Tenenbaum and Raffaman on vague projects, the Self-Torturer, and the sorites

Article

Accepted Version

Tenenbaum and Raffaman on Vague Projects, the Self-Torturer, and the Sorites*

Luke Elson; luke.elson@reading.ac.uk

Forthcoming in *Ethics* — please cite published version.

Abstract. Sergio Tenenbaum and Diana Raffman contend that 'vague projects' motivate radical revisions to orthodox, utility-maximising rational choice theory. Their argument cannot succeed if such projects merely ground instances of the paradox of the sorites, or heap. Tenenbaum and Raffman are not blind to this, and argue that Warren Quinn's Puzzle of the Self-Torturer does not rest on the sorites. I argue that their argument both fails to generalise to most vague projects, and is ineffective in the case of the Self-Torturer itself.

1 Self-Torture and Rational Choice

We often attach value (utility) to *vague projects* which—like 'is a heap'—lack precise satisfaction conditions, but such projects engender puzzles for rational choice theory. You may wish for a restful night's sleep, but to stay up as late as possible consistent with that. Since *restful* is vague, one minute of sleep apparently couldn't make the difference between a restful and a non-restful night, and you ought to stay up for another minute. But foreseeably, if you keep thinking that way, you will stay up all night. To get a restful night, you must at some point reject such momentary calculations. But simply saying 'don't even go down that road' is a bad response to this puzzle: you would never leave your bed.

The challenge is to *balance* or *weigh* your competing goals. Philosophers—most recently Sergio Tenenbaum and Diana Raffman[1]—have argued that orthodox utility-maximising ra-

*For comments and discussion, I am indebted to several anonymous reviewers, John Broome, Ryan Doody, Thomas Hofweber, Brad Hooker, Douglas MacLean, Katherine Meehan, Julia Nefsky, Ram Neta, David Oderberg, C.D.C. Reeve, John Roberts, Geoffrey Sayre-McCord, Keith Simmons, Susan Wolf, and audiences at the North Carolina Philosophical Society, Chapel Hill, Reading, and York.

[1]Tenenbaum and Raffman (2012).

tional choice theory cannot do so, and that it fails under conditions of vagueness. In this paper, I defend orthodoxy against Tenenbaum and Raffman's argument.

Their argument appeals to a classic of this genre, Warren Quinn's *Puzzle of the Self-Torturer*, which I restate only briefly here:[2]

> **Puzzle of the Self-Torturer.** A torture device is attached irreversibly to a person 'ST'. The device has a dial—currently set to 0, with settings up to 1,000—which can only be turned up, in single increments, every week; each turn of the dial permanently increases the amount of electricity running through ST's body. The difference between adjacent settings is very small, but higher settings are agonising. Every time she turns the dial, ST gets $10,000.

In the original puzzle, 'the self-torturer cannot feel any difference in comfort between adjacent settings …[and] …appears to have a clear and repeatable reason to increase the voltage each week'.[3] A natural question is how putatively *imperceptible* differences in electrical current could amount to severe differences in pain. But our main problem—that of vague projects—arises even if each setting is perceptibly but slightly more painful than the previous one. With the right preferences, even fairly severe lifelong pain could be 'worth' $10,000. We set aside questions of imperceptibility for the moment, and return to them later.

Now, if ST has turned the dial k times, she must decide whether to advance to setting $(k+1)$. Orthodoxy says she must: doesn't the utility from $10,000 outweigh the disutility of a small increase in electric current? Since k was arbitrary, by parity of reasoning, she must advance to the end. ST indeed seems to have 'clear and repeatable reason' to turn the dial each week.

But this is foreseeably the route to disaster. If ST is like most of us, she would far rather turn the dial ten times and gain $100,000, than advance all the way into agony. Utility maximization has led him astray. This seems to show that you do better by deviating from a central tenet of orthodoxy:

> **Tenet.** At every choice point, act to maximize (expected) utility.

[2] See Quinn (1990), especially p. 79, for the canonical presentation.
[3] Quinn (1990), p. 79.

That is the Puzzle: ST seems always required by Tenet to advance one more stage. But she can see that if she continues to do so, she will eventually disprefer the outcome. Similarly, you would prefer to get 7.5 hours of sleep and feel rested, but Tenet apparently keeps you up all night. And if we do better by disobeying Tenet, how can an account of rational choice that includes it be correct?

2 Heterodox Views

On this basis, many philosophers defend strikingly *heterodox* accounts of rational choice. Quinn himself argued that the Self-Torturer has genuinely intransitive preferences, and that the Puzzle 'reveals a quasi-deontological aspect to a fully adequate theory of rational choice'.[4]

More recently, Sergio Tenenbaum and Diana Raffman have argued that Tenet does not apply to vague projects, including ST's goal of avoiding pain:

> 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.[5]

They argue that a refusal by ST to turn the dial cannot be justified on the grounds of utility, but 'the pain-free life project issues permission to stop turning the dial, independently of what maximises utility in light of ST's momentary preferences'. Even if utility is maximized by advancing one stage, ST has permission to refuse.[6] Sometimes, even though it would maximize utility for ST to turn the dial now, she is rationally permitted (perhaps required) to refrain. Utilities do not serve as a guide to life:

[4] Quinn (1990), p. 87.
[5] Tenenbaum and Raffman (2012); p. 102.
[6] Tenenbaum and Raffman (2012), p. 106.

> 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.[7]

Clearly, this view is supported by the Puzzle of the Self-Torturer only if utility-maximisation indeed fails in that case, by leading ST to disaster.

Since these cases hang on vagueness, a natural thought is that they are somehow instances of the paradox of the sorites, or heap. In that paradox, it is compelling that removing one grain from a heap of sand leaves a non-heap: that if *n* grains form a heap, then *(n-1)* grains form a heap. But if 10,000 grains form a heap, this implies (via repeated modus ponens) that five grains form a heap, even in contexts where this is clearly false. Such reasoning is a sorites on 'is a heap'. It is vague where we tip from a heap into a non-heap, and this renders the predicate 'sorites-embeddable'. Sorites hang on *tolerance* principles such as 'one grain couldn't be the difference between a heap and a non-heap'; tolerance principles (at least in their universally quantified forms) are widely agreed to be false, though compelling.

Theories of vagueness fall into two broad camps. *Indeterminists* claim that there is an indeterminate minimum number of grains required for a heap, perhaps because meaning depends on use and our use has not fixed a precise threshold. The relevant tolerance principle has a false instance, but it is indeterminate which. On *epistemic* views, some instance of the tolerance principle is determinately false, but we don't (perhaps can't) know which.[8] It is common ground that universally quantified tolerance principles are false: one of their instances is false, but it is indeterminate or unknowable which.

If the Puzzle is simply an instance of the sorites, then the challenge to orthodox rational choice theory is liable to dissolve. If claims such as that turning the dial is always required on utility-maximising grounds, and that ST has 'clear and repeatable reason' to turn the dial—which are at the heart of the putative counterexample to orthodoxy—are equivalent to a a tolerance principle, then they are false.

Of course, Tenenbaum and Raffman are not blind to the challenge of the sorites. But in their long and rich paper, they mention the paradox only once:

[7] Tenenbaum and Raffman (2012), p. 111.

[8] For indeterminism, see Fine (1975), Keefe (2000), Dorr (2003), and Barnes (2010). For epistemicism, see Sorensen (1988) and Williamson (1994). The 'tolerance' terminology is due to Wright (1975).

> 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 $10,000 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.[9]

This argument involves two claims: that vague projects involve distinctively *comparative* tasks, and that they therefore cannot be seen as a sorites. In the next sections, I'll argue that neither stands up to scrutiny.

3 The Shepherd in a Practical Sorites

The quoted argument appeals to distinctive features of ST's situation which put 'pressure on her rationality'. But since Tenenbaum and Raffman are defending a general thesis that vague projects are exceptions to orthodox rational choice theory, an appeal to distinctive features of one particular case of vagueness—such as the 'comparative' nature of ST's plight—is a non-sequitur. Indeed, many of the vague projects they discuss—such as that of writing a book—*do* involve trying to decide where the books end and the non-books begin, and are akin to that of sleeping restfully.

In this section, I'll argue that the decision-theoretic puzzles arising from such non-comparative cases are most naturally seen as sorites. The challenge of vagueness was first articulated by Richard Tuck, in a somewhat overlooked example:

> He could be a shepherd who wishes to build a cairn of stones by himself to guide him in the hills. On setting out in the morning, he can reason as follows. If I work all day, I will have a suitable pile of stones by nightfall. But one stone added to a collection of other stones makes a negligible difference – it can never

[9] Tenenbaum and Raffman (2012), p. 88 fn. 3.

be enough to tip it over the edge and into a heap. It takes a certain amount of time and effort to find a spare stone. If I do not start immediately, I will still have a heap of stones at nightfall, since the stone I could have picked up in the next few minutes would have made no difference to the outcome. But the same applies to the next stone, and the next: there is no point in ever beginning. Moreover, at some time in the day it will be clear that I have passed the stage where I will have enough time to build a cairn, and after that point there is certainly no benefit to be gained by piling up stones.[10]

Once again, it appears that utility-maximisation foreseeably leads to a dispreferred outcome, since the shepherd will not build a cairn. Maximisation seems to prevent him from starting the job.

To bring out how vagueness impacts the shepherd's predicament, consider how his situation looks when made artificially precise:

The precise cairn-builder. Another shepherd, John, will get 200 utils from building a cairn by nightfall. A cairn consists of 15 stones; once he starts building, he cannot stop. Each stone takes 30 minutes to move, and costs 4 utils (he will miss one episode of his favourite television show). He has ten hours before nightfall; the cairn will be useless if not completed by then.

Intuitively, it is clear that John ought to enjoy 2.5 hours of television, and then build a cairn with precisely 15 stones before nightfall. He will gain 140 utils: 200 utils for the cairn, minus 60 utils from moving 15 stones. (We set aside questions about whether it is really rational for him to leave it to the last minute like this, without setting aside a 'margin of safety', which form a different set of issues.)

If John starts earlier or later, then he is irrational: either he fails to build a cairn, or he piles up an excessive number of stones, to no additional benefit. Starting late and then building a partial cairn is clearly the worst strategy. He not only gets no value from a partial cairn, but also loses value from the television programmes he misses, for no compensating benefit. If he has left it too late to build a cairn, then he should not start.

[10] See Tuck (1979) p. 154. Cairns are heaps of rocks or stones, often used to mark mountain trails.

As expected, Tenet explains each of these judgements. Every thirty minutes, John must start moving stones, or wait. If he has already waited *k* stages of thirty minutes, then applying Tenet, he ought to wait for stage *(k+1)* just in case utility is maximized by doing so. In utility terms, stage *(k+1)* differs from *k* in just two relevant ways—more leisure time, and one less stone on the cairn.

Stage *(k+1)* always involves more leisure time; *(k+1)* is worse in cairn-terms only if the cairn can be built at *k*, but not at *(k+1)*. Utility is maximized by waiting for *(k+1)* rather than starting to build at *k*, *unless* the cairn is buildable at *k* but not at *(k+1)*.[11] And 'the cairn is buildable at k but not at (k+1)' is false at all points except one—the point when there is enough time left to carry exactly 15 stones. In this precise case, Tenet correctly requires that John start work at the last point when he can still finish the cairn by nightfall.

But what about the original, vague case? As Tuck notes, this case is clearly parasitic on the vagueness of 'is a cairn', which is sorites-embeddable. The relevant tolerance principle is:

(Cairn-Tolerance) If *n* stones form a cairn, then *(n-1)* stones form a cairn.

We thus get a sorites series: 20 stones form a cairn; if 20 stones form a cairn, then 19 stones form a cairn; … if 6 stones form a cairn, then 5 stones form a cairn. But, 5 stones do not form a cairn. Since it is better to build a cairn with fewer stones if possible, we also get a sorites on 'better':

It is better for John to build with 19 stones than 20;
If 19 stones are better than 20, then 18 are better than 19;
…
If 6 stones are better than 7, then 5 are better than 6.

I use 'better' subjectively, to mean that an option is preferable utility-wise. By the transitivity of 'better' on such a characterisation, and via a sorites series, we reach the false

[11] I'm assuming here that the following situation does not obtain: stage 2 is worse than stage 1, but stage 3 is better than both stage 2 and stage 1. Such 'darker before the dawn' cases are a little more complicated, but not fundamentally different.

conclusion that it is better for John to build with 5 stones than with 20. This is how the cairn-builder is caught in a sorites series.

We may call this a *practical* sorites. Unlike the standard 'theoretical' case, John cannot simply withhold judgement, or say 'it's borderline': he must act according to the application of a predicate. Is this the minimum number of stones required for a cairn, or not?

Instances of the tolerance principle (Cairn-Tolerance) fall into three groups. For high numbers of stones (clear cairns), it is clearly true. In a third group, with low numbers of stones (clear non-cairns), (Cairn-Tolerance) has a false antecedent and is clearly true. In between, (Cairn-Tolerance) has a false instance. So much is common ground between the precise and the vague cases.

But when it is vague which instance is false, the central instance becomes a *penumbra*, where there are borderline-cairns composed of middling numbers of stones. Here instances of the principle are borderline, construed neutrally between indeterminism and epistemicism: indeterminate or unknowable. Whatever the theory of vagueness, the tolerance principle (Cairn-Tolerance) has a false instance in this penumbra.

We can now see that Tenenbaum and Raffman's argument doesn't support the *general* thesis that vague projects do not engender sorites. The shepherd *is* 'proceeding along a sorites series of [piles of rocks] from a [clear cairn] to a [clear non-cairn], attempting to decide where the [cairns] end and the [non-cairns] begin'. Some vague projects can be seen as sorites.

4 The Self-Torturer as Practical Sorites

But utility-maximization is a general theory, and just one counterexample would falsify it. Even if Tenenbaum and Raffman are wrong about many other vague projects, what about their argument that the sorites could not be 'the source of the puzzle' for ST?

It has some force. In the paradox of the heap, 1,000,000 grains form a heap, and if *n* grains form a heap, then *(n-1)* grains form a heap. The Self-Torturer doesn't seem to be like this. What predicate corresponds to 'is a heap'? They imply that it would be something like 'is a bearable level of pain'. A sorites on *this* predicate would, as they rightly say, misdescribe

the puzzle. ST is making a comparative judgement—is this additional pain worth $10,000?—which doesn't seem reducible to a fruitless search for the edge of a predicate's extension.

But it is a mistake to think that a sorites account of the Self-Torturer must have this form. It is not *quite* explicit that Tenenbaum and Raffman do think this, but if not, then their argument is manifestly ineffective. A rational ST is certainly not looking for the last bearable pain: intuitively, she should stop long before that point. If there is a sorites here, it is not one on a predicate like 'is a bearable level of pain', but on something more comparative: is this pain increment worth this money increment? In focusing on non-comparative predicates, Tenenbaum and Raffman underplay the resources of the sorites view.

Further consideration of the shepherd's case suggests how ST may also be caught in a sorites. The shepherd is actually performing both of the tasks distinguished by Tenenbaum and Raffman. He is proceeding along a sorites series of [piles of rocks] from a [clear cairn] to a [clear non-cairn], attempting to decide where the [cairns] end and the [non-cairns] begin; he is also trying to decide whether a certain incremental difference in [the number of stones] can be compensated by [extra leisure time] at that point in the spectrum of [stones].

For the shepherd, the outcome of the latter 'comparative' task is determined by that of the former 'categorical' task. Tenenbaum and Raffman are correct that ST doesn't seem to face such a categorical task. But their mistake is to think that therefore the case cannot be seen as a sorites: we can exhibit a sorites 'directly' on the comparative task.

4.1 Two Kinds of Vague Project

To do show that ST is caught in a sorites, I wish to introduce a distinction between two kinds of desires, or projects. First:

Binary desire. A *binary* desire that a is F is unsatisfied if a is not F, and satisfied if a is F.

Binary desires divide worlds or states of affairs into those where they are satisfied, and those where they aren't. My desire to go to Churchill, Manitoba, and see the polar bears, is binary: it is satisfied in worlds where I go, and unsatisfied in worlds where I do not. The desire for a cairn is similarly binary, despite its vagueness: the shepherd's desire is satisfied *iff* he has a cairn. Your desire for a restful night's sleep is also binary and vague.

But some desires are not like this:

Essentially comparative desire. An *essentially comparative* desire is for things that are more F rather than less F. There is no 'Fness threshold' beyond which adding Fness is not desired.

Essentially comparative desires do not divide worlds, but *rank* them. Instead of desiring to see a polar bear, I might prefer to see larger mammals over smaller ones. I prefer seeing an elephant to seeing a polar bear, to seeing a Scottish Wildcat. One cannot say outright whether such a desire has been satisfied—satisfied compared to what?

The shepherd's desire to watch more television is comparative, as are ST's desires for more money and less pain, and your desire for more waking minutes. In the rational choice literature, comparative desires or projects are often marked with phrases like 'utilities linear in dollars'.[12]

Crucially, in the essentially comparative case, given a situation, there is both a better and a worse situation. For some amount of money, there is a preferred (all else being equal) situation where you have more, and a dispreferred (all else being equal) situation where you have less. (Perhaps this is not so in extreme cases: one might be genuinely indifferent between one trillion and two trillion dollars, given actual facts about the world.)

This distinction explains a structural difference between the cairn-builder and the Self-Torturer. If the shepherd waits too long, then he ought to watch more television, and at least salvage something of the day, since he gets no utility from a partial cairn. Thus he seems to pass gradually from 'there's no point starting now—it would be overkill' to 'there's no point starting now—it's too late'. The shepherd's binary desire for a cairn becomes rationally inert once it can no longer be satisfied.

But we would not say that once ST 'has gone too far' and regrets turning the dial so many times, she ought to keep taking the deal. That would just make things worse. Even once ST has passed the optimum trade-off of pain and money, there are many settings that are clearly *worse*, with respect to her essentially comparative desire for less pain.

[12] See, for example, Elga (2010) p. 4.

4.2 A Model of a Self-Torture Sorites

Now, we can construct a model of ST's plight as a sorites series. Our main aim is to use a minimal set of assumptions to adequately explain the key features of the Puzzle, and to show that the comparative structure identified by Tenenbaum and Raffman is no barrier to seeing the case as a sorites.

Again, let's imagine a fully precise version of the puzzle. From Quinn's description, we obtain several intuitively plausible constraints:

(I) ST cares only about money and pain: her net utility is the sum of the utility of the money and the disutility of the pain. At setting zero, she has none of either, so her net utility is zero.

(II) The marginal utility of taking the deal is initially positive: at setting 2 with $20,000, for example, ST's net utility is positive. This is plausible, and implicit in Quinn's discussion of *filtered series*.[13]

(III) ST eventually has *negative* total utility, since she prefers setting zero to some later stages.

For precise utility functions, then, here is how we might graph ST's utility:

This graph might need some justification. Here I have for ease of presentation treated machine settings as continuous rather than discrete (this is wholly unrealistic, but does not substantially affect the case). Money utility is modelled as 5 times machine-setting, and pain disutility is modelled as approximately 1.333 raised to the power of the machine setting.[14] These functions are arbitrary, but other utility functions which respect constraints (I)-(III) will—assuming no gross discontinuity—generate a graph with a similar structure.

We now see that when ST crosses from zone 1 (where the slope of the money-utility line exceeds that of the pain-disutility line) into zone 3 (where the relative slopes are reversed), there is a point—call it zone 2— where the net-utility line peaks and then turns down. (Since earlier dial-turns have positive marginal utility, but later ones negative, her utility functions for machine-settings and for money cannot both be linear.[15]) After this, the

[13] Quinn (1990), p. 86.

[14] More precisely, the pain-utility function is e to the power of (W(50)x/10), which is chosen for neatness: ST ought stop at setting 10.

[15] I am grateful to Richard Yetter Chappell for forcing me to be clear on this point.

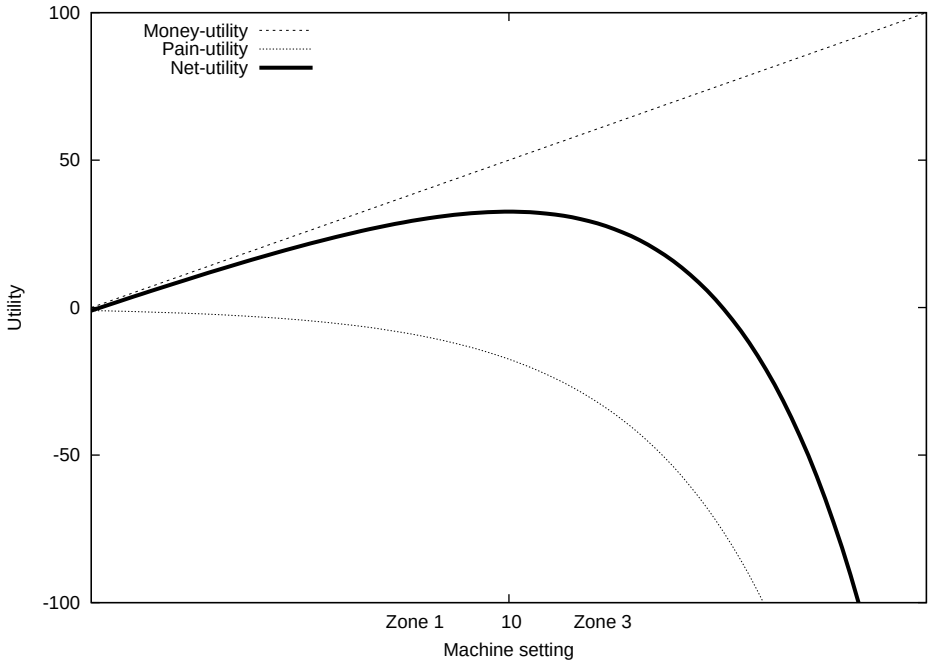

Figure 1: The precise Self-Torturer's utility functions.

relative gradients change, and advancing further—into zone 3—decreases net utility. Here the slope of the pain-line exceeds that of the money-line.

We are most interested in the point where net utility 'tops out' (zone 2), where the difference between accumulated money-utility and pain-utility is maximized. Clearly, ST ought stop here, and orthodoxy explains this. Suppose utility is maximized at setting 10, as it is on the graph above. Before that point, the marginal utility of turning the dial is positive; but at stage 10, marginal disutility of the extra pain at stage 11 outweighs the marginal utility of the extra money, and turning the dial is impermissible. Orthodox utility-maximization correctly mandates that ST accept the deal until setting 10—the point with the highest net utility—and then stop.

In this precise version, ST should, according to both intuition and Tenet, stop at stage 10, maximizing utility. Unlike in the precise shepherd case, there is nothing akin to the last stage at which a cairn is buildable. Here there is no simple categorical project that we can 'look through' to, so we must speak directly in terms of weighting ST's two projects, of less pain and more money. Nevertheless, the orthodox view renders the correct verdict.

In a vague version of the Self-Torturer, ST's utility functions are vague: it is vague at what setting the slope of the pain-line overtakes that of the money-line. Speaking metaphor-

ically, zone 2 is 'smeared' into a region, just as the zone of where a cairn becomes a non-cairn is smeared.

We thus get a sorites on 'maximises utility': for low settings (zone 1), it is false that utility is maximized; in a penumbra (zone 2), it is vague whether utility is maximized; later (zone 3), it is again false. The tolerance principle is:

Torturer-Tolerance. If setting *k* does not maximize utility, then setting *(k+1)* does not maximize utility.

This formulation is a material conditional, so it is true except at that point when the pain-utility begins to dominate money-utility (where the derivative/slope of the former overtakes that of the latter). The tolerance principle is false only at that point at which utility is maximised; I'll argue that there is such a point—though its location may be vague.

As we saw, the difference between the binary and essentially comparative desires explains how one ought behave once one has already 'gone too far'.

In zone 1, it is determinately true that the marginal utility of turning the dial is positive: the net-utility line has a clearly positive slope. In zone 2, it is borderline whether the marginal utility of turning the dial is positive or negative. In zone 3, the marginal utility of turning the dial is clearly negative: each stage is worse than the previous.

Thus we can explain why once ST has proceeded too far, into zone 3, repeatedly advancing is clearly disadvantageous. In the binary cairn builder case, the net utility slope 'after the borderline zone', so to speak, is positive. In the comparative self-torture case, it is negative.

The Puzzle of the Self-Torturer is only superficially simple: besides the imperceptibility we have set aside, it requires a non-trivial strategy of ST: stop when the derivative of her pain-utility function exceeds that of her money-utility function. This opaque structure grounds an objection: why believe Torturer-Tolerance? It's not as intuitively compelling as 'if *k* grains are a heap, then *(k-1)* grains are a heap'. But the sorites story works only if the principle is plausible enough to explain the judgement that each particular stage is preferable to the previous one.[16]

[16] I owe this objection to an anonymous reviewer.

The first response to this objection is to dodge it: we *do* find such a judgement compelling—otherwise the Self-Torturer scenario wouldn't be so troubling—and in this sorites, Torturer-Tolerance is equivalent to that judgement. So the compelling plausibility of the judgement/tolerance principle is agreed background.

But we might hope that a theory of vague projects would explain the claim's plausibility, not just appeal to it. The plausibility of tolerance principles is an important question in theorising about vagueness. In the original paradox, for example, when we imagine that the penumbra is about 100 grains 'wide', we can imagine why the relevant tolerance principle is compelling: for any particular grain, it seems incredible that *that* grain could be the cut-off. By parity of reasoning, so it seems for any grain.

We might say something similar here. Plausibly, our preferences are not completely determinate and knowable, and it is vague which setting is optimum, or where *pain begins to outweigh money*, so to speak. Anyone who accepts this, and the structural constraints on the Puzzle in the precise case, must accept that zone 2—where utility is maximised, which depends on our preferences—will be a penumbra rather than a point. Given that there are a thousand settings on the device, this penumbra may be rather wide: for any setting therein, it seems incredible that *that* setting is the tipping-point. After all, the penumbra of 'is a cairn' only perhaps five stones 'wide', and that was sufficient to ground a sorites.

With this machinery, we *can* somewhat artificially cast ST's plight in non-comparative terms. Though she is not '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', ST is proceeding along a sorites series of dial-turns from those with clearly positive marginal utility to those with clearly negative marginal utility, attempting to decide where the positive/required ones end and the negative/impermissible begin.

5 Conclusion

If the foregoing is correct, then Tenenbaum and Raffman's case for heterodoxy is undermined.

It is false that the shepherd always maximises utility by waiting, and thus false that orthodox rational choice theory condemns him to never building the cairn. Those claims

are equivalent to the principle (Cairn-Tolerance), since he always ought wait *iff* no stone makes the difference between a cairn and a non-cairn. Given that such tolerance principles are false, orthodox rational choice theory does not so condemn him, and his predicament doesn't falsify the orthodox view.

And since Torturer-Tolerance has a false instance, there is a k such that setting k doesn't maximise utility, but setting *(k+1)* does. According to orthodox rational choice theory, ST ought turn the dial to setting *(k+1)* and no further. The decision-theoretic constraints are similar to the precise case, except that k is vague: indeterminate or unknowable.

Since the universally-quantified tolerance principle is equivalent (given our plausible constraints about the structure of the puzzle) to the claim that ST always maximizes utility by turning the dial, we know that the latter is false. The central challenge of the Puzzle was this: can we reconcile the intuitive need to stop before *somewhere* with the apparent utility-maximizing requirement to advance at every point? But since the Puzzle is a sorites, there is no such requirement, and heterodox 'permissions' to diverge from utility considerations are otiose.

This line of argument, of course, hangs on the falsity of universally-quantified tolerance principles, which could be denied. The argument is therefore conditional on the truth what we might call a 'tolerance-denying' account of vagueness.[17] But such accounts are both overwhelmingly plausible and overwhelmingly popular. In any case, doubts here do nothing to support Tenenbaum and Raffman's argument that vague projects do not ground sorites.

If seeing the case as a sorites undermines the motivation for heterodox views, can it contribute a positive story? A full account is beyond the scope of this paper, but here we may sketch an orthodox, utility-maximising explanation of the key intuitive facts about ST: she ought to turn the dial a few times and then stop; she is rational if and only if she stops in some central zone: the borderline zone 2.

Epistemicists can easily explain this, as a mere case of action under ignorance (in this case, of her own preferences).

Indeterminists can also show that the borderline-zone is the only rationally permissible place for ST to stop advancing, and that no point therein is determinately preferable to any

[17] I'm grateful to Julia Nefsky for discussion of this point.

other.

At each setting in zone 2, it is indeterminate whether utility is maximised. In that zone, somewhere but indeterminately where, the slope of the pain-line exceeds that of the money-line, and net utility peaks. At each stage in that zone, it is indeterminate whether this turning point has not yet been reached (ST should advance), is at hand (ST should stop), or has already been passed (ST should stop, and retreat if possible). Each point in the zone is borderline-optimal, and none is determinately superior to any other.

So if ST stops in that penumbra, it is indeterminate whether she maximizes utility.

But if ST stops outside the zone, then it is determinately false that she maximizes utility. If she stops beforehand, then she could have done better by advancing further; if she stops afterwards, then it is determinate that she has gone too far. For each point outside the borderline-zone, there is determinately a point within that zone with higher utility.

Plausibly, if her options are (i) make it indeterminate whether she maximises utility, or (ii) make it determinate that she does not, then she ought to (i). This is not *in general* true—if one can either make it indeterminate whether one gets 100 or 10 utils, or determinate that one gets you get 80 utils, then one ought to do the latter—but the 'aggregative' considerations that ground exceptions don't seem to apply here.[18] If she ought to make it indeterminate whether she maximizes in this manner, then on the sorites account, the Self-Torturer is akin to an indeterminist Ass of Buridan: she should simply pick some point in the second zone.

None of this amounts to a general account of the Puzzle of the Self-Torturer. In particular, the issue that we set aside at the beginning—the apparent imperceptibility of the torture-increments—poses trouble for any view. If the increase in electrical current is genuinely imperceptible, then it is hard to see how refusing an extra $10,000 could *ever* be justified: pain is only bad because it feels bad, and if the difference is imperceptible then one feels no worse after accepting the money.

The central question is whether it is really coherent to suggest that the adjacent settings are indiscriminable: together with the claim that indiscriminability is transitive, it leads to the particularly nasty *phenomenal sorites*. Phenomenal sorites can be constructed for all kinds of predicates, with 'is loud', 'is red', and 'is cold' being particularly common examples. So

[18] For a representative denial that such aggregation can be justified, see Williams (2014).

the move to the perceptible case—not only in this discussion, but also in that of Tenenbaum and Raffman[19]—was a substantial weakening of the Puzzle. A full account of ST's plight has not been provided here or elsewhere, in terms of the sorites or ortherwise.

Nevertheless, Tenenbaum and Raffman have failed to show that the challenge of vague projects to orthodox rational choice theory is not, as they put it, 'just an especially picturesque instance' of the sorites.

References

Barnes, Elizabeth (2010). "Arguments Against Metaphysical Indeterminacy and Vagueness". In: *Philosophy Compass* 5.11, pp. 953–964.

Dorr, Cian (2003). "Vagueness Without Ignorance". In: *Philosophical Perspectives* 17, pp. 83–113. ISSN: 0301-1526.

Elga, Adam (2010). "Subjective Probabilities should be Sharp". In: *Philosophers' Imprint* 10.5.

Fine, Kit (1975). "Vagueness, truth and logic". In: *Synthese* 30.3, pp. 265–300.

Keefe, Rosanna (2000). *Theories of Vagueness*. Cambridge University Press.

Quinn, Warren S (1990). "The Puzzle of the Self-Torturer". In: *Philosophical Studies* 59.1, pp. 79–90. DOI: 10.1007/BF00368392.

Sorensen, R A (1988). *Blindspots*. Oxford University Press.

Tenenbaum, Sergio and Diana Raffman (2012). "Vague Projects and the Puzzle of the Self-Torturer". In: *Ethics* 123.1, pp. 86–112. DOI: 10.1086/667836.

Tuck, Richard (1979). "Is there a free-rider problem and if so what is it?" In: *Rational Action*. Ed. by Ross Harrison. Cambridge University Press, pp. 147–156.

Williams, J Robert G (2014). "Decision-Making Under Indeterminacy". In: *Philosophers' Imprint* 14.4, pp. 1–34.

Williamson, Timothy (1994). *Vagueness*. London: Routledge.

Wright, Crispin (1975). "On the coherence of vague predicates". In: *Synthese* 30, pp. 325–365.

[19] See, for example, Tenenbaum and Raffman (2012), p. 94.