepts monetary tradeoffs for all imperceptible increases in voltage, but not for at least some larger increases? More precisely, the challenge appears to come from the conjunction of three plausible claims:

(1) Were ST confronted with any of the choices in the series in isolation, i.e., with any single pair of adjacent settings, she would always choose to accept more money. (Thus she exhibits a behavioural preference for more money at each choice situation).

(2) At least one stage in the series, a_k, is such that ST would prefer the initial stage of the series, a_0, over a_k. In fact, she would willingly pay a premium in order to return from a_k to a_0. (Thus she exhibits a behavioural preference for relief from the pain experienced at a_k over retaining her financial gains all the way up to a_k.)

(3) Considered either singly or jointly, the behavioural preferences described above are not irrational.

Orthodox RCT cannot accept all of (1)-(3); taken together, these assumptions amount, or at least seem to amount, to endorsing the rationality of a nontransitive set of preferences.

In one sense, (1) and (2) are non-negotiable, since we can simply stipulate that some agent behaves in this manner. We suppose that, in general, solutions that retain orthodox RCT will reject (3), and are likely to take one of three forms:

A. The *Obstinate* Solution: ST's preferences are nontransitive; this by itself shows that ST is irrational, and no further explanation is needed. Consequently, ST must revise her preferences so as to make them conform to the axioms of RCT.

B. The *ST is a Freak* Solution: ST's preferences do not correspond to ordinary preferences.[6] They seem like ordinary preferences, but we are just confused about this.

C. The *ST Is Confused* Solution: although ST's preferences do correspond to ordinary preferences, they result from bias and/or illusion and thus are formed in an irrational way.[7]

An unqualified obstinate solution is plainly unsatisfactory: we want to understand what's wrong with evaluating outcomes the way ST does, so it doesn't help to be told simply that it must be wrong. Any satisfactory solution must explain why these preferences *seem* rational even if they're not. Also, note that B-type solutions need to do more than show that ST's preferences are abnormal; a freak might be perfectly rational, after all. They need to show that the "freakish" formation of ST's preferences justifies, or at least makes plausible, a verdict of irrationality. Alternatively, strategies (A-C) might be combined in some way; for instance, one might argue that ST's preferences differ somewhat from ordinary preferences and that the difference results from a bias or illusion in reasoning.

Let's look more closely at an orthodox solution, and see the sorts of difficulties it faces. An especially interesting approach that preserves RCT is proposed by Frank Arntzenius and David McCarthy.[8] According to their view, the assumption that adjacent settings of the device are indiscriminable is incorrect, and once we notice this we can see that ST's preferences are indeed irrational. Arntzenius and McCarthy propose the following modified version of the thought experiment:[9]

Self-Torture 2

> The setup is the same as Quinn's, except that ST has a trial period in which she experiments with various settings of the device in different orders. Each time she experiences a given setting, she describes it with terms like "not painful", "slightly painful", "moderately painful", etc. At the end of the trial period ST gets a report of the frequencies with which she uses each description at each setting.

Arntzenius and McCarthy point out that since the first and last voltage settings are clearly discriminable, there must be some adjacent settings that differ in the frequencies with which ST uses the various descriptions. For instance, two adjacent settings a_n and a_{n+1} might be such that the first is described as "slightly painful" 96.3 % of the time and "moderately painful" 3.7% of the time, while the second is described as "slightly painful" 96 % of the time and "moderately painful" 4% of the time. Arntzenius and McCarthy argue that given everything ST knows, she would be unreasonable not to treat this difference in the frequencies of the different descriptions as evidence of a difference in pain levels; hence ST should take the expected level of pain to be higher at a_{n+1} than at a_n. As a result, since *ex hypothesi* ST cares about differences in levels of pain, she should assign different utilities to the two settings. But since the assumption that adjacent voltage settings would feel the same was essential to the argument that it was rational always to choose the monetary reward, we no longer have reason to think that ST's nontransitive preference ordering is rational. Given that ST prefers a_0 over a_k irrespective of the monetary reward at a_k, she should conclude that the increase in monetary reward will eventually stop compensating the increase in expected pain.[10]

Ingenious as this solution is, we think it faces several problems. First, it assumes that the described probability distribution somehow tracks what ST cares about. However, one might have doubts about this claim. In particular, at each stage ST might care only about the character of her current subjective experiences, the way she feels right now. The fact that in other circumstances, or even simply at other times, she would apply different descriptions to her experiences of adjacent voltage levels does not imply that there is now a felt difference between them. Nevertheless, let us grant that the probability distribution reflects what ST cares about. Still, the Arntzenius-McCarthy solution works only if we assume that ST's situation involves some hidden uncertainty—in this case, uncertainty about the location of the tipping point, viz., the first point in the series at which the increased monetary reward is less valuable (or has lower utility) than avoiding greater expected pain. (The uncertainty is contingent because it can be removed by showing ST the statistical record of her descriptions.) But a parallel puzzle can be generated even in the absence of such uncertainty. Consider a pattern of preferences often exhibited by smokers.[11] Many people think that the pleasures of smoking, no matter how many times repeated, do not compensate for the disease and premature death often caused by heavy smoking. On the other hand, for each next cigarette, these same people prefer smoking it (and perhaps quitting immediately afterward) over not smoking it (what difference could one cigarette make?). Like ST, they have a nontransitive preference ordering that seems to result from ordinary, not obviously irrational, preferences.

Now consider two hypotheses as to how smoking causes cancer. On the first hypothesis, smoke causes cumulative damage to one's lungs, so that although one cigarette will not cause cancer, and any single extra cigarette will cause only incremental damage, repeated smoking of single cigarettes will eventually almost certainly result in lung cancer. (Assume for the moment that no single cigarette ever puts a smoker over the threshold for lung cancer.[12]) Call this the 'incremental hypothesis". On the second hypothesis, each cigarette has a low probability of causing a relevant cell mutation. Although smoking one cigarette is unlikely to result in lung cancer, repeated smoking over many years raises the probability of developing lung cancer near to 1.[13] Call this the "stochastic hypothesis".

The Arntzenius-McCarthy solution might work if the incremental hypothesis is correct. One might think that a person who *always* prefers to smoke the next cigarette is not properly taking into account its expected harm; a rational agent who appreciated the expected harm would find a point at which the tradeoff between the pleasure of a cigarette and the prevention of cancer shifts in favour of the latter. However, their solution makes no sense on the stochastic hypothesis. Assuming that the pleasure of smoking each cigarette is independent of the pleasure of smoking any others (that is, pleasure neither increases nor decreases the more one smokes), and that the smoker's preference ordering does not change, each choice situation the smoker faces is identical. The probabilities are independent, so the trade-off at each choice situation is the same: a very small chance of developing cancer against the pleasure of smoking a single cigarette.[14]

On the incremental hypothesis, the Arntzenius-McCarthy solution allows some latitude in one's choice of cigarettes over health consistent with the hypothesis that one's preferences are stable; one could locate one's tipping point at any total number of cigarettes. But given that their solution does not work on the stochastic hypothesis, assuming stable preferences, orthodox rational choice theory appears to allow only two rational sets of choices: never choosing to smoke, and always choosing to smoke.[15]

One might question whether the choice situations would differ so radically depending on which hypothesis is correct. But leaving that worry aside, the idea that a rational agent in such a situation is confined to these two sets of preferences is badly in need of justification. Surely a person whose stable preferences dictate that she'll smoke only a few cigarettes and then quit so as not to endanger her life unduly is rational in light of her ends. If one's theory of rationality implies otherwise, some independent justification is needed. It's worth noting that orthodox RCT is committed to the same verdict in many analogous cases. For example, one can coherently avoid the inconvenience of strapping a child into a car seat for a short trip only if one would make the same choice in every similar situation, thereby putting the child at unacceptable risk; and one can slightly exceed the speed limit to avoid being late for a meeting only if one is always willing to speed in similar circumstances, despite knowing that in the long run such a policy would almost surely be disastrous; and so forth.[16] Arntzenius and McCarthy's paper does not discuss any “stochastic” cases, so maybe we should accept their solution to the self-torturer puzzle and seek a different treatment of the stochastic cases. But it’s difficult to evaluate such a “divide and conquer” strategy apart from a solution to the stochastic hypothesis case; and anyway, all else being equal, a solution to the ST puzzle that generalizes to other, apparently similar puzzles is preferable to one that does not.

It’s also worth noting that the puzzle doesn’t require adjacent settings of the dial to be indiscriminable. It seems equally rational to prefer large sums of money over *nearly* imperceptible, or even just slight, differences in pain, and yet prefer abject poverty over sustained agony; again, these seem to be the preferences of most ordinary agents.

Arntzenius and McCarthy anticipate this version of the puzzle and question whether such preferences would in fact be rational. They reason that, given the diminishing marginal utility of money, avoiding even very slight differences in pain will eventually be preferred to gaining $100,000. However, there is no guarantee that ST will arrive at her tipping point before she arrives at the point at which she would pay back all of the money to return to a_0. Whether she arrives at her tipping point first, depends on the rate at which slight increases in pain add up to unbearable pain, on the one hand, and the rate at which the marginal utility of money declines, on the other. Moreover, as long as the number of increases needed to reach unbearable pain is not extremely large, one can solve this problem by increasing the monetary reward at each choice node, slowing the decline to a crawl.[17]

Arntzenius and McCarthy's solution also faces a related problem that will afflict any orthodox solution to the puzzle. On their view, there is a certain point in the series at which ST's preference function changes from choosing the higher monetary value between two adjacent options, to being indifferent between two adjacent options, and then a subsequent point at which ST now prefers the lower monetary value between two adjacent options. More precisely, given ST's series of possible outcomes [o_1, …, o_n], if her choice situations always involve two adjacent outcomes, then there will be either a precise point at which she now strictly prefers not to accept the money, or there will be at least some triad of adjacent outcomes o_{m-1}, o_m, and o_{m+1} such that $o_m\mathbf{P}o_{m-1}$ but $o_m\ \mathbf{I}o_{m+1}$, and a certain later triad of adjacent outcomes such that $o_{m+k-1}\mathbf{I}o_{m+k}$ but $o_{m+k}\mathbf{P}o_{m+k+1}$.[18] And the trouble is that since each of these choices is independent, a rational agent must have the same preferences even when she is not threatened with the continuation of the series. But this seems counterintuitive.[19] Surely if an agent with ordinary preferences were to face only one of the choices ST faces, in isolation, she would be permitted, and likely required, to choose the option that increases her wealth significantly rather than avoid an insignificant increase in pain. On Arntzenius and McCarthy's view, indeed on any view according to which the choice situations are independent, if the choices happen to be between o_{m+k} and o_{m+k+1}, the agent cannot secure the outcome with the highest monetary value (*viz.*, o_{m+k+1}).[20]

2(b). Self-Torture as a Purely Intuitive Puzzle

Consider now the other way of interpreting the puzzle, according to which it follows simply from certain intuitive judgments. Thus understood, the puzzle arises for any theory of rational choice, not just for orthodox RCT. One could of course argue that there is nothing necessarily wrong with nontransitive preferences, or even with preferences that force you to choose in cycles. And in fact, a growing literature argues that nontransitive preferences might be rational.[21] However, even if we conclude that choosing in cycles is sometimes rational, it is hard to deny that ST is *not* rational if she does the following: continue to the last setting; then, should the option be available, pay back all of the money she has gained in order to return to the initial setting; and then be prepared to start again from scratch with the same choices as before. Moreover, intuitively it seems that ST is required stop at some point before the end of the series. Yet on the other hand, this intuitive response seems unstable; for at any proposed stopping place, wouldn't it be rational for ST to turn the dial up to the next setting, given that she clearly prefers it?

A proposal by Duncan Macintosh is instructive here.[22] According to Macintosh, an agent with nontransitive preferences can still make rational choices—namely exactly those choices dictated by her preferences. At least when it comes to pairwise choices, an agent who prefers A to B, and B to C, and C to A can always choose in accordance with her preferences. She might achieve an outcome that she would gladly trade back for her initial state, but why must this be irrational? Such a result is simply a consequence of the structure of her preferences.

We can grant that nontransitive preferences may not be irrational; our argument does not depend on the axiom of transitivity in its full generality. However, the description of ST's preferences as pairwise nontransitive does not fully capture her predicament. For she also has attitudes about the series of choices as a whole; in particular, she classifies some outcomes as clearly unacceptable. Among other things, were ST asked to make a single choice of one among all the settings, she would never choose the last (highest) one, yet that is where she would end up were she to follow her pairwise preferences.[23]

The "stochastic" version of the puzzle suggests that the problem revealed by the self-torturer puzzle is not a problem merely about nontransitivity. Although the agent's preferences in the smoking case can be characterized as forming a nontransitive set, the problem seems rooted, more deeply, in a tension between the agent's attitudes towards individual pairwise choices, on the one hand, and her attitudes towards cumulative outcomes that might result from those choices. In the stochastic case, the problem stems from the tension between between the willingness to accept the risk at each individual choice point, and the unwillingness to accept the greater risk that emerges from those individual choices in their totality.

In sum, even those who reject RCT must say what the self-torturer should choose to do given the compelling arguments both to raise the voltage at each stage and to stop at a stage early enough to avoid unbearable pain, and explain what makes this choice (or set of choices) rational. Here it is helpful to contrast Quinn's puzzle with the following variant:

The Amnesiac Self-Torturer

> The amnesiac self-torturer is faced with the same choices as ST. However, at each stage he does not remember the previous stages (he justifiedly believes that the pain he feels is caused by an accidental injury) and he has no reason to believe that further stages will ensue.[24]

On the face of it, the amnesiac self-torturer, who persistently raises the voltage and gains more money, seems subjectively rational even if he ends up wanting to pay back all of the money to return to a_0. The same cannot be said of ST; if ST ends up in unbearable pain, she seems irrational. Why the intuitive asymmetry, when both self-torturers face the same choices at each stage? Why should knowledge of how you arrived at a certain stage be relevant to the subjective rationality of your choices, given that the values or utilities of the objects of your choices in this case do not depend on your history of choices? By the same token, how can the knowledge that I will be offered further options if I choose option A provide reason, by itself, not to choose A, given that I know that I will be free not to exercise those further options when they become available?[25] Any theory of rational choice needs to answer these questions.

Since ST's predicament is puzzling even for opponents of orthodox RCT, an analogous obstinate solution should be equally unsatisfactory. Strictly speaking, those who reject RCT are free to dig in their heels and insist that ST's predicament shows only that rational preferences need not obey the axioms of RCT; they can say that ST ought simply to stop at a certain point without revising her preferences. But like the RCT obstinate solution, this kind of solution fails to explain what's wrong with the intuitively compelling arguments that favour raising the voltage and taking the money at each stage.

3. The Rational Structure of Vague Projects and Ends

Our discussion of orthodox solutions brings to light two desiderata for an acceptable solution to the puzzle of the self-torturer. First, an acceptable solution must explain the rational requirements on ST's decisions in light of the fact that she cares only about financial gain and freedom from pain; an acceptable solution should not postulate preferences other than those two. In particular, it must accommodate the fact that ST's preferences are, or at least appear to be, *discontinuous*: her preferences are such that an increase in pain can be compensated by certain financial benefits—*but not indefinitely*. Her pain can be compensated by money up to a certain vague threshold of pain.[26]

The second desideratum harks back to the example of the amnesiac self-torturer.

An adequate solution to the puzzle must respect a principle that we shall call "Non-Segmentation":

> *NON-SEGMENTATION*
>
> When faced with a certain series of choices, the rational self-torturer must choose to stop turning the dial before the last setting; whereas in any isolated choice, she must (or at least may) choose to turn the dial.[27]

Consider Barry, whose preferences at t_0 require him to stop no later than the setting a_n. As time goes by, Barry develops chronic back pain (but his preferences remain unchanged). Right now, at t_1, he is at the same level of pain he would be at were he to perform the self-torturer's sequence of choices and stop at setting a_m Barry is invited to participate in an experiment to determine whether a certain drug enhances the absorption of vitamin A. The experiment involves injecting a single dose of the drug VitA. Barry neither suffers from vitamin A deficiency nor is at risk of "overdosing" on vitamin A, so the effects of the drug on his vitamin A level are irrelevant to his choice situation. However, for reasons unknown, the VitA injection will slightly exacerbate the condition causing his back pain. Barry cannot discern the difference in pain caused by a single of dose of VitA, but we can assume that if the scientists were to inject more and more of the drug, the pain would become progressively worse and, eventually, be so much worse than his initial pain that he would gladly return all of the money to go back to his initial state before he had taken any of the drug. (We assume also that Barry knows all of this.) However, the experimenters, who are trustworthy, want to inject only a single dose; and they offer Barry $100,000 to take the drug. It seems that whatever you want to say about the rationality of stopping before or at setting a_m, if Barry were offered the $100,000 at t_1, he would be rationally permitted (arguably required) to accept it and take the single injection of VitA. Non-Segmentation demands that a solution allow (perhaps require) the self-torturer to accept the monetary offer in any one-shot version of the puzzle.

An adequate solution to the puzzle should comply with Non-Segmentation or else explain why the principle seems plausible. In fact, in light of Non-Segmentation we can see why orthodox RCT is unable to resolve ST's predicament. From the viewpoint of orthodox RCT, there can be no difference between the serial and one-shot choice situations. Hence a solution that preserves orthodox RCT is bound to yield counterintuitive results regarding at least one of the two.

It is worth noting that ST's predicament is not an isolated anomaly to be accommodated by minor revisions or exceptions to a theory of rational choice. On the contrary, as Quinn himself notes, the structure of the self-torturer puzzle is present in many ordinary actions. In particular, any instance of what we'll call a "vague project" or a "vague end" will generate an analogous puzzle, and most of our projects and ends are vague. In what follows we will develop a certain conception of the demands that practical rationality makes on vague projects or ends.[28] We will then show how this conception helps to resolve the puzzle of the self-torturer. Broadly speaking, our view is that orthodox RCT imposes a substantive restriction on our pursuits: they cannot include vague projects or ends. Puzzles like that of ST cast doubt on the legitimacy of such a restriction, since many projects that seem rationally innocent are vague.[29] In what follows, we will use the term "vague project" broadly to include not only pursuits with highly complex structure involving planning etc., such as writing a book, but also those involving simpler ends that one adopts or pursues, such as, for instance, baking a cake, whose realization or completion is also vague in at least some aspects (what counts as an acceptable cake, what counts as an acceptable timeframe for baking the cake, and so forth).

Suppose you are writing a book. The success of your project is vague along many dimensions. What counts as a sufficiently good book is vague, what counts as an acceptable length of time to complete it is vague, and so on. In particular, the following assertions seem true of this project:[30]

(i) Its completion requires the successful execution of many momentary actions.

(ii) For each momentary action in which you execute the project, failure to execute that action would not have prevented you from writing the book.

(iii) On many occasions when you execute the project, there is something else that you would prefer to be doing, given how unlikely it is that executing the project at this time would make a difference to the success of your writing the book.

(iv) Had you failed to execute the project every time you would have preferred to be doing something else, you would not have written the book.

(v) You prefer executing the project at every momentary choice situation in which you could work on the project, over not writing the book at all.[31]

Now suppose that you have completed your book. Looking back, you recall a certain Sunday afternoon on which you sat on the porch reading *The New York Times*. Around 4 o'clock you toyed with the idea of reading one more article in the paper, but decided instead that you had better get back to work. You didn't find the prospect of going back to writing intrinsically rewarding; in fact your sole reason for setting the newspaper aside was the need to complete your book. Looking back, you now realize that, had you read another article instead of getting back to work, this would have not prevented you from completing your book. But you are unlikely to regret your decision, and your lack of regret would not be irrational. The fact that you could have read one more article and still finished your book doesn't show that the decision to work was irrational, since there is no precise set of momentary actions that are strictly necessary and sufficient for the completion of the project, i.e., no set of momentary actions such that if even one had been subtracted from the set, the project would not have been completed (or, if there is such a set, it is not epistemically accessible to you and thus not relevant to determining your [subjective] rationality).

Vague projects or ends are problematic for orthodox RCT not only because what counts as success is vague in such cases, but also because the success of a vague project cannot be measured by the success of specific momentary actions.[32] For example, suppose one's project is to provide free hair growth products to all and only those who are bald. What counts as success in this project would be vague, but it could conceivably depend upon a single momentary action (e.g., one could press a button that would legally transfer the ownership of a bottle of hair growth product to bald persons across the universe). In contrast, success in a vague project depends upon a series of momentary actions and is measured in terms of patterns of activity extending through time; in the case of a vague project, there is no measure of the rationality or success of any particular momentary action with respect to the project. Moreover, the comparative evaluative judgments among various outcomes of your book-writing project have no obvious, perhaps even no systematic, implications for the evaluation of your momentary actions. Consider for instance the following comparative judgments:

(1) It is better that you write the book than that you don't write the book,

(2) It is better that you write the book and visit Auntie Mary regularly in the nursing home than that you write the book but don't visit Auntie Mary regularly.

(3) It is better that you write the book and train enough to do well in the NYC marathon than that you write the book and barely finish the NYC marathon.

None of these judgments dictates what you should do at any given moment.[33] You could work on your book, go for a run, call Auntie Mary to schedule a visit; all of these actions are compatible with the above evaluations. Ideally, you would end up writing a book, seeing Auntie Mary regularly, and recording a personal best in the marathon. And insofar as these things are within your rational control and do not conflict with your other projects and preferences, you are rational if you accomplish all of them. More to the point, there seems to be nothing *else* that rationality requires of you; as long as you succeed in all of these projects, and as long as your success in these projects is non-accidental, you are perfectly rational.[34] However, since executing these projects successfully is compatible with failing to pursue any particular momentary action considered in isolation,[35] there can be no rational requirement stemming from these projects to undertake any particular momentary action.

If the preceding line of thought is correct, we should let go of the tempting idea that the irrationality of failing to take adequate means to execute these projects supervenes on the irrationality of momentary actions.[36] In other words, we should accept the possibility of what may be called "top-down irrationality", i.e., the possibility that certain patterns of activity can be irrational without any of the momentary actions that compose those patterns being irrational.[37] On the other hand, given that *never* taking the means to execute these projects would be irrational, an adequate account of rational agency must make some recommendation with respect to their execution. The pursuit of such projects must contribute in some way to the assessment of the rationality of momentary actions.

We propose that a vague project issues in a requirement and a set of permissions. The requirement is just an instance of the instrumental requirement: insofar as one is rational one must adopt (what one believes to be) the means (including constitutive means) necessary to execute one's project. The permissions are permissions to execute the project in some momentary actions rather than simply maximizing utility in light of one's preferences for momentary actions considered in isolation. This proposal should not be construed on the model of theories that allow for agent-centered permissions in ethics.[38] Whereas on the latter views one often has a permission to deviate from bringing about the best outcome, we are proposing that some goods cannot be pursued by always pursuing what would be *best at a given moment*. It's not that an agent has the permission to pursue something other than what's best overall in a particular situation, but that the evaluation of the alternatives becomes more complex when the agent is engaged in vague projects. (We will return to this point.) The trouble is that *a standard utility maximizing theory forces vague projects to figure in one's preferences directly* *in momentary choices*.[39] In contrast, vague projects are temporally extended actions and their proper execution can be assessed only in light of the entire period during which they were, or ought to have been, executed.

We suggest that two perspectives are relevant to the rationality of your actions while writing your book. The first perspective we will call *extended.* The extended perspective issues the requirement mentioned above; in this case, the requirement that, roughly, your actions overall can be expected to bring about the writing of your book, or that your actions constitute your writing the book in an acceptable way, i.e., in such a way that you are (or should be) satisfied that this is the kind of book that you're aiming to write. Obviously, many (perhaps indefinitely many) possible sequences of momentary actions would satisfy the requirement imposed by the extended perspective. If you did go on to finish writing the book in an acceptable manner,[40] the possibility that you could have written a slightly better book had you spent less time reading the newspaper, or the possibility that you would have written the same book had you read an additional article, does not show that you violated the extended requirement. If you wrote an acceptable book without unduly compromising your other projects for the sake of the book, then you satisfied the requirement of the extended perspective. The second perspective we will call *punctate*. The punctate perspective determines which momentary actions are open for you at a given moment, as well as how each action on its own can contribute to the satisfaction of your ends.

We will say that your project of writing a book is *implicated* at certain momentary choice points; at certain points in time, one of the momentary actions open to you is to take (constitutive or instrumental) means to complete your project of writing a book. That is to say, at those moments, one rational action open to you is to exercise the permission to execute your long-term vague project, i.e., to take (some) means to writing the book.[41] Here we require a notion of means to an end that is broader than the traditional view according to which a means is either necessary or sufficient (or both) for the attainment of the end. On our view, the momentary actions of executing a vague project are neither necessary nor sufficient; rather, they are merely *constituents* of a means that is necessary and/or sufficient to bring about the completion of the book.[42] We can say that token actions of the relevant type are *general*, as opposed to necessary or sufficient**,** means to their ends.

It is worth distinguishing between general means and what can loosely be called sets of "indifferent" means. Suppose my end is to buy milk on my way home, and that between my office and home are seven grocery stores that sell milk. *Prior to my choosing any action(s)*, none of the stores is such that buying milk there is a necessary means to buying milk on my way home. Although buying milk on my way home will require that I visit at least one of the stores, any store is as good as any other. At the same time, if I don't stop at any of the first six stores, then when I arrive at the seventh store, the only rational action is to stop. This is because the project of buying milk on the way home is not, in the relevant sense, vague: it can be accomplished by a more or less momentary action. Although no *particular* action of stopping at the store is a necessary means, performing at least one these actions is a necessary and sufficient condition. We can say that in this case the set of all possible actions of stopping at one of these stores forms a set of indifferent means. In the case at hand, if I haven't stopped at any of the first six stores, stopping at the last one will be necessary and sufficient for achieving my end. On the other hand, when my project is vague, like writing my book, no particular moment is such that, given my previous choices, typing at that moment is necessary for the acceptable completion of my project. If I do not write enough, the book will not be completed, or will not be completed acceptably, but no particular moment is such that choosing to write at that moment is necessary, or choosing not to type at this moment will render the book unacceptable. Momentary actions of writing are *general means* to the execution of the project.

In analogous fashion we can distinguish between a project's being *implicated* and its being *generally implicated*, at the time of a certain momentary action:

(1) A project of yours is ***implicated*** at a time **t** if and only if (you know that) one of the momentary actions open to you at **t** is to take means to executing that project, where a means is a necessary and sufficient condition, or at least INUS (insufficient but a necessary part of an unnecessary but sufficient) condition,[43] of completing the project.

(2) A project of yours is ***generally implicated*** at a time **t** if and only if one of the momentary actions open to you at **t** is to perform a token of an action type T such that:

(i) In the time interval (that includes **t**) in which you are executing your project or pursuing your end, there will be several choice situations in which one of your options is to perform a token of action type T.

(ii) Were you never to perform tokens of T, your project would not be successfully completed or brought about.

(iii) No particular token of T is (or could be known to be) necessary (or sufficient, or an INUS condition) for the completion of the project.[44]

(iv) You know that (i) - (iii) obtain.[45]

The fact that an end or project is generally implicated is relevant to the rational options open to an agent in virtue of the structure of that kind of project as delineated above. You need to be engaged in typing on your computer multiple times while you are writing your book, but not at any *particular* times. More specifically, tokens of the relevant action type T, considered singly, are neither necessary nor sufficient to bring about one's end. It is necessary and (in conjunction with other actions or means) sufficient that we perform enough tokens of the relevant action types to achieve the end or complete the project in question. However, suppose I complete my project of writing a book partly by choosing often enough to type on my computer. Given the vagueness of the project, I would never be in a position to say that any of these choices was necessary to complete the project. Had I chosen to type for one less second, I would also have completed my project of writing a book.

The extended perspective issues only a requirement to perform sufficiently many momentary actions of the relevant type. However, an agent can satisfy the requirement of the extended perspective only if the punctate perspective permits performance of actions that are general means to a project that is generally implicated at a certain moment, even when they are neither necessary nor sufficient for the completion of the project. Of course, that permission will not extend to all actions in which the project is generally implicated; if you must act now to save your beloved from a burning building, there is probably no rational permission to type a few more pages of your book instead. Also, the permission does not apply where your project is *not* implicated because you can no longer pursue it. If you are lying on your deathbed having written only a few pages of your book with no hope of finishing it, the project of writing it cannot generate a permission to spend your last hours typing away. [46]

Note that we have not introduced any new basic requirement; each perspective is generated by the instrumental requirements that attach to any actions. However, given the extended nature of vague projects, certain violations of (and instances of compliance with) the instrumental requirement cannot be located at particular moments but only in extended periods. Moreover, where vague projects are concerned, no particular actions considered in isolation would constitute a violation of the instrumental principle.[47] A *requirement* of particular momentary actions would be stronger than what follows directly from the instrumental principle. On the other hand, lacking permissions to execute the long term project would guarantee that the principle is violated.

4. Back to the Self-Torturer

With these materials in hand, let us see how our discussion of the rationality of vague projects bears on the puzzle of the self-torturer. What vague projects, if any, are implicated when ST makes her successive choices in the original version of the puzzle? First ST has the vague project[48] of leading a life relatively free of pain, as one might characterize it: if the self-torturer continues to accept the monetary offers, she will fail in this project. Her project of living a pain-free life is generally implicated insofar as there is an action such that repeated performance of that type of action (*viz.,* refraining from taking the money) under the same circumstances would normally be necessary for its realization. This pain-free life project requires that the self-torturer not always perform tokens of this type of action; it can be realized only if the self-torturer does not go too far. Accordingly, the pain-free life project issues permission to stop turning the dial, independently of what maximizes utility in light of ST's momentary preferences.

Considered in isolation, the project of leading a pain-free life permits ST to refuse every monetary offer from the experimenters. However, exercising the latter permission would violate the extended requirement generated by ST's other vague project of making enough money; hence the requirement of the money-making project confines the permission generated by the pain-free life project to the later stages of the series of choices. ST is rational just in case she exercises the permissions issuing from the extended perspective in such a way that she completes both projects in an acceptable manner (which is also *required* by the extended perspective).

If we are right to attribute these projects to ST, our solution satisfies the first desideratum expressed above: it explains the rational requirements faced by the self-torturer given that she cares solely about financial gain and avoidance of pain and without postulating new preferences. However, one might object that we did add something that was absent from the original puzzle. After all, the original formulation says nothing about ST's projects; in fact it seems enough to generate the puzzle that ST prefers always the higher setting of two adjacent settings and yet prefers the first setting over the last setting.   We have now introduced, if not a new preference, a new relevant attitude; in particular the project of a relatively pain-free life.

Nonetheless, simply *postulating* these preferences cannot generate the puzzle. We can postulate acyclical preferences to our hearts' content. What makes *ST*'s preferences puzzling is that they seem perfectly rational; in fact, they seem like rather ordinary and unproblematic preferences or attitudes. They seem unproblematic because most of us recognize that we would (reasonably) not give up a (relatively) pain free life for any amount. In our view, this expresses our commitment to a project of leading a (relatively) pain free life that we would not be willing to abandon (even if this project were to prevent us from earning princely sums). In other words, we contend that these attitudes are unproblematic in part because they manifest ST's commitment to a project of a (relatively) pain free life. Of course, one could challenge our characterization of what makes ST's attitudes rational and ordinary. A challenge along these lines would have to generate the puzzle in such a way that ST was still intuitively rational (neither "confused" nor a "freak" in the language of our earlier setup), but incapable of being characterized as having the project of a relatively pain-free life or a similar (and similarly vague) project. Our hypothesis here is that this cannot be done.

This solution also satisfies the second desideratum, *viz.*, to comply with Non-Segmentation. To see how, suppose that Gary, who has the same ordinary preferences as ST, is presented with a single choice from ST's series. Is he rationally required to accept the money and turn the dial to the next setting? We think so. Recall the case of Barry, who suffers from mild chronic back pain and is offered $100,000 to take the vitamin A drug just once. What does our view say about Barry? Since his choice is a one-shot deal, the project of leading a pain-free life is not generally implicated (either condition [i] does not obtain, or if it does, then condition [ii] doesn't). Hence no permission is generated by the latter project, and either only Gary's momentary preferences are relevant, or else the only project implicated in this choice is the project of making enough money;[49] either way, Gary is rationally required to accept the$100,000. One could of course elaborate the example in such a way that the project of leading a pain-free life *is* implicated; for example, maybe scientists are running similar experiments all around Barry's university. In that case Barry is permitted to refuse the money, but the existence of such a permission is not counterintuitive.

5. Plan Solutions

What we'll call "plan solutions" to the puzzle of the self-torturer respect Non-Segmentation: they require that the agent make a plan and then stick to it. Arbitrary revisions of the plan are disallowed even if the revision would not take the self-torturer to an unacceptable combination of pain and money. Plan solutions seem to entail that an agent who stops at an acceptable but unplanned point is irrational, even if the stopping point is preferred over the planned one. For instance, suppose ST plans to stop turning the dial at setting a_{25}, but then when she arrives at a_{25} she decides to go one step further and stop at a_{26}. Also, suppose a_{26} is a setting that ST prefers over the initial setting a_0. Arguably, according to planning solutions, ST has acted irrationally despite the fact that she ends up with more money and an acceptable level of pain—indeed a level of pain that is pairwise indistinguishable from the planned one. Clearly, a charge of irrationality here stands in need of defense.

Michael Bratman's "planned" solution works in essentially this way.[50] In a nutshell, Bratman imposes a "no regret" condition on plan revision: it is rational for an agent not to revise a plan even in light of her preference to do so, just in case she reasonably judges that she will regret her choice if and only if she chooses to revise her plan. So suppose that, before she straps the device to her wrist, ST plans to stop at setting a_{25}. As she gets there, she realizes she would prefer to go farther, and wonders whether she should revise her plan. According to the no regret condition, ST needs to figure out whether she now believes that were she to revise her plan according to her current preference, she would later regret it. Bratman thinks that ST can reasonably expect to regret revising her plan:

> She can ask: "If I abandon my prior intention to stop at [a_{25}], what will then transpire?" And it seems she may reasonably answer: "I would then follow the slippery slope all the way down to [$a_{1,000}$][the last setting]". [51]

According to Bratman, ST would now realize that were she to revise her plan, she would later regret having done so. Thus, revising her plan would violate the no regret condition.

Granted, if ST has reason to believe that she will either stick to her plan or continue to the end of the slippery slope, i.e., that these are her only options, then she shouldn't revise her plan.[52] But why should she believe that? By hypothesis some of the settings above a_{25} are acceptable but the final setting is not; so a rational agent could certainly stop at a point between a_{25} and the final setting if she didn't stop at a_{25}. Of course someone's *psychology* might be such that he can stop at a setting short of the final setting only if he stops at a planned setting (e.g., a_{25}), but why think this is true of rational agents as such? Bratman writes:

> His prior decision to stop at [a_{25}] was his best shot at playing the game without going all the way; if he does not stick to that decision, there is little reason he would stick with any other decision short of the bottom of the slippery slope .

On the contrary, if ST is rational, she has good reason to believe that she *will* stop short of the end, since the end point is plainly unacceptable. So why must she make a prior decision and stick to it? Surely she could abandon her initial plan and later exercise a permission to stop without forming any plan to do so.

Naturally, if one rejects the possibility of top-down irrationality, it may seem that if ST doesn't stop at the planned setting, she will have no reason to stop later, since whatever reasons she would have at a later point would also be reasons to stop at the planned setting; and if those reasons were inadequate at the planned setting, they would be inadequate at any other point. We find this line of reasoning problematic in many respects, but for now we need only repeat that rejecting top-down irrationality imposes an extra requirement on rational agency—one that would need to be justified, and that seems implausible on independent grounds. Its implausibility is evident in the ST case itself. Suppose again that ST plans to stop at a_{25}, but at a_{25} decides it's worth going a little further, and then stops at a_{26}. Assuming that a_{26} is still an acceptable stopping point, how could it have been irrational for her to continue to a setting that is acceptable and that she prefers over the planned setting? Perhaps someone will object that ST would stop at a_{26} by sheer luck; she certainly couldn't have counted on that. But ST makes her decision under certainty; she fully controls the outcome. So it is hard to see how her choice of stopping point could be merely lucky.[53]

Macintosh makes a similar proposal concerning cases in which the agent faces vague boundaries.[54] On his view, ST should pick at random (by, say, flipping a coin or using a random number generator) a number within "a vaguely bounded safe region" . Let us assume that ST can identify such a region prior to engaging in the experiment. The proposal then is to have ST select at random a place to stop, and then ensure that she stops at that point. When we express Macintosh's view this way, we can see that it's isomorphic to Bratman's, and faces the same problem. Suppose the agent's technique dictates that she stops at setting a_n, but she stops instead at setting a_{n+1}. If a_{n+1} is an acceptable outcome, why is she irrational?

Perhaps someone will ask how ST knows when to stop without having a plan or method of selection. We contend that exercising one of the permissions of the extended perspective without violating any of its requirements is all there *is* to knowing when to stop. If such knowledge seems mysterious, consider that it's an instance of a common phenomenon. Suppose you take a break from writing an important memo and start surfing the web. Surely surfing for one additional second will not prevent you from completing the memo, but if you surf for long enough you won't have time to finish it. The point is that while you often find yourself in such situations, you don't employ any symmetry breaking technique to determine when to stop surfing. You simply do stop with enough time left to finish the memo.[55] For that matter, if you did employ such a technique, that would be irrelevant to determining the rationality of your action. Instrumental rationality dictates only that you stop at an acceptable point.

6. A Potential Objection: Weighing the Projects

One might argue that the present solution sidesteps the most difficult task facing the self-torturer, namely how to determine the relative importance of her various projects. At first blush the best (most rational) strategy would be to find a point at which the tradeoff between pain and financial gain is optimal. But if the present account is correct, there is *no such optimal point* where the projects at issue are vague.

Of course, what counts as an acceptable financial gain and an acceptable amount of pain depends partly on how the pain and gain interact: but for the pain generated, ST would be rationally required to accept the money *ad indefinitum*.[56] Our view may seem to leave no room for such an interaction; in our view, the extended perspective generates requirements and permissions from each long-term project, but the requirements and permissions of each project are independent of the requirements and permissions of any other. However, distinct projects do constrain each other in significant ways. Suppose that Harry has two projects—making as much money as possible, and leading a life completely free of pain. Once Harry learns that he'll face the series of choices in the self-torture puzzle, he ought to realize that his projects are mutually incompatible, and revise at least one of them. By reflecting on the nature of vague projects, we learn that in such cases we cannot simply plug weights in to various ends to generate a preference ordering; rationality is not always purely calculative. As a matter of fact, since vague projects or ends do not allow for preference orderings that satisfy the axioms of expected utility theory, there may be no procedure for determining what an agent with such projects ought to do in every possible circumstance. Instrumental rationality dictates only that an agent ought not to pursue incompatible ends, ends that she knows (or expects) are not jointly realizable. Hence where vague projects conflict, all instrumental reasoning tells us is that the agent must revise at least one of them. In particular, there is no guarantee that *how* the project(s) should be revised will be determined by the proper employment of instrumental reasoning. At any rate, it is the revised projects, rather than the incompatible ones, that generate the puzzle of the self-torturer and, therewith, the requirements and permissions relevant to assessing ST's rationality.

7. Conclusion

While orthodox rational choice theory provides a compelling and powerful conception of practical rationality, it relies upon at least two importantly different types of idealizations. On the one hand are idealizations intrinsic to the subject matter of RCT: as a theory of rational choice, it presupposes that its agent is fully rational, thus ignoring the many ways in which human beings fall short. On the other hand are what might be called "simplifying" idealizations: in particular, the axioms of orthodox RCT do not apply to vague projects or ends. The reasons to presuppose full rationality are clear enough, but why should a theory of *instrumental* rationality require that an agent's projects be non-vague? Shouldn't such a theory apply to projects, such as the writing of a book, that do not specify precisely how they ought to be completed, or specify a precise comparative valuation of the various ways of completing them and other competing ends that the agent might have?

The puzzle of the self-torturer shows not that ST's ends or preferences must be flawed, as RCT would have it, but rather that since ST's ends or preferences are incompatible with some of RCT's simplifying assumptions, RCT cannot tell us what an agent in ST's predicament is rationally required or permitted to do.[57] In the absence of an alternative theory, it is tempting to conclude that the problem must lie with ST; but we think such a conclusion is unwarranted. We hope to have taken a step toward understanding how principles of instrumental rationality apply to vague projects, and, in so doing, quieted the suspicion that an ordinary agent with such projects must be irrational.

Bibliography

Andreou, Chrisoula. "Temptation and Deliberation." *Philosophical Studies* 131, no. 3 (2006): 583-606.

Arntzenius, Frank, and David McCarthy. "Self Torture and Group Beneficence." *Erkenntnis* 47, no. 1 (1997): 129-44.

Benbaji, Itzhak, and Dan Ariely. "Transitivity as a Metapreference." Jerusalem, 2011.

Bratman, Michael. "Toxin, Temptation, and the Stability of Intention." In *Faces of Intention*. Cambridge: Cambridge University Press, 1999.

Carlson, Erik. "Cyclical Preferences and Rational Choice." *Theoria* 62 (1996): 144-60.

Fishburn, P.C. "Non-Transitive Measurable Utility for Decision under Uncertainty." *Journal of Mathematical Economics* 18, no. 2 (1989): 187-207.

MacIntosh, Duncan. "Intransitive Preferences, Vagueness, and the Structure of Procrastination." In *The Thief of Time*, edited by Chrisoula Andreou and Mark White, 68-87. New York: Oxford University Press, 2010.

Mackie, J.L. *The Cement of the Universe*. Oxford: Clarendon Press, 1980.

Quinn, Warren. "The Puzzle of the Self-Torturer." In *Morality and Action*, edited by Warren Quinn, 198-209. New York: Cambridge University Press, 1993.

Raffman, Diana. *Unruly Words: A Study of Vague Language*. New York: Oxford University Press, forthcoming.

Tenenbaum, Sergio. "The Vice of Procrastination." In *The Thief of Time*, edited by Chrisoula Andreou and Mark White, 130-50. New York: Oxford University Press, 2010.

Voorhoeve, Alex, and Ken Binmore. "Transitivity, the Sorites Paradox, and Similarity-Based Decision-Making." *Erkenntnis* 64, no. 1 (2006): 101-14.

Weirich, Paul *Realistic Decision Theory*: Oxford University Press, USA, 2004.

NOTES

[*] We would like to thank Chrisoula Andreou, Hagit Benbaji, Itzhak Benbaji, Dalia Drai, David Enoch, Yuval Eylon, Edmund Henden, Thomas Hurka, Olav Gjelsvik, Jennifer Nagel, Julia Nefsky, Christian Piller, Jim Pryor, Levi Specter, Michael Thompson, Ruth Weintraub, Jonathan Weisberg, and audiences at CSMN, at the University of Oslo, the University of Toronto, University of St. Andrews, Simon Fraser University, Universidade Estadual do Rio de Janeiro, Universidade Federal do Rio de Janeiro, University of Sheffield, Groningen University, University of Bern, two anonymous referees, Donald Hubin, and three anonymous editors for *Ethics* for many helpful comments on earlier versions of the paper. This research was supported by the Social Sciences and Humanities Research Council of Canada and the Institute for Advanced Studies, Jerusalem, Israel.

[1] Warren Quinn, "The Puzzle of the Self-Torturer," in *Morality and Action*, ed. Warren Quinn (New York: Cambridge University Press, 1993), 198-209.

[2] We have slightly modified Quinn's presentation of the puzzle.

[3] As will emerge, the puzzle does not depend on the axiom of transitivity in its full generality. Readers familiar with the sorites paradox may wonder whether the self-torturer puzzle is just an especially picturesque instance of it: perhaps ST is proceeding along a sorites series of pains from a clearly bearable one to a clearly unbearable one, attempting to decide where the bearable ones end and the unbearable begin. However, this way of thinking about ST overlooks a crucial element of her situation: at each step of the way she is also trying to decide whether

a certain incremental difference in pain can be compensated by $100,000 at that point in the "spectrum" of her pain. The latter task is what appears to put pressure on her rationality and is, at bottom, the source of the puzzle.

[4] We understand by 'orthodox rational choice theory' (or, simply, 'rational choice theory', or 'RCT') theories that are committed to the classical axioms of decision and game theory or subjective expected utility theory; that is, theories committed to the Von-Neumann-Morgenstern axioms or the Savage axioms, and possibly some variations thereof. Rational choice theory has many applications, both descriptive and normative; and not everyone committed to some application of it will face all of the difficulties presented here. ST presents a problem for those who take RCT to be a normative theory and believe that an agent whose preferences cannot be represented in the theory (or at least cannot be represented as a fragment of a utility function that conforms to its axioms) is to that extent irrational.

[5] Quinn intends the latter.

[6] Frank Arntzenius and David McCarthy, "Self Torture and Group Beneficence," *Erkenntnis* 47, no. 1 (1997), 129-44 endorse this view to some extent. We discuss it below.

[7] Alex Voorhoeve and Ken Binmore, "Transitivity, the Sorites Paradox, and Similarity-Based Decision-Making," ibid.64(2006), 101-14.

[8] Frank Arntzenius and David McCarthy, "Self Torture and Group Beneficence."

[9] We simplify their version slightly, but our objections will not depend on that.

[10] Quinn thinks that a version of the self-torturer puzzle can be constructed in which adjacent voltage settings are absolutely indiscriminable, i.e., not discriminable even (e.g.) by triangulation with other settings. Quinn says:

> Surely it ought to be an open empirical question whether such triangulations are possible; if there are increments of voltage just small enough to be directly undetectable, it seems there might be even smaller increments that cannot be detected by triangulation. And I want such a case. (Quinn, "The Puzzle of the Self-Torturer," 201)

Tempting though the idea may be, such absolute indiscriminability is not in fact logically possible. (See Erik Carlson, "Cyclical Preferences and Rational Choice," *Theoria* 62(1996), 144-60, for a similar claim.) The puzzle stipulates the following:

(1) Pairwise Indiscriminability

For every two adjacent settings a_j and $a_{j+1,}$ the Self-Torturer cannot discriminate between a_k and a_{k+1} in pairwise comparisons.

Quinn is proposing to add the following:

(2) No Triangulation

For any settings a_j and a_k, a_j and a_k are pairwise indiscriminable if and only if a_k and a_{j+1} are also pairwise indiscriminable.

However, (1) and (2) entail that *all* settings of the dial are indiscriminable, contradicting the stipulation that the first and last settings are discriminably different. Consider for example the first three dial settings a_1, a_2, and a_3. *Ex hypothesi* a_1 and a_2 are indiscriminable, and likewise a_2 and a_3. Given the No Triangulation rule, a_1 and a_3 are indiscriminable (otherwise a_1 would be discriminable from a_3 but a_2 would not, contradicting the stipulation that a_1 and a_2 are indiscriminable). Given Pairwise Indiscriminability, a_3 and a_4 are also indiscriminable. But then given No Triangulation, a_1 and a_4 too are indiscriminable; and so forth. Hence at least some adjacent settings must be discriminable by triangulation.

[11]For simplicity, we will ignore the complication that smoking is addictive.

[12] If lung cancer is an on/off condition with no vague threshold, we can change the case to a fatal condition that is gradually or incrementally debilitating. Since the main point of the example is the contrast with the stochastic hypothesis, this will not affect our argument.

[13] We make no claims about the scientific plausibility of either hypothesis. However, recent debates about body scanners vs "pat downs" at US airports have a similar structure. Experts often say that frequent fliers and flight crews should perhaps prefer the "pat down" even though the choices they face on each occasion are identical: there is no chance of cumulative harm, only a small risk of harm in each

exposure. See, for instance, http://www.npr.org/2010/11/19/131447056/are-airport-scanners-safe.

[14] Of course, at some point one might have already developed cancer, and past choices might raise one's credence that that is so. But assuming that smoking does not introduce further risks to those who already have lung cancer, the expected utility of smoking only goes up for later choices (provided the agent has smokers' usual preferences).

[15] RCT doesn't say that the utility of smoking *n* cigarettes is a function of the utility of smoking one cigarette. That's why we need to assume that the utility of smoking the next cigarette does not vary given that one has smoked a cigarette in the past, and that the preference ordering is stable through time. We assume also that the agent is not indifferent between the pleasure of smoking and the relevant probability of developing cancer (otherwise any combinations of choices would be consistent with RCT). The same caveats apply to the other cases we discuss below.

[16] See caveats at note 15 above.

[17] Voorhoeve and Binmore, "Transitivity, the Sorites Paradox, and Similarity-Based Decision-Making," suggests that we fall victim to well-known heuristic biases when we judge that small differences in pain are always compensated by monetary rewards. This strikes us as implausible. First, many of the biases they discuss are not recalcitrant; subjects recognize their mistake once they understand how their choices result in a nontransitive preference ordering. In contrast, in ST's

predicament we do not revise (or at least don't know how to revise) our general preferences once we are aware of the difficulty. Moreover, it is not clear how our intuitive reasoning about ST could count as a case of (mis)using certain heuristics. After all, the preferences in question are not supposed to be a method used to discover the truth about *something else*; they are supposed to *determine* our basic attitudes regarding acceptable tradeoffs between money and pain. Of course one could deny this last point, and say that our preferences reveal only the heuristics we are using in trying to conform to some deeper attitudes regarding acceptable tradeoffs between money and pain. But then the claim that these heuristics are (mis)used would need to be supplemented by some account of what determines the correct tradeoffs between money and pain.

[18] It is compatible with Arntzenius and McCarthy's solution that ST go directly from $o_m\mathbf{P}o_{m-1}$ to $o_m\,\mathbf{P}o_{m+1}$ for some setting a_m, with no indifference point. But this would make the present problem worse, if anything.

[19] This counterintuitive result violates a requirement we'll later call *Non-Segmentation*.

[20] If you think it makes a difference that we are looking at three rather than two stages in the series, shrink the increments by half. Also, we don't mean to deny that a rational agent could have preferences making it rational for her to reject the money in one shot versions of the puzzle; this much is true even if triangulation can be ruled out. ST doesn't have such preferences, though.

[21] See, for instance, Duncan MacIntosh, "Intransitive Preferences, Vagueness, and the Structure of Procrastination," in *The Thief of Time*, ed. Chrisoula Andreou and Mark White (New York: Oxford University Press, 2010), 68-87., and Itzhak Benbaji and Dan Ariely, "Transitivity as a Metapreference," (Jerusalem2011). Of course there are also various formal proposals for measuring utility that do not presuppose transitivity of preference; see for instance P.C. Fishburn, "Non-Transitive Measurable Utility for Decision under Uncertainty," *Journal of Mathematical Economics* 18, no. 2 (1989), 187-207

[22] MacIntosh, "Intransitive Preferences, Vagueness, and the Structure of Procrastination."

[23] In fairness to MacIntosh, he thinks that the original setup of the ST puzzle is incoherent. See note 54 below.

[24] See also the discussion of Non-Segmentation below. “Amnesiac” is a bit of a misnomer since the subject in the case must also be ignorant of his *future* choices, not just the past ones.

[25] To be sure, the mere existence of certain options may be undesirable in some cases; for example, I might not want to have the option of receiving $1,000,000 if I betray my best friend, or if I enslave someone. However, while we do not rule out the possibility, argument would be needed to show that ST’s situation is such a case.

We raise the above questions only to emphasize that an “obstinate” solution will be unsatisfactory to friends and foes of RCT alike.

[26] For simplicity, we’ll assume that no amount of money can compensate ST above this vague threshold; but all we need to generate the puzzle is a discontinuous function.

[27] Donald Hubin has pointed out to us that Non-segmentation can be thought of as a “one-person game” equivalent to the denial of subgame perfection. It is an interesting question whether similar issues would lead us to deny that rational strategies in multiplayer games must always be in subgame perfect equilibrium. We think the answer is “yes”, but we won’t pursue the matter here.

[28] Some elements of this conception appear in Sergio Tenenbaum, "The Vice of Procrastination," in *The Thief of Time*, ed. Chrisoula Andreou and Mark White (New York: Oxford University Press, 2010), 130-50.

[29] Certainly classical decision theory may be correct within a certain subdomain—specifically, it may prescribe the correct actions where no vague projects or ends are involved.

[30] This example and the characterization of vague projects below appear first in Tenenbaum, "The Vice of Procrastination."

[31] (v) is stronger than needed and perhaps implausible in this strong form. Life might be very unpleasant if every moment at which I could be writing my book, I

would be writing it. All we need is that this (v) be true for a large set of momentary choice situations. But the stronger version makes for simpler presentation.

[32] Except in certain degenerate cases in which a single momentary action can guarantee failure of a project; for instance, I jump off a bridge, guaranteeing that I'll either die or suffer enough brain damage that I won't be able to finish my book.

[33] This claim obviously simplifies matters somewhat; see note 32.

[34] Non-accidental success is a sufficient, but not necessary, condition for rationality with respect to these projects. Of course, you might be perfectly rational and yet fail to complete these projects due to unforeseen circumstances, bad luck, etc. One might argue that instrumental rationality should also determine what counts as an acceptable tradeoff. We discuss this issue in section 6.

[35] Again, with some notable exceptions such as avoiding an oncoming bus, etc.

[36]We believe that this assumption of supervenience is widespread though usually tacit. . A notable exception is Paul Weirich who spells out the assumption (and endorses it): "A wide variety of acts, such as swimming the English Channel, take more than a moment to complete... An agent cannot perform such acts at a time. How does the principle of optimization address such acts? It does not compare the extended acts to alternative extended acts. Instead, it recommends an extended act only if each step in its execution is an optimal momentary act. Swimming the English Channel is recommended only if starting, continuing moment by moment, and

finishing are all optimal" (Paul Weirich, *Realistic Decision Theory* (Oxford University Press, USA, 2004), p. 18).

[37] The possibility of top-down irrationality is discussed also in Tenenbaum, "The Vice of Procrastination."

[38] For the notion of an agent-centered permission, see Scheffler 1994.

[39] See again the remarks by Weirich cited in note 36.

[40] What counts as "acceptable" is determined by the nature of your end or project. Roughly, we can say that if your project has been completed in a way that you (correctly) recognize to have achieved the end of the book, the completion is acceptable. Strictly speaking, "acceptable" just reiterates that your actions do count as a proper completion of the project (or achievement of the end). Thus, for instance, having my friend who owns Vanity Press, Inc. decide to publish my collected facebook posts as a book won't count, since my project was not to write a book *in this way*. Given that the project is vague, there might be cases in which it is not clear whether my completion of the project is acceptable.

[41] Of course, in certain situations this permission might be cancelled. Working on your book might not be rational if one of your options at the time is to keep your dialysis appointment.

[42] Similarly, the atoms of a car engine are not individually necessary for the engine to function properly. Rather, they are all parts of the engine, which is itself necessary for the proper functioning of the car.

[43] For the notion of an INUS condition, see J.L. Mackie, *The Cement of the Universe* (Oxford: Clarendon Press, 1980),

[44] Here again condition (ii) is stronger than necessary (since *some* of the tokens of T might be necessary), but the stronger condition makes the presentation much simpler. A full account of these conditions would require an account of individuation of actions that singles out the right action types, but this a difficult issue on its own and trying to settle it would lead us away from our main purposes in the paper.

[45] A broader definition would have “believe” replacing “know” in condition (iv) and in (1). But we can’t go here into the complexities introduced by false and, in particular, unjustified beliefs; however, nothing in our argument depends on the use of “know” over “believe”.

[46] Of course, exercising too many permissions generated by a project might prevent you from completing another project and thus result in your failure to satisfy a requirement from the extended perspective. But suppose that exercising too many permissions in completing one project doesn’t prevent you from completing any other. Could you be irrational simply because you did significantly more than was necessary and never pursued your most preferred option (e.g., surfing, or even

resting) from the punctate perspective? We are inclined to think the answer is “yes”, but nothing we have said so far commits us to that; further argument would be needed. Since our analysis of the self-torturer puzzle is independent of this issue, we leave it aside here.

[47] More precisely, certain sets of momentary actions are such that none of the actions considered in isolation would violate the instrumental requirement, but the set of all taken together does violate it. Certainly *some* possible momentary actions are violations of the instrumental requirements considered in isolation; we use the imprecise formulation for simplicity.

[48] As we explained above, “vague projects” is being used broadly. See p. 18.

[49] Whether it is implicated will depend on whether Barry is already rich enough.

[50] Michael Bratman, "Toxin, Temptation, and the Stability of Intention," in *Faces of Intention* (Cambridge: Cambridge University Press, 1999). We believe that our criticisms of Bratman will extend to any planning solution. For another interesting planning solution, see Chrisoula Andreou, "Temptation and Deliberation," *Philosophical Studies* 131, no. 3 (2006), 583-606.

[51] Bratman, "Toxin, Temptation, and the Stability of Intention," 95.

[52] In drawing this inference one employs evidential, rather than causal, decision theory (or some similar evidence-based theory of rational choice). This might be cause for concern, but Bratman contends that even those unsympathetic to

evidence-based decision theory should accept the inference. For present purposes we assume that Bratman is right on this point.

[53] Of course, we do not deny that having a plan might be prudent if, for instance, ST prefers not to have to deliberate each time she needs to decide whether to move to the next setting. We claim only that ST is neither required to make a plan nor prohibited from reconsidering her plan if she makes one. Our view is compatible with the claim (in fact it implies) that ST is *permitted* to make a plan and to act on it.

[54] MacIntosh thinks that the original ST problem does not follow this pattern, on the ground that as ST turns the dial back, there must be a first setting at which she finds the pain tolerable. However, perhaps when ST reaches the first setting she considers tolerable, she would now consider the previous setting tolerable too. (Such a "revisionary" pattern of judgments might constitute a so-called hysteresis effect; see Diana Raffman, *Unruly Words: A Study of Vague Language* (New York: Oxford University Press, forthcoming), , Chapter 5). We thus discuss MacIntosh's proposal as if applied directly to the original self-torturer puzzle.

[55] This is not to say that your stopping where you do is inexplicable: presumably some brute psychological mechanisms determine when you actually stop. Thus brute mechanism causes you to stop where you do, but it doesn't *justify* your stopping there.

[56] Again, assuming that she cares only about financial gain and freedom from pain.

[57] One might want to know just which axioms of RCT are simplifying idealizations. For instance, perhaps the axiom of transitivity is one: perhaps there is nothing intrinsically irrational about an agent who has nontransitive preferences or even chooses in cycles. For present purposes we are non-committal on the matter. All we claim is that *some* axioms of RCT must be simplifying idealizations insofar as they rule out the preference structure of ST.