The Puzzle of the Self-Torturer

Author(s): Warren S. Quinn

Source: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, May, 1990, Vol. 59, No. 1 (May, 1990), pp. 79-90

Published by: Springer

Stable URL: https://www.jstor.org/stable/4320117

WARREN S. QUINN

THE PUZZLE OF THE SELF-TORTURER*

(Received in revised form 18 July, 1989)

Suppose there is a medical device that enables doctors to apply electric current to the body in increments so tiny that the patient cannot feel them. The device has 1001 settings: 0 (off) and 1 . . . 1000.[1] Suppose someone (call him the self-torturer) agrees to have the device, in some conveniently portable form, attached to him in return for the following conditions: The device is initially set at 0. At the start of each week he is allowed a period of free experimentation in which he may try out and compare different settings, after which the dial is returned to its previous position. At any other time, he has only two options — to stay put or to advance the dial one setting. But he may advance only one step each week, and he may *never* retreat. *At each advance he gets $10,000.*

Since the self-torturer cannot feel any difference in comfort between adjacent settings, he appears to have a clear and repeatable reason to increase the voltage each week. The trouble is that there *are* noticeable differences in comfort between settings that are sufficiently far apart. 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.[2]

The self-torturer is not alone in his predicament. Most of us are like him in one way or another. We like to eat but also care about our appearance. Just one more bite will give us pleasure and won't make us look fatter; but very many bites will. And there may be similar connections between puffs of pleasant smoking and lung cancer, or between pleasurable moments of idleness and wasted lives.

In all these cases, we find a mix of transitive and intransitive preferences. The self-torturer's *step-wise* preferences are intransitive. All things considered, he prefers 1 to 0, 2 to 1, 3 to 2, etc. . . . but certainly not 1000 to 1. This is why he cannot say that any setting is

Philosophical Studies **59**: 79—90, 1990.

better than the previous one. "Better than" is, while his step-wise preferences are not, transitive. But when he compares settings so far apart that he prefers the earlier setting, his preferences are transitive. If he prefers 500 to 1000 and 0 to 500, then he prefers 0 to 1000. This seems to permit us to say that 1000 is worse for him than 0.

The self-torturer's preferences are considered and well-informed. Before forming them, he freely experiments with all the relevant settings. And he is well-informed about the pleasures and advantages that different amounts of money can buy. Despite this, many theorists would condemn his intransitive preferences as irrational.[3] But this response may be too hard on the self-torturer and too easy on the theorist. The self-torturer's intransitive preferences seem perfectly natural and appropriate given his circumstances. They are the very ones most of us would have in his place (and *do* have in structurally similar, everyday situations). To insist that he get new, more 'rational' preferences might well invite bad faith. He wants to know how he should act on the ones he actually has. Intuitively, this question does not seem to be one that must lack a satisfactory answer. So the theoretical convenience of rejecting the question must be weighed against the failure to address what seems to be a genuine problem. In any case, I shall assume provisionally that the self-torturer has, as he is, a real problem of rational choice: How to take reasonable advantage of what the device offers him without ending up the worse for it.

I

Let's begin by considering some objections to my description of the case — objections that would point the way to some familiar way of solving the puzzle or some further grounds for rejecting it.

(1) *The self-torturer's preferences are changing*: On this objection, he is like an addict. At 0, the thought of 1000 appalls him, but at 999 it looks good. The changes wrought in him between 0 and 999 affect his outlook. Since he now (at 0) prefers not to change, he should resist any advance that has that effect.

But his preferences do not change. Even at 0 he prefers 1000 to 999. The same holds for plans of sequential choice. He always prefers a

plan that would take him to setting s to one that would take him only to $s - 1$. This is why he finds it so hard to set an initial overall plan.

(2) *We are neglecting behavioral evidence*: Someone might not be able to notice introspectively that his comfort had declined in a single step, even though his demeanor or behavior indicated otherwise — he might look less comfortable, or be grouchier. But while this is a possibility, it is also possible that the individual increments of current are too small even to have these effects. And this is the case I want us to consider.

(3) *We are ignoring the measures of his discomfort*: If we assign the self-torturer a discomfort-index, quantifying his discomfort at each setting, the numbers will have to change as he advances. Since he starts, we may suppose, at 0 discomfort and ends in great pain, there must be some *first* setting s with a positive discomfort index. So, whether he knows it or not, his comfort would decline in stepping from $s - 1$ to s.

But the measure of the self-torturer's discomfort is *indeterminate*. There is no fact of the matter about *exactly* how bad he feels at any setting. And if so, we cannot argue that the measure of his discomfort must increase in some single step.[4]

(4) *We are ignoring the effects of "triangulation"*: The self-torturer can triangulate a difference between s and $s + 1$ if he can find some third setting s' that feels the same as s but better than $s + 1$. And if he can use 0 to triangulate such a difference, then it is obvious that his comfort, at least compared to 0, declines in stepping from s to $s + 1$.

But 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.[5] The self-torturer (and his observers) try to triangulate differences between adjacent settings but honestly cannot — not because they are inattentive, but because we have made the increments of current too small to make *any* difference in comfort, even one that can be detected only by triangulation.

(5) *We are ignoring the reversal of his preferences*: The self-torturer starts out preferring early settings to 0 but ends up preferring 0 to late settings. At some intermediate setting his preferences must reverse.

There must be some s such that he prefers s to 0 but 0 to $s+1$. And that must mean that the added money at $s+1$ isn't worth the extra discomfort.

This argument goes wrong, I think, in presupposing that for any positive setting s, the self-torturer (counting both pain and gain) determinately prefers s to 0 or 0 to s. But, empirically speaking, his preferences as between s and 0 can exhibit various kinds of *indeterminacy*. Not only is there no empirically determinable *first* setting that he disprefers to 0, there is no empirically determinable *first* setting at which these preferences become indeterminate. There is simply nothing in the way a single increment of current affects him to warrant such precise line drawing. This is implicit in the failure of triangulation. A first setting at which his overall preferences (as between 0 and s) reversed or went indeterminate, would have to be a setting at which his comfort (relative to 0) suddenly declined.[6] There could be no other explanation. But then the self-torturer himself, or observers, could detect the decline, if only by the evidence provided by his reversing preferences. And we are considering a case in which the increments of current are too small to have any such effect.

(6) *Because the self-torturer's preferences are paradoxical, they cannot present genuine problems of rational choice*: He feels no worse at 1 than at 0, and the comfort-comparison between 1 and 0 is the same for him as the comparison between 2 and 0. From this it seems to follow that he feels no worse at 2 than at 0. Reiterated enough, the argument implies that he feels no worse at 1000 than at 0. Here we find a kind of *sorites puzzle*. Some of his clear and immediate judgments about his comparative comfort are true only if others are false. Surely preferences based on such paradoxical discriminations are not to be taken seriously. The self-torturer must either change his preferences or give up the mix of vague and precise terms that generates the puzzle.

It would be fair enough to discard a practical puzzle that *depended upon* the empirically false conclusions of a sorites argument. But the self-torturer's problem doesn't depend on any such conclusion. What naturally matters to him is that the comfort status of s and $s+1$ are, introspectively and behaviorally, no different — either in direct comparison to each other or in oblique comparison with any third setting. It is enough for him that the empirical data give him *no* reason to

suppose that his comfort declines, either directly or relative to some fixed point, in any single step. His predicament in no way depends on his supposing something that he can see to be empirically false or dubious.

And to say that the self-torturer should give up thinking in the vague terms that give rise to the sorites puzzle (terms such as 'more painful than', 'no less comfortable than', etc.) is to say that he should give up thinking about his real predicament. Whether something will or will not be less comfortable than something else is precisely what matters to him. It might be barely plausible to recommend that he purge such terms from his scientific psychology. But it seems bizarre to advise him to remove them from his practical deliberations.

II

The self-torturer's situation seems to defy conventional solutions. Let's look at a couple. It has been suggested that he can solve his problem in something of the style of a utility maximizer. Since 'just as comfortable' is intransitive, he cannot infer from the fact that 1 is just as comfortable as 0 and that 2 is just as comfortable as 1, that 2 will be just as comfortable as 0. So he cannot infer that his comfort won't decline in the overall move from 0 to 2. To make things simple, we may imagine an ultra-conservative self-torturer who is always eager to collect $10,000 so long as he remains *just as comfortable* as he was at 0, but regards $10,000 as too little to justify *any* increase of discomfort relative to 0 (i.e., any entry into the discomfort zone). This gives him a reason to move from 0 to 1. But the next move is a different matter. Then he must reckon with the probability that in moving one more step he may in fact begin to experience a decline relative to 0. And as he continues to advance, that probability must rise, until it cancels the expected advantage of the next payment. At that point a rational self-torturer would stop, well short of any disaster.

Of course, there is no solution here for a less conservative self-torturer who feels, as most of us might, that a minimal decline relative to 0 would be worth $10,000. But it is far from clear that even the ultra-conservative self-torturer's problem is solved. The alleged solution asserts that there is some chance that in taking a further step (e.g. from

1 to 2, or from 2 to 3), he will suddenly feel slightly worse than he did at 0; and this seems to presuppose that, for all we know, there is a step at which this actually happens. In the case we are imagining, however, there is simply no evidence of any such change. This is implicit in the failure of triangulation. The increments of current have been made so small that not only is there absolutely no indication of a subjective change from s to $s + 1$, there is absolutely no indication that any third point p has a different subjective relation to s than to $s + 1$.[7]

To press the alleged solution in the face of this is to suppose that there can be subjective contrasts that the self-torturer, and those observing him, cannot identify. But if this is true, the "solution" is needlessly subtle. Why suppose that the self-torturer's ability to identify these contrasts breaks down only when he tries to triangulate? Why not say that it already breaks down in his direct comparisons of adjacent settings? Indeed, why not say that he immediately feels worse whenever there is *any* undetectable increase in current, no matter how small? But this line of thought threatens to undermine the distinction between the physically objective and the experientially subjective. For given the possibility of an undetectable subjective contrast, that distinction would surely presuppose exactly what we seem to lack: a principled way of distinguishing stimuli differences that are too slight to be felt at all from those just strong enough to be felt undetectably.

Another possible risk-based solution might seem to side-step this problem. When deciding whether to advance a step, any self-torturer must take into account not just the danger of encountering unacceptable discomfort at the next setting, but also the danger that, if he takes the step, he will *eventually* find himself (as a result of further moves) in some unacceptable discomfort. At some point short of disaster, this latter danger may be too great to justify taking even one more step.

The trouble with this solution is that it cannot work for a self-torturer who assumes that he will always act rationally. If such an agent supposes that the risk of eventually finding himself in unacceptable discomfort is too great to warrant taking the next step, it must be because he thinks it will be rational for him to go on taking enough more steps to get himself into trouble.[8] But this means the next step is irrational only if *no subsequent step* that would occur before he found himself in unacceptable discomfort is irrational in the same way. For if

some such future step would be irrational, he could foresee that he wouldn't take it and hence wouldn't be risking trouble by taking the next step. So there is a good objection to the next step (from s to $s+1$) only if there won't be a good objection to, say, the step after that (from $s+1$ to $s+2$). But how could this be, given that both steps are intrinsically innocuous and that stopping at $s+1$ is just as sure to prevent disaster as stopping at s.

Perhaps risk-based solutions won't work. But if so, what is to prevent the self-torturer from adopting an even simpler solution? Why not pick a reasonable looking stopping point, proceed to it, and then *really stop*? If he could execute such a strategy, he could enjoy his material gains at a cost in discomfort that would seem to him well worth it. The trouble is that, even if we waive the theoretical difficulty of determining such a stopping point, he would be tempted to formulate a *new* forward-looking strategy once he reached it. And it is far from clear how conventional accounts of rational choice could oppose this temptation. For they see as irrelevant everything about a present choice except the way it serves the agent's preferences for outcomes. They are therefore ready to dismiss any past strategy that now requires him to forego something that he would in fact prefer to get all things considered. Strategies continue to have authority only if they continue to offer him what he prefers overall. Otherwise, they should be changed. We might call this natural idea the *Principle of Strategic Readjustment.*

This principle is compatible with all familiar desiderata of rational choice. It in no way speaks against long-term planning or temporally extended policies.[9] Policies need to be monitored to see whether they are still serving our preferences. And when they are not, they need to be adjusted. But it is precisely this familiar and seemingly innocuous idea of strategic readjustment that leads to trouble in the kind of case we are considering.

III

How then should the self-torturer proceed? There are, I think, two different kinds of case. To distinguish them we need a new notion. Suppose that instead of moving one step at a time (which is, in fact, his only way of advancing), the self-torturer could move from 0 to 1000 in

roughly equal-sized hops. At each landing point he would collect all the money attaching to the settings he had traversed. Call the sequence of positions he would occupy in such hops *a filtered series* of the original 1001 positions. Over some of these series the self-torturer's preferences would be transitive. In one kind of self-torturer case, some of these transitive series would have a position better than 0. Let's consider such a case first.

The self-torturer must begin by setting an initial strategy. This will consist in selecting a *reasonable* stopping point — a final goal. But this is not an easy task. For if s looks like a reasonable goal, won't $s + 1$ look better? And won't he then be on the slippery slope? Yes, but since he can see this, he can see that this is no way to select a reasonable goal. Instead, he might imaginatively restructure his problem. He could look for some orderly way of constructing a set of increasingly refined filtered series (each new series in the set containing more positions than the preceding one) such that one of them is identifiable as the most refined series in the set that clearly preserves transitivity of preference *and* contains a position better than 0. One natural way to try to do this would be to divide the overall range into two steps as nearly equal in size as possible (giving in our case the filtered series 0, 500, 1000), then into four smaller such steps (0, 250, 500, 750, and 1000), then into eight, and so forth.[10] If the most refined transitive series containing a member better than 0 contains only one such member, that should be his goal. If the series contains more than one best member, then either would be an equally reasonable goal. Suppose, for example, that the most refined such series was (0, 250, 500, 1000). If 250 is best, it should be his goal. If 250 and 500 are tied for best, then either (or perhaps any setting in between) will serve.[11]

Of course, this is only a sketch of a possible solution for setting a reasonable goal. But suppose we press ahead to a second problem. Suppose he has chosen a reasonable goal by this procedure, and has finally reached it. How can he now rationalize adhering to his original strategy? As we have seen, the idea of adopting a new strategy of continuing, say, only one more step has this in its favor: it will take him to a setting he prefers, all things considered, to the setting he is at.

But of course he could foresee all this when he chose his original

goal. And it is not that he now sees something wrong with that choice. It was, let us suppose, as good as any he could have made. Intuitively, his recognition of these facts should keep him from abandoning his original strategy. He should be stopped by 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, the Principle of Strategic Readjustment is suspended. *An agent is not rationally permitted to change course even if doing so would better serve his preferences.*

The other kind of self-torturer case offers a more radical challenge at the very outset. Since in these cases there is *no* transitive filtered series with a member better than 0, the procedure cannot be used to set a goal. But if the self-torturer simply picks some setting that he prefers to 0 as a reasonable goal, e.g., 1, he will find himself in trouble. For when he reaches 1, his preferences will importune him to change strategy and pick a *new* goal, e.g., 2. And so on.

It therefore seems that there is no way for him to take advantage of the opportunities that confront him at the start. If he could proceed, for example, to 1 and then stop, he would get a $10,000 cost-free gift. Surely a plausible theory of rationality would give him a way of seizing that or some even more attractive prize. But, again, it could do this only by setting aside the Principle of Strategic Readjustment — by giving his choice of a reasonable strategy that makes no mistakes about later developments binding authority over later choices. The self-torturer's predicament thus reveals a quasi-deontological aspect to a fully adequate theory of rational choice. It is this aspect that provides the theoretical 'brake'.

IV

Such a conception may remind one of David Gauthier's idea that it can be rational for an agent to honor a past agreement (in which the other party has performed his part) even though it would be disadvantageous for him overall to do so.[12] Gauthier has cases in mind that, simplifying somewhat, satisfy two conditions: First, the agent benefitted more from

securing the agreement than he would now lose by honoring it. And second, since he was unable to deceive the other party, he secured the agreement only by having the sincere intention to honor it. Since it is doubtful that rational agents can intend to do what they think it will in fact be irrational to do, either they cannot enter into such advantageous agreements or they are fully rational in honoring them. Gauthier is loathe to think that reason might stand in the way of these important advantages and therefore chooses the second alternative.

But perhaps it is not intolerable to suppose that rationality might get in the way of someone burdened with certain *inabilities* to maximize advantage. Gauthier's type of agent is unable to hide his true character. If he were a better actor, he could secure advantageous agreements by seeming to be transparently faithful while actually being opaquely faithless. So perhaps we should simply say that his inability to dissemble combines with his rationality to create a problem for him. That is unfortunate (from his point of view), but his misfortune might seem a doubtful reason to qualify the theory of rationality.[13]

The problems of the self-torturer are different. No inability stands in his way. It isn't that he lacks the will-power to stop at some reasonable initial goal. It is that a certain specious ideal of rational choice drives him onward — by advising him always to bring about some preferred outcome. It is the Principle of Strategic Readjustment itself that prevents him from getting and keeping the real advantages that his situation offers him. The self-torturer's predicament therefore invites, not a compromise theoretical accommodation to a standard human limitation, but a real revision of the theory of what it is for the best-positioned kind of agent to act rationally. His problem warns us that it would be a bad mistake for even the most advantageously endowed to see every moment as a possible new beginning in their practical lives.

NOTES

* I've profited from discussions with Alan Nelson, David Erikson, Yoram Gutgelt, Tony Martin, members of a seminar at New York University, and, especially, Kit Fine.

[1] The device derives from Derek Parfit's case of the Harmless Torturers in *Reasons and Persons*, Clarendon Press, Oxford, 1984, pp. 80—82. Parfit uses the devices to present, as he puts it, a puzzle in moral mathematics. I have appropriated them in order to present a puzzle of rational choice. The idea was provoked by Parfit's denial that the

self-interest theory of rationality could be, as he puts it, directly self-defeating. See p. 55.

[2] But I am supposing that even at 999 he will want things that an extra $10,000 can buy. Suppose he is a devoted philatelist who, even in severe pain, would value important new acquisitions — or a philanthropist who would still care about helping others. And if it seems too hard to believe that the marginal attractiveness of each new dollar would remain the same, imagine the payments compensatingly increased as he grows richer and more uncomfortable.

[3] For a classic discussion of some difficulties arising in other cases of intransitive preferences see Donald Davidson, J. C. C. McKinsey and Patrick Suppes, "Outlines of a Formal Theory of Value, I," *Philosophy of Science* 22 (1955), p. 146. Prominent among these difficulties is the way someone with intransitive preferences can become a "money pump". The self-torturer is quite different in this regard: he is more of a money vacuum. Amos Tversky, in "Intransitivity of Preferences," *Psychological Review* 76 (1969), p. 45, introduced cases in which intransitive preferences seem natural because the subject is understandably indifferent to very small differences in some variable even though he cares very much about larger differences. Our case is certainly of this sort. Note that other writers have suggested various ways of ruling out various allegedly offensive aspects of intransitive preferences. See, for example, Thomas Schwartz, "Rationality and the Myth of the Maximum," *Nous* 6 (May 1972), pp. 97—117; and Dennis Packard, "Cyclical Preference Structures," *Theory and Decision* 14 (1982), pp. 415—466.

[4] The objection could not be met by assigning determinate measures *with determinate margins of error* — saying, for example, that his comfort index at setting s was n plus or minus m. For it would still seem arbitrary to say that the range in which his comfort index lies has n as its precise center and $2m$ as its precise outer limits.

[5] Remember that, technology aside, we can make the increments as small as we like. We might, for example, have 100,000 settings with each increment reduced to 1/100th its original size. Then we might offer the self-torturer $100 for each advance and allow him to make 100 advances every week. And so on. Also note that, strictly speaking, I need claim only that the increment is too small to make a subjective difference *for the worse*. For it seems possible that a barely noticeable change might be completely unobjectionable in itself even though wholes composed of many such changes are very bad.

[6] Sharply enough to make it questionable whether an initial plan to stop at s was really worth $10,000 more than a plan to stop at $s - 1$!

[7] Nor are they able to find subjective differences in more complex contexts that vary only by the substitution of $s + 1$ for s. And this is not because the self-torturer or his observers are rushed or distracted. There is no reason not to allow the self-torturer and his observers to make their comparisons and observations with all due care and concentration.

[8] Or if he sets mixed strategies for advancing, he thinks it will be rational for him to adopt mixed strategies for future decisions whose combined effect will be an unacceptably high chance that he will get into trouble.

[9] And it is in no way incompatible with the familiar idea that an overall policy or plan (in which there must be a succession of choices) is best regarded as a *single* decision. See, for example, Leonard J. Savage, *The Foundations of Statistics*, John Wiley and Sons, New York, 1954, pp. 15—16. Savage's "look before you leap" advice is, as I understand it, directed not at all to the special kind of case we are considering here but to the altogether familiar situation in which long-term planning is required by the fact that the utility of a present option depends on the utility of the future options it makes possible.

[10] When the steps have to be slightly unequal, there will be alternative subdivisions that

are equally eligible. For example, in the series with sixteen steps, the first step could be either to 62 or 63. In such cases, it won't matter which series is chosen. Intuitively, there are many positions that represent acceptable compromises between wealth and comfort.

[11] Will there always be a clearly identifiable most refined series that preserves transitivity? That seems to be an empirical question. The intransitivity in these cases results, as we have seen, from the original steps being so small. As we increase the size of the steps (by progressively coarser filters on the original problem) we eventually reach steps that are clearly big enough for the discomfort to be registered and to be taken account of. In such series preferences will be transitive. In between, there is an (indeterminately limited) range of step-sizes for which the self-torturer's preferences are neither determinately transitive nor intransitive. Call this the cross-over range. If one or more of the filtered series have steps whose sizes fall within the cross-over range, then it may be indeterminate which progressively more refined series is the last to preserve transitivity of preference. In such a case, the self-torturer will simply have to pick some clearly transitive filtered series as refined enough for his purposes.

[12] *Morals by Agreement*, Clarendon Press, Oxford, 1986. See Chapter VI pp. 157–189. The case we are considering is one in which noncompliance will have no bad further effects on the agent's ability to reach agreements. The obligation might, for example, have been assumed in secret and the promisee might now be dead.

[13] Of course I am thinking of rationality (as I have been throughout) as *instrumental* — as something that is and ought to be the slave of the agent's preferences. On a more comprehensive *moral* picture of rationality, the disposition to comply with a fair agreement would no doubt count as a virtue and therefore something desirable to possess and, derivatively, to act upon. This is the kind of moral solution to Gauthier's problem I favor. But note that there is no moral question raised by the case of the self-torturer.

Department of Philosophy,
University of California,
Los Angeles, CA 90024,
U.S.A.