Chapter 5: Self-Torture and other Vague Projects

Luke Elson

June 9, 2025

Abstract. I turn to Warren Quinn's famous Puzzle of the Self-Torturer, which is emblematic of cases where vagueness seems to engender intransitive or cyclic preferences. I defend a vagueness model of the Puzzle as what I call a repeating practical sorites. 'Solving' the Puzzle is important both as an application of the material from earlier chapters and to resist arguments that cyclic preferences are a rational response to the Puzzle and other vague projects.

Thus far I've focused on the assumption made by many—including Savage in his (P1)—that the weak preference relation $\preccurlyeq$ is complete. I've argued that incompleteness is vagueness, explained how our preferences could be vague, and defended a decision theory for action under vagueness. I've tried to avoid excessive departures from standard decision theory, because—I claim—incompleteness as vagueness can largely be accommodated therein. Supersharp says that preferences may be vague and thus incomplete in that sense, but it's supertrue that we have full preferences.

I now turn to transitivity. I wish to reject the possibility of rational *cyclic* preferences. In a preference cycle, there are some options such that $A \prec B \prec C \prec A$. Most of the putative cycles we'll see are longer than this, but it's clear in outline how such cycles could engender value pumps. Suppose you'd pay £10 for B over A, £5 for C over B, and £7 for A over C, then in a series of choices you could be led to pay £22 to move from A to ... A?

Such cycles are particularly threatening for two reasons. Unlike the earlier examples we've seen the trades in this pump (for example from A to B) seem non-optional. If it really is true that $A \prec B$ and that this preference is not irrational then—given the orthodox construal of such states, which I am defending—you are required to move from A to B. Similarly from B to C, then from C to A. This is a *requiring* value pump, not a *merely permissive* one. Rationality not only not only fails to save us from a sure loss (as was the threat in Sally's Bets, for example), but seems to impose one on us.

Moreover, the surgery on orthodox decision theory to avoid this consequence would be severe. If there are incomplete preferences then in a sense the orthodox view must simply be expanded to cope with them—this is what I've tried to do

with vagueness—but the fix doesn't require us to act *contrary* to the clear verdicts of orthodoxy under the existing relations of preference and indifference. But if there is a rational preference cycle then to avoid being value-pumped we must at some point refuse to do something—to refuse to move from B to C, for example—*even though we rationally, determinately prefer C to B*. Orthodox decision theory is not just incomplete, it's wrong in particular cases.

The unorthodox draw that conclusion. Sergio Tenenbaum and Diana Raffman argue that vague projects provide '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'. Chrisoula Andreou argues that 'disorderly' (in particular, cyclic) preferences are rational in certain cases, but that it's not rational to follow them into a value pump. So sometimes rationality requires us to violate our rationally-held preferences, to do something that determinately fails to maximise expected utility. These are decisive breaks with orthodoxy. [1]

I'll argue that those certain cases hinge on (you won't be surprised to read) vagueness. They are *practical sorites*: sequences of actions (compound actions) where each action is E-admissible and the sequence is foolish, but where *each action is individually compelling*.[2] That psychological force is the main difference between cases of incompleteness (like those we've seen already) and of apparent preference cycles.

In practical sorites both the psychology and the decision theory seem to face a challenge: to stop before you complete the foolish sequence, to vindicate stopping. As in the ordinary predicate sorites, it seems clear that we should stop at some point... just not at this point. *Pace* the opposition, I'll argue that the intransitivity in such cases is an illusion borne of vagueness, so standard decision theory needs supplementation to cope with that vagueness but not major surgery.

This is all very abstract, so here's an example—the earliest of which I'm aware, due to Richard Tuck.[3] A shepherd has the project of building a cairn, which is a heap or pile of stones. This project has vague completion conditions, because it's vague how many stones are needed for a cairn. The shepherd has another project, of minimising how much time he spends carrying heavy stones up the hill; each stone is a severe effort.

Consider a particular stone. It'll take some effort to move and, as (Tuck 1979, 154) puts it in the shepherd's voice, 'one stone added to a collection of other stones makes a negligible difference – it can never be enough to tip it over the edge and into a heap'. So he shouldn't move this stone; but of course the reasoning is repeatable and 'there is no point in ever beginning'. This can be so even if he would rather move all the stones than move none. Tuck rejects this reasoning; his focus is on groups and soritical free-riding which I consider in the next chapter.

[1] Tenenbaum and Raffman (2012), p. 102; Andreou (2023). Tenenbaum's view is further developed in Tenenbaum (2020).

[2] I was very proud of the term 'practical sorites', but Richard Tuck independently beat me to it. See (Tuck 2008, chap. 3).

[3] Tuck (1979), p. 154.

Practical sorites can arise when we have two competing goals, ends, or projects. If one of them (the vague project) has vague completion-conditions, but the other (the precise project) doesn't, then the sorities tolerance principle of the vague project can make it seem like some small actions—abandoning a stone, for example—cost nothing in terms of the vague project (we'll still have a cairn) but bring benefit in terms of the precise (we'll save effort). Thus these small actions or omissions seem required. But the small actions add up and soon the vague project is endangered. If we care more about the vague project than the precise project, then this is something we should seek to avoid. Academics may identify more with Sergio Tenenbaum and Diana Raffman's example of writing a book yet minimising time spent writing.[4]

As I mentioned, practical sorites like these have lead some to major revisions to standard decision theory: the opposition claims that we must act against our preferences, determinately failing to maximise utility. I'll argue that this is not correct, because it is vague which actions maximise utility in isolation. Like the tolerance principle for 'is a heap', the claim that skipping one stone or one hour of writing is always preferred may be compelling, but is false. There's a number of stones the shepherd should move and a number of hours I should write, and I maximise utility by doing that and no more, but the number is vague. Thus we needn't depart from the determinate verdicts of standard decision theory about whether to move the stone or write the paragraph. There *are* no determinate verdicts.[5]

1 The Cairn-Builder

Let's focus on the shepherd as Tuck stipulates it. The shepherd's only reason to move stones is to build a cairn, and he would prefer to move as few as possible. Given the vagueness of 'cairn', the following is a tolerance principle:

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

We can all agree that instances of tolerance principles ('if 17 stones form a cairn, then 16 stones form a cairn') seem plausible. If Cairn-Tolerance is true, then the shepherd always prefers to move one fewer stone.

But suppose that the shepherd cares more about having a cairn than he does about moving fewer stones, in the sense that he'd prefer to be the exhausted but proud owner of an over-engineered cairn with 17 stones than to relax the day away with an empty space where the cairn should be. Then we appear to have a preference cycle:

> he prefers moving 17 stones to moving 0 stones (better an excessive cairn than no cairn); he prefers moving 16 stones to moving 17 stones (saving some effort at no cairn-cost); ...; he prefers moving 0 stones to moving 1 stone (saving some effort at no cairn-cost).

Since I'm assuming a tolerance-denying account of vagueness, I'm assuming

[4] Tenenbaum and Raffman (2012), especially pp. 99–100.

[5] Some of the argument of this chapter builds on Elson (2016).

that Cairn-Tolerance is false and there is no such cycle. Given indeterminism about vagueness, instances of Cairn-Tolerance fall into three groups. For high numbers of stones (clear cairns), Cairn-Tolerance has a true consequent and is clearly true. For low numbers of stones (clear non-cairns), Cairn-Tolerance has a false antecedent and is clearly true. In between (penumbra of borderline-cairns), Cairn-Tolerance has a false instance—but it's indeterminate which one.

Suppose, for example, that the threshold is indeterminately 9, 10, or 11 stones: 8 stones is determinately not a cairn, and 11 stones is determinately a cairn. After how many stones should the shepherd stop and call it a day? Stopping after 9, 10, or 11 stones are each E-admissible actions. Other stopping places are not E-admissible. No sharpening sanctions stopping after 3 stones or 13 stones, for example: not enough for a cairn and wasted effort, respectively.

Let's suppose that moving one stone costs him 1 utile, but that a cairn is worth 15 utiles. Thus a cairn is worth more than the effort needed to build it, which must be so otherwise it would never be rational to build a cairn. There are 17 stones scattered in walking distance and a cairn requires c stones. I just stipulated that it's indeterminate whether c is 9, 10, or 11. If he has moved n stones, then he has $\lfloor n/c \rfloor$ complete cairns (n divided by c with no remainder), because any 'surplus' stones moved above c are wasted effort (there definitely aren't enough stones to make a second cairn). So when he has moved n stones, his utility is:

$u(n) = 15\lfloor n/c \rfloor - n$

This is 15 utiles for a complete cairn if he has one, minus n utiles for the effort of moving that many stones. As n increases by 1, $u(n)$ decreases by 1 almost every time—*except* when he completes a cairn, and his utility increases by 14 (15 utilies for the cairn minus 1 for the stone). His utility is decreasing except at the false instance of Cairn-Tolerance, where it sharply climbs.

Figure 1 graphs his utility function, with each line representing one sharpening of 'cairn'. Each sharpening has the same shape, but the climb occurs in a different place. The heights vary, because on different sharpenings it takes more or less effort to get a cairn, but on all of them a cairn brings 15 utiles.

At each point where Cairn-Tolerance is borderline (ie, around 9, 10, and 11 stones), the shepherd is in the penumbra of 'is a stage that maximises utility'—it's indeterminate whether net utility is maximised—so if he stops in that penumbra it's indeterminate whether he maximises utility. In other words, each stage therein is merely E-admissible (it is the peak of at least one utility function but not all of them).

If my arguments in Chapter 3 against compromise under indeterminacy were correct, then under Supersharp it's never permissible to make it *determinately false* that you maximise utility. My preferred rule Hierarchical Liberal says that any E-admissible stopping point is permissible (because there are no T-admissible ones available here). If we comply with that rule, then in coping with the vague project we never determinately fail to maximise expected utility.

The Shepherd faces what I'll call a *single-threshold* practical sorites, because there is one penumbra (around 'cairn'), and the challenge for rational choice is to

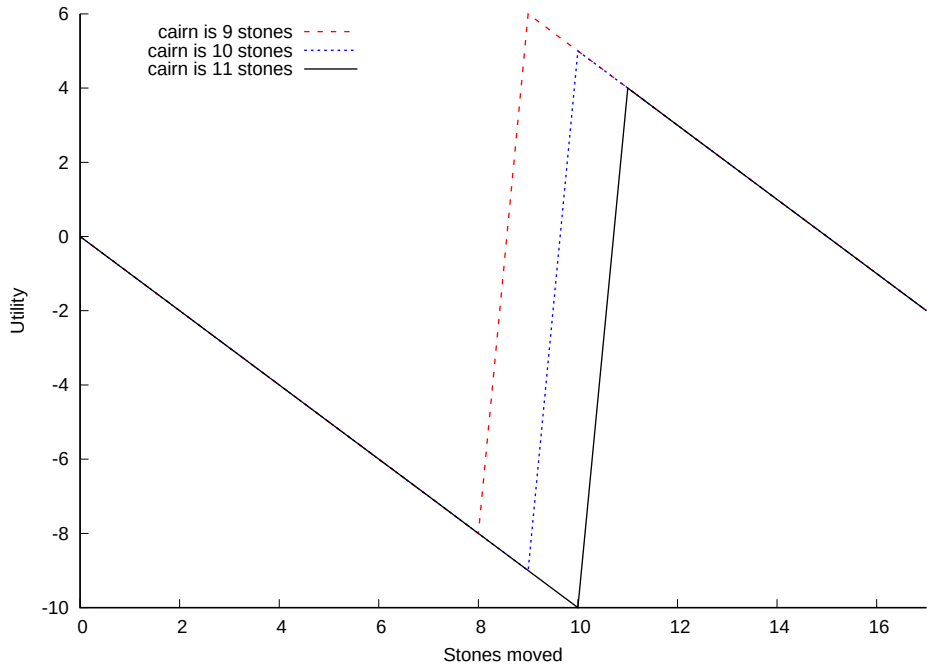

Figure 1: Building a cairn

stop in that penumbra. Or before, depending on his utilities. As can be seen in Figure 1, if he moves 12 stones he determinately fails to maximise his utility, and the slope from 11 to 12 stones is determinately down. So moving 9, 10, or 11 stones are each permissible actions according to Hierarchical Liberal.

2 The Puzzle of the Self-Torturer

But other vague projects are *repeating*, and far more challenging. The classic is Warren Quinn's notorious 'Puzzle of the Self-Torturer'; his description is unmatched:

> Suppose that there is a medical device that enables doctors to apply electric current to the body in increments so tiny that the patient cannot feel them. The device has 1001 settings: 0 (off) and 1 ... 1000. Suppose someone (call him the self-torturer) agrees to have the device, in some conveniently portable form, attached to him [...] The device is initially set to 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 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.*[6]

Let's say that stage n is the package of device at setting n and n bundles of $10,000: stage 500 is setting 500 with $5m, and so on. ST begins at stage 0, and is offered the chance to advance stage-by-stage. The puzzle comes from the com-

[6]Quinn (1990), p. 79. Emphasis in original.

bination of two apparent facts:

(1) ST prefers each stage to the one before—an extra $10,000 for only a tiny additional amount of electrical current!—and he looks rational to do so. As (Quinn 1990, 79) puts it, ST seems to have 'clear and repeatable reason to increase the voltage each week'.

(2) ST prefers earlier stages to later stages. Stage 10 is preferable to stage 800, for example. Again, this looks rational.

Together (1) and (2) engender intransitivity: ST prefers stage 1 to stage 0, and stage 2 to stage 1, and ... and stage 800 to stage 799, but prefers stage 0 to stage 800. Moreover, all of this is foreseeable even at stage 0.

For Quinn the Puzzle is non-reversible—once ST has advanced, there is no going back to earlier stages—which adds drama and even tragedy, but the puzzle remains even without this constraint, which costs us generality. Even if ST *could* retreat by turning the dial back and returning some money, the decision-theoretic puzzle would remain: where to stop?

Quinn assumes that the differences in electrical current between adjacent stages are *imperceptible*. This assumption—like all 'phenomenal' sorites, such as those involving indistinguishable colour patches—raises puzzling issues in the philosophy of mind and perception. Is it even coherent to talk of imperceptible differences in pain, for example? The sense that it might not be gives the Puzzle particular force.

If the *only* reason not to turn the dial would be that there is more pain at the next setting but we cannot perceive an increase in pain, surely that means there is no more pain at that setting and thus no reason to turn the dial? It seems clearly coherent to talk of imperceptible differences in electrical current as Quinn does, but here electrical current matters only because it causes pain.

Conceding a lack of ambition, I don't have a solution to the phenomenal sorites. But I don't think we need one: the central issue is vagueness, and vagueness arises both from the phenomenal and from the non-phenomenal. I'm concerned more with how to deal with the vague preferences, not so much with how those preferences arise from the phenomenal.

Here are two related lines of defence for my less-ambitious approach. First, even a neutered version of the Puzzle where each dial-turn brings *barely-perceptible* increases in pain raises similar decision-theoretic issues: a barely-perceptible difference in pain seems clearly worth it for $10,000 (or $100,000). And so we have a puzzle once more, albeit perhaps less striking than the original. There are many other non-phenomenal vague projects. Perhaps the cairn has phenomenal aspects (can he see it from a distance?) but the book doesn't, and to foreshadow some other examples, whether some piece of land or the atmosphere is wrecked or spoilt involves its biodiversity levels.

Second, my solution to the puzzle below does *not* rely on a phenomenal concept. My solution is that it's vague how many dial-turns cause a 'life-altering' increase in pain. Whether a pain increase is life-altering can be vague in a non-

phenomenal sense even if adjacent settings are phenomenally indistinguishable. Whether some pain increase is life-altering will depend on how much it affects your ability to carry out daily activities or saps your motivation to cook meals from scratch over the long term, for example. It is plausible that even an imperceptible increase in electrical current could push you over a threshold here. I can't say that I hear the traffic noise from my bedroom on a main road in Reading, but when I visit my family home in Llangennech I'm always struck by how much more peaceful I feel in the morning.

It might be thought a fall-back victory for orthodoxy if we could show that there are indeed rational preference cycles, but only in cases of genuine imperceptibility. But it would be hollow, because there are so many cases of imperceptibility and because almost every practical sorites can be made into such a case by increasing the number of steps involved. So to really meet the challenge, orthodoxy must deal with both kinds of case.

Confronted with the Self-Torture puzzle, nobody should advance all the way to the end, and almost nobody *will* advance all the way to the end. Nearly everybody would stop at a reasonably early stage. The vast majority of us do not proceed all the way to the end of the everyday self-torture scenarios that we face. When we set the thermostat in the winter, each 0.1C reduction brings a determinate financial benefit in return for an imperceptible or barely-perceptible reduction in comfort. But almost nobody who can afford to run the heat keeps it off all year: they find a setting that seems a reasonable trade-off of money and comfort.

As I mentioned, some people think that the Puzzle has revisionary implications for decision theory. Andreou (2023) claims that (1) and (2) are true, that ST has cyclic preferences, and moreover that such preferences are *appropriate* to his situation. Then to avoid disaster ST must refuse the deal at some point, even though he would prefer to take it at that point. Quinn himself argued that the Puzzle 'reveals a quasi-deontological aspect to a fully adequate theory of rational choice', and this rests on the intransitive nature of his preferences: '*an agent is not rationally permitted to change course even if doing so would better serve his preferences*'.[7]

You won't be surprised to learn that I think that the Puzzle is a practical sorites: there is no intransitivity because it's not true that ST determinately prefers each stage to the previous one.

In Elson (2016) I gave a model of the Puzzle as (what I now call) a single-threshold practical sorites, along the lines of the cairn, with the following tolerance principle:

Torturer-Tolerance. If stage k doesn't maximize utility, then stage $(k+1)$ doesn't maximize utility.

The core thought was that there's an optimum trade-off of pain and money, but it's indeterminate at what stage that occurs. Initially, ST's utility curve is clearly rising as the marginal money outweighs the marginal pain of turning the dial; eventually, ST's utility curve is clearly falling, as the marginal pain outweighs

[7] Quinn (1990), p. 87. Emphasis in original.

the marginal money. 'Maximises utility' has a penumbra, and at at each stage in that penumbra it's indeterminate whether net utility is maximised. Each stage in the penumbra in the borderline-optimal, and none is determinately superior to any other. So if ST stops in that penumbra, it is indeterminate whether she maximizes utility—in other words, each stage therein is E-admissible. The case is fundamentally similar to the shepherd's.

The most obvious mechanism for this to arise is through the diminishing marginal utility of money: rather than attaching constant utility to a dollar, how much you value each extra dollar depends on how many you already have. Each $10,000 is less valuable than the one before it, and from some stage onwards—it is indeterminate where—the $10,000 is not worth the extra pain.

To quote myself, this model implies that after the penumbral stages 'the marginal utility of turning the dial is clearly negative: each stage is worse than the previous' (Elson 2016, 485). Hrafn Asgeirsson, Bennet Francis, and Sergio Tenenbaum have convincingly argued that this is an inadequate model of ST's predicament.[8] It is correct for the Shepherd—once it's too late to build a cairn he'd lose utility by wasting time moving stones—but the following principle due to Sergio Tenebaum and Diana Raffman looks far more plausible in the case of the Self-Torturer:

Nonsegmentation. 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. (Tenenbaum and Raffman 2012, 98)

Because my earlier account gives ST a determinately downward-sloping utility curve after the penumbra, it is incompatible with Nonsegmentation's final clause. Assuming plausibly that stage 750 is beyond the penumbra, for example, it says that ST *clearly* prefers staying put to turning the dial to stage 751. So Nonsegmentation is false. This is the opposite conclusion to Andreou, who claims that at every stage—and thus at 750—ST prefers to turn the dial.

There are three possibilities for Nonsegmentation. First, it could be false as I argued in previous work, but as I say I have been convinced that this is psychologically and normatively implausible. Suppose that you *begin* the puzzle at stage 300: the device is attached to you (without your consent) and set to that stage, and you are now in very serious pain. You are also given $3mn, again without your consent and offered the chance to turn the dial to 301 for an extra $10,000. Notice that Nonsegmentation is disjunctive about such an isolated choice, so let's disambiguate:[9]

Nonsegmentation-Permissive. ... whereas in any isolated choice, she **may** choose to turn the dial.

Nonsegmentation-Requiring. ... whereas in any isolated choice, she **must** choose to turn the dial.

[8] Asgeirsson (2019); Tenenbaum (2020), pp. 94–100; personal communication from Bennet Francis.

[9] I'm indebted to an anonymous reader for OUP for comments here.

These Nonsegmentation claims are deontic, about permissible actions. Preferences are also relevant, because if the Puzzle is to engender cyclic preferences it must be the case that you *prefer* the next stage. In standard decision theory, including the version I'm developing in this book to cover indeterminacy, a determinate strict preference for the next stage would imply that it's permissible and required to advance to the next stage (in any choice, isolated or not), at least if the preference isn't irrational. That's why rational cyclic preferences would pose such a threat to the standard view, and why I need to explain away their appearance. Revisionary views, on the other hand, accept cyclic preferences as real and as appropriate but break their link with permissibility: it's impermissible to follow those cyclic preferences, beyond a few stages at least.

Let's try to inject a little psychological realism. It's not quite clear to me what you would or should prefer if you are 'dropped' into self-torture. If you begin the puzzle at stage 300 as I just described, we may suppose that you are facing a lifetime of serious but not debilitating pain, but also have a lifetime to spend your three million dollars. For only a little more pain, perhaps a barely-noticeable or imperceptible increment, you could have a substantial amount of extra money. $10,000 is not negligible for someone with wealth in the mere low millions. Do you prefer stage 301 to stage 300? I think it would be quite defensible to have preferences either way.

If adjacent settings are perceptibly different, then given the declining marginal utility of money there are likely some later stages (consider stage 850!) where both versions of Nonsegmentation are false, where utility has a determinately downward sloping curve as in my original model. This is where imperceptibility bites, supporting Nonsegmentation even at those high-pain much-money stages where it is hard to accept.

Even without the imperceptibility assumption, I think Nonsegmentation-Permissive is quite plausible at earlier stages. If 'dropped' at 300 and offered a one-shot choice to move to 301, I think many of us *would* take the deal and consider ourselves rational. Nonsegmentation-Permissive is plausible, especially early in the Puzzle, and doesn't require strong assumptions. I will thus defend it, with the possible exception of any final stages where the utility curve is determinately downwards.

Nonsegmentation-Permissive is all we need to get the puzzle going. It seems to engender a merely-permissive value pump: if turning the dial is permissible at every single stage, how could doing so impermissibly lead to disaster and make us rationally wish that we'd never taken the deal at all?

So why does my earlier solution fail? Because the Puzzle *repeats*. Even if ST reaches stage 250 and regrets doing so, doesn't she still face an apparently overwhelming reason to move to stage 251? At the very least it looks extremely tempting and we wouldn't judge any such move irrational in a one-shot choice. This may not be true in the later stages, but in the earlier ones it does seem clear that the next stage is at least *arguably* permissible even well past any point of maximal utility. There does seem to be 'clear and repeatable reason' to take the deal as Quinn puts it, perhaps not all the way to agony, but at least in moderate pain. "I

shouldn't have come this far, but it's too late to worry about that now—and the next deal still looks desirable." This goes beyond Nonsegmentation-Permissive, because we are no longer talking only about *isolated* choices. My previous solution failed to capture this, implying instead that after the point of maximal utility, taking the deal is determinately undesirable.

One way to patch things up would be to say that the Puzzle of the Self-Torturer is fundamentally similar in structure to that of the cairn-builder, but whereas there are only three E-admissible stages for the shepherd (9, 10, or 11 stones), the peak of ST's utility function is smeared out far further, and so several hundred stages in the self-torture series are E-admissible. But that wouldn't respect the phenomenology of the case. In the central stages of the Puzzle (the 300s, in the example I've been using), it seems plausible both that we might think we have gone too far, past the point of maximal utility (even if not yet worse-off than when we started), and at the same time think that turning the dial once more is permissible.

The parallel claim is *not* true of the Shepherd. Once he has gone too far, past the point of maximum utility, he won't make things worse by carrying on in the same way. If he's wasted most of the day and it's too late to build a cairn, then he rightly doesn't feel tempted to move just one or two stones. Unless they form a cairn, moving stones is a waste of time and effort.

We need a more complex model of the self-torture series, one that respects Nonsegmentation-Permissive for *any* isolated choice—or at least most of them—and captures the fact that whereas the Shepherd faces one locus of irrationality (failing to build a cairn) ST faces many, because things can keep getting worse for her.

3 Repeating Puzzles

Here is some terminology, for a general discussion:

n-trade. In an n-trade, the dial is turned n times and ST is given $10,000n.

ST is repeatedly offered 1-trades, and the central puzzle (given Nonsegmentation-Permissive) is why each 1-trade looks rationally permissible in isolation, yet the whole series—a 1000-trade—rationally forbidden.

The core idea of the new model is that longer trades (higher 'n') are impermissible, but below some threshold number for 'j', j-trades are merely E-admissible. 1-trades appear rationally required for the same basically psychological reason that sorites tolerance principles are compelling.

It's part of the setup of Quinn's case and seems undeniable that 1000-trades are impermissible. This is accepted by all parties. We all agree that setting 1,000 with $10mn is dispreferred to setting 0 with no money, and that if you move from the latter state to the former then you've done something impermissible.

But since we are accepting Nonsegmentation-Permissive, each 1-trade is permissible in isolation. As a preview, this will be because each 1-trade is E-admissible

but the 1000-trade is not admissible. Since 1-trades are E-admissible but 1000-trades aren't, there must be a boundary: some number k such that k-trades and longer are not admissible, but shorter trades are E-admissible. If k is 6, then anything up to a 5-trade is E-admissible (indeterminate whether \$10,000 is worth one dial-turn ... indeterminate whether \$50,000 is worth 5 dial-turns) but a 6-trade is not admissible, because six dial-turns are determinately not worth it for \$60,000.

We'll then bring in Chapter 4's Compound: turning the dial k times completes a foolish sequence, so Compound forbids the k^{th} turn, even if that turn would be permissible in isolation. As ST proceeds through the stages, she keeps violating Compound.

Complications may have already occurred to you. First, it's unlikely that k will be constant. Perhaps at stage 50 a 10-trade (10 dial-turns for \$100,000) is E-admissible but at stage 500 a 5-trade (5 dial-turns for \$50,000) not. Granted, but I think we can set this aside as a complication that obscures the underlying structure.

Second, what about the start and end of the Self-Torture sequence where Nonsegmentation-Permissive is less plausible, as we've discussed? There's plausibly an initial set of stages where turning the dial is clearly rational. It may well be clearly preferable to live at stage 10 (\$100,000 and a slightly painful tingling in one's arm) than at stage 0 for the rest of one's life. As we'll soon see, something similar can also happen at the *end* of the series on the repeating model without being built in as an assumption, which I count as a significant virtue.

Third, aren't I wishing away second-order vagueness, by assuming that there is no indeterminacy in k? I'll return to this issue in a few paragraphs.

Let me restate the contrast with the single-threshold practical sorites such as the cairn. That had a simple utility curve: a determinately upwards slope before a penumbra, a single peak at an indeterminate stage, and a determinately downwards slope afterwards. The repeating case is more complex. At each stage it's indeterminate whether the slope is increasing or decreasing. Thus there remains no intransitivity, because no determinate comparison of each stage to the previous one. But over the long run—over each span of k stages or more—the slope is determinately downwards.

This is all rather abstract, so it will be useful to once again illustrate with our shepherd, making the numbers smaller to make them more tractable and the pattern clearer. We will make his predicament repeat by putting more than one cairn in play, illustrating along the way that repeating practical sorites are not limited to rather odd puzzles such as that of self-torture. Suppose that procrastination has bankrupted the shepherd and he's looking to sell some of his materials:

> **Bankrupt Shepherd.** The shepherd has 51 stones. A cairn requires (indeterminately one of) 9, 10, or 11 stones and he attaches 15 utils to each cairn. His utility is linear in pounds, and he attaches no utility to 'spare' stones. A trader offers him £1 per stone. The trader will

buy as many stones as the shepherd is willing to sell, but can only take one at a time.

The shepherd ought to sell a stone unless it costs him a cairn. At first, he ought to sell his 'spare' stones:

- On sharpening s_9 a cairn is 9 stones, so he initially has 5 cairns and 6 spare stones;
- On sharpening s_{10} a cairn is 10 stones, so he initially has 5 cairns and 1 spare stone;
- On sharpening s_{11} a cairn is has 11 stones, so he initially has 4 cairns and 7 spare stones.

There is an initial segment—he determinately ought to sell one stone. Then things get interesting: on s_9 and s_{11} a second stone is also spare, but on s_{10} selling the second stone would cost him a cairn.

Let's say that he begins at stage 0; at stage n he has sold n stones for £n, so his utility is n utiles from the money received (the precise project), plus 15 utiles for each remaining cairn (the vague project). But it's vague how many cairns he has left. On sharpening s_k, his utility $u(n, s_k)$ at stage n is

$$u(n, s_k) = 15\lfloor (51 - n)/k \rfloor + n$$

Here $\lfloor (51 - n)/k \rfloor$ means "the number of stones he still has, divided by k with no remainder"—ie, the number of complete cairns he still has when a cairn is k stones. Since the value of k is indeterminate, his utility function is indeterminately increasing or decreasing at many stages, but the long-range slope is clearly downwards. Figure 2 shows the sharpenings of his utility function.

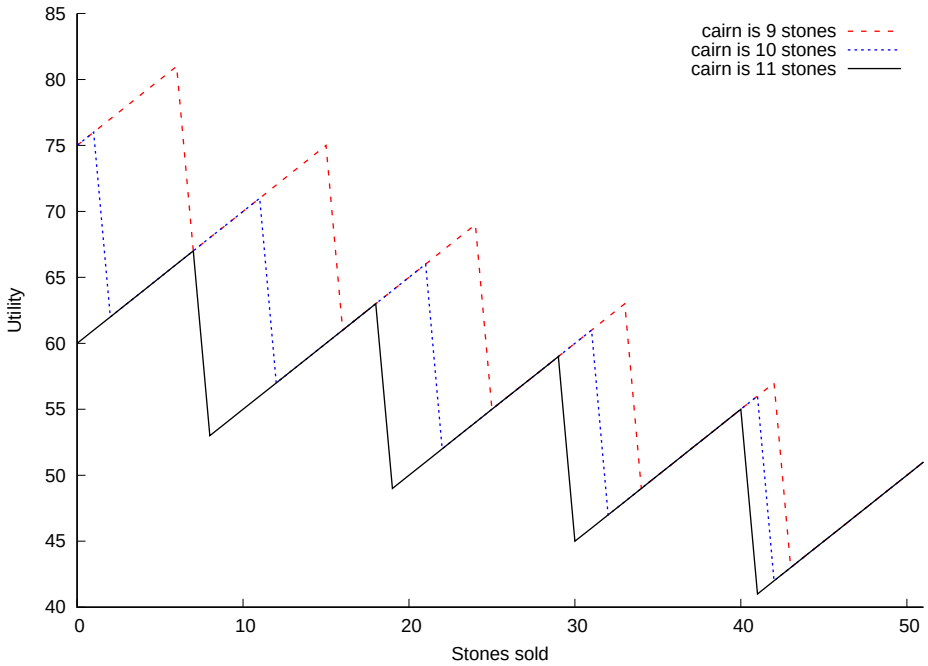

Figure 2: Bankrupt Shepherd utilities

The initial segment—the first stone he determinately ought to sell—is repre-

sented by his higher utility at stage 1 than stage 0 on every sharpening. Similarly, at the final stages he has no cairns on any sharpening, only surplus stones to sell, so the slope is determinately upwards.

But in between these extremes, you can see that it's indeterminate whether the utility curve is sloping upwards or downards: at many stages, both selling a stone and not selling a stone are merely E-admissible. (With more sharpenings this could easily have been the case at *every* stage, at the cost of cluttering the example.)

At longer ranges, the curve is determinately downwards. Every 11-trade is impermissible: across any gap of 11 stages, every sharpening falls. This is because a cairn will have been determinately sacrificed for its constituent stones, which are worth less separately. All 11-trades and longer are impermissible, and so Compound forbids completing 11-trades. We thus vindicate Nonsegmentation-Permissive for this case. Outside the initial and final segments, in any isolated choice he may sell a stone because doing so is E-admissible, but he must not complete a series of eleven 1-trades. (In this case at the start and end Nonsegmentation-Requiring is true.)

A word about Nonsegmentation. First, its 'in isolation' is to be interpreted thus: considering selling a particular stone as a genuinely isolated choice. This is not merely a psychological matter. What makes Nonsegmentation true is genuine path-dependence in the sense of Chapter 4. The shepherd's previous actions can change the rational status of this one: selling 10 stones makes selling this eleventh one impermissible; the previous sales make this final one the completion of a foolish (and thus impermissible) sequence. Assuming Liberal, Nonsegmentation-Permissive is vindicated because selling in isolation is E-admissible. Nonsegmentation-Requiring is *not* vindicated, because even in isolation refusing to sell is also E-admissible. Isolated sales may be psychologically compelling (they seem compelled by a sorites tolerance principle) but are not rationally required because that tolerance principle is false.[10]

Other than the specific numbers, the only major assumptions in this model are that the shepherd values a cairn more than he values the individual stones (a cairn is more than the sum of its parts), and that the vagueness of 'cairn' is reflected in his preferences. It's very difficult to deny that this pattern is commonplace. If anything is unrealistic about the model, it's the *precision* in terms of the range of indeterminacy of 'cairn'—that is, the lack of second-order vagueness—and in the amount of utility he attaches to each object.

Isn't assuming away second-order vagueness here more than elsewhere just dodging a central objection to a vagueness-based view? No. Second-order vagueness might make it vague *where* Compound tells us to stop, because *where* a foolish sequence is completed will also be (second-order) vague. But the aim here is to show the *existence* of a stopping-point, and we've done that, even if as a second-order supertruth. In particular, once we get away from any second-order penumbra (far above k), we'll be in the realm of clearly

[10] In this paragraph especially we are not talking about the very start and end of the series, where things may be different.

impermissible compound actions. I should note that second-order vagueness is one area where epistemicism shines, because it becomes another layer of ignorance and thus less distinctively threatening.

But we are here for self-torture, not cairn-building. Let's as before make the simplifying assumption that k (ie, the length of the shortest foolish sequence) is constant, say 10. When what we are looking for is that—outside of any initial segment where taking the deal is determinately preferable—the following structure obtains:

- every 1-trade is merely E-admissible (it's indeterminate whether the small extra pain is worth $10,000);
- similarly, every 2-trade, 3-trade, ..., and 9-trade is merely E-admissible;
- every 10-trade is inadmissible (it's determinate that the sizeable extra pain is not worth $100,000).

If we can construct this structure for the Self-Torturer, then we can explain why ST shouldn't turn the dial 10 times: it would complete a foolish sequence, and so be forbidden by Compound.

In the Self-Torturer, there are no cairns to build or stones to sell. This is about trade-offs between money and pain. A simple model has constant marginal utility of money; this preserves some of the analogy with the cairns, as the utility of selling a stone is constant. But just as there are thresholds when a cairn is sold (or built), there are threshold pain increases. I'll use the concept of a *life-altering* increase in pain. Whether a pain-increase is life-altering need not depend purely on the felt experience of the pain, but could involve shortened concentration span, irritability damaging personal relationships, and so on.

Just as a cairn is worth more than the sum of the stones, so a life-altering pain increase costs more than the sum of the increases in pain taken separately. Whether a pain increase is life-altering is not detached from the phenomenal character of the pain. It's a question of when that character and its consequences add up to something that outweighs the amount of money on offer.

In the model, let's stipulate that increasing the self-torture device ten settings determinately causes a life-altering increase in pain. Moreover, given your preferences, $100,000 is *not* worth a life-altering increase in pain. Any 10-trade is inadmissible and thus (according to Hierarchical Liberal) impermissible, because you determinately prefer to sacrifice that much money to avoid a life-altering pain increase.

But it's indeterminate how many dial-turns it takes to impose a life-altering pain increase: it's indeterminate whether it takes 8, 9, or 10 dial-turns. An 8-trade—taking $80,000 for 8 dial turns—is E-admissible, and permissible in isolation, because it's indeterminate whether the pain is life-altering. And if an increase of eight settings is not life-altering as I'm using the term, then it is worth enduring for $80,000.

Unlike the stone sales which involve steep collapses in utility, a life-altering pain increase doesn't need to make things *terribly* worse: it (perhaps together with the more mundane disutility of the pain) simply needs to determinately outweigh

the $100,000. It's easy to imagine, for example, that even a fairly minor caffeine headache for the rest of one's life could be dispreferred to that amount of money, if it stops one from truly relaxing and enjoying music.

A 10-trade determinately brings a life-altering pain increase, so is determinately not admissible. Taking the deal 10 times is a foolish sequence and thus forbidden by Compound, no matter how you break it up: it would be impermissible to accept a 10-trade synchronically, and so Compound forbids its completion diachronically. You may turn the dial 9 times but not 10.

Let's put some numbers on ST's preferences. We can split his utility function into three components:

- Utility from money is linear in thousands of dollars, so at stage n money-utility is $10n$.
- Disutility from 'mundane' pain is linear in stages, and at stage n is $-9n$.
- A 'life-altering' pain-increase contributes -12 utils, so at stage n the utility for that is $-12\lfloor n/s \rfloor$, where s is the number of dial-turns for such an increase (which is indeterminately 8, 9, or 10).

As can be seen from Figure 3, this model has fundamentally the same structure as the bankrupt shepherd: indeterminately rising or falling at many points, but down in the long run on all sharpenings.

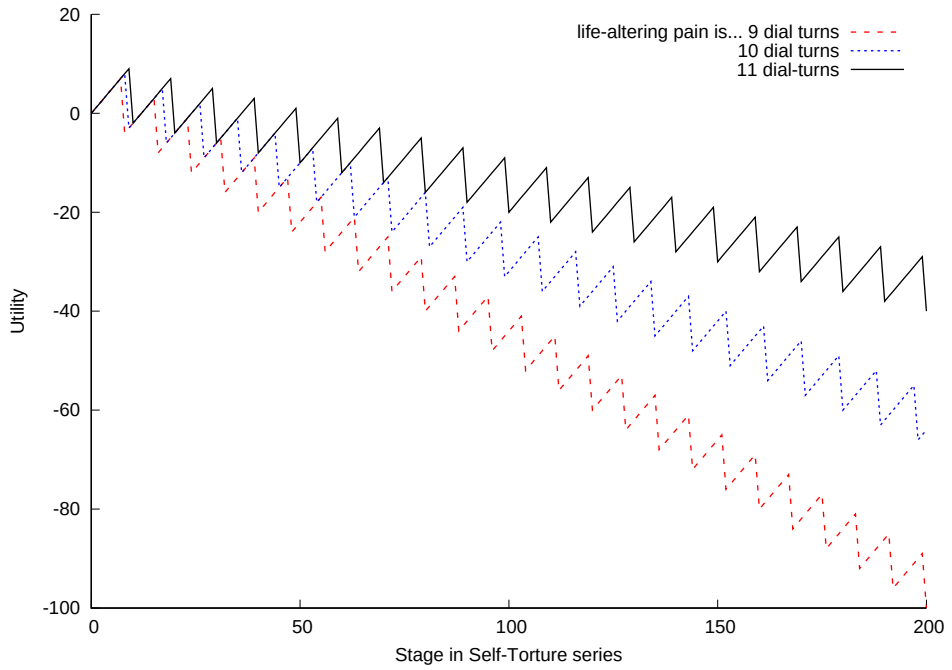

Figure 3: Self-Torturer's Utility Function

Teasing apart the disutility of dial-turns into two components like this is obviously an idealisation, but it allows us to show one way the repeating self-torture situation could arise. On this model, were it not for the risk of life-altering pain, turning the dial would be determinately required every time: ignoring life-altering pains, turning the dial is always preferred. (This is the picture behind

the revisionary intransitive-preferences accounts of ST's plight.)

The model could also be a simplification in two ways. As I've said, later in the series it's plausibly impermissible to turn the dial. Second, we might want taking the deal to be merely E-admissible even considering only the mundane aspects of pain; this could be done by making the linear disutility of each dial-turn indeterminately somewhere in the range [-11, -9].

But the goal here is not exactly psychological realism. It is to show how Compound can patch-up standard decision theory to explain plausible verdicts about ST's plight, such as Nonsegmentation. That this can be done even with a grossly *simplified* model is a bonus, both for tractability and for showing that the model captures the underlying problem without too many epicycles.

But why accept this model? Aren't we attributing too much psychological realism to utilities, especially given my professed attachment to the choice-based construal of preferences? No: nothing hangs on the utility functions being psychologically real. What matters is whether they represent the choices we think ST should make. In particular, to get something like the structure of the model, we don't need the full machinery of life-altering pain increases: we simply need places where ST prefers not to turn the dial and refuses to do so, along with those places being indeterminate.

Before moving on, I'll consider an illustrative objection. Isn't it obvious as an intuitive datum about the case that because every 1-trade is an offer of the same amount of money ($10,000) for the same pain increment (one dial-turn), the 1-trades must stand or fall together, permissibility-wise?[11] The short answer is No. Though it's the same pain increment every time (let's suppose), some pain increments cross significant thresholds, such as what I'm calling the *life-altering*.

But even without vagueness this inference would not follow. Suppose that setting 10 is the first setting in which I experience a constant lifelong ache, whereas at setting 9 it was intermittent. It may well be that moving from stage 8 to stage 9 is required, but from stage 9 to stage 10 is forbidden. At one level the pain increment is the same (one extra dial-turn) but at another the later increment induces a life-altering change: constant ache. Similarly, at one level the stone increment is the same (another stone on the pile) but at another a later increment induces a change (having a cairn).

In my model the second level (that of cairns and life-altering pains) has vagueness, and so it's vague when we cross those thresholds. And so in one sense the 1-trades do stand or fall together: they are all (outside any initial and final segments) E-admissible if any is. Hence in isolation, they are all permissible if any is, given Liberal. That's the truth behind Nonsegmentation-Permissive.

But at the level of a compound action, they do *not* stand or fall together: Compound says that two merely E-admissible 1-trades differ in permissibility when one is the completion of a foolish sequence and the other isn't. In self-torture, whether some trade is the completion of a foolish sequence usually depends on how far along in the series it is, because ST is offered later deals only if she takes

[11] I owe this objection to Teruji Thomas.

the earlier ones. The exception would be if ST is 'dropped' somewhere into the series, in which case the 1-trade is permissible. Sally's Bets were different, because she is offered Bet B no matter what she did in Bet A.

Teasing apart the senses in which 1-trades stand or fall together (or not) illustrates why the puzzle has been so obstinate: there is a sense in which they do, but there is another sense—specific to unsharpness, and captured by Compound—in which they do not. As with Sally's Bets, the latter sense becomes apparent only when a *sequence* of 1-trades is considered.

Relatedly, an anonymous reader objects that it seems implausible that we don't care all that much about the qualitative character of the pain, but do care a lot about whether the pain is 'life-altering' (which is vague). In response, on the current model we do care an awful lot about the character of the pain. In a 10-trade we gain 100 utils from the money, lose 90 utils from the 'mundane' pain, and lose 12 utils from the life-altering increase. These numbers are all somewhat arbitrary of course, but as you can see the life-alteration is a mere sprinkling of additional disutility on top of that from felt pain.

And I think this is psychologically realistic: pain harms us in many ways. As well as being nonlinear at a qualitative level, there will be step-changes: the pain level that stops me going for a run, that stops me relaxing when playing with my son, and so on. And it is quite plausible that there will be vagueness in these levels. A focus on the 'life-altering' also offers a model for the Self-Torturer with imperceptibility, because if adjacent settings really are indistinguishable then assuming any difference at all in utility between all adjacent settings is problematic. Nevertheless an increase in pain will become noticeable at some vague point—that is the nature of the phenomenal sorites, paradoxical as it is—which we can call the 'life-altering' point or use some other predicate, and attach disutility to. And that will happen again, repeating throughout the series.

Here's a quotidian cousin of the model. In language that might now be regarded as a little fatphobic, (Quinn 1990, 79) says that we want to eat but don't want to 'look fatter'. But notice that this also repeats: it's not as if we are like the Shepherd with just one cairn to build, where we either look fatter or we don't. Instead once we look fatter we can keep eating and look fatter again. A repeating model of this predicament attaches disutility to 'looking fatter', and if you look fatter after j bites, a j-trade is forbidden. But quite plausibly, j will be indeterminate (and of course not constant).

Once we see the Puzzle as a repeating puzzle of vagueness, its force as a challenge to standard decision theory largely evaporates. Given Supersharp, there is an optimum stage (near the start if we use my stipulated numbers) but it's vague what stage that is. Transitivity is a supertruth because as all sharpenings of 'is a cairn' induce a transitive utility function, and so do all sharpenings of ST's preferences, if she is rational. Indeterminacy creates the appearance of intransitivity. So standard decision theory—augmented to cope with vagueness, in the form of Supersharp—is preserved.

4 Filtered and Refined Series

Quinn's 'filtered series' break down ST's plight into a series of n-trades:

> Suppose that instead of moving one step at a time [...] 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.[12]

The opposite of filtration is *refinement*: for example, a series which contains only 10 stages (0, 100, 200, ..., 900, 1000) is a filtered version of one which contains 100 stages (0, 10, 20, ..., 990, 1000), and the latter is a refinement of the former. (Quinn 1990, 86) advises that we look for 'the most refined series in the set that clearly preserves transitivity of preference *and* contains a position better than 0', if one exists.

For him, filtered series are a heuristic, a way for ST to 'imaginatively restructure his problem'.[13] This is because after using such a series to identify a permissible stopping place n, it remains the case that stage n+1 would be preferred. The problem remains, and so more radical deontic surgery is needed to prevent him from advancing to stage n+1:

> On such a theory of rationality some contexts of choice fall under the authority of past decisions. [...] *An agent is not rationally permitted to change course even if doing so would better serve his preferences.*[14]

As we've seen, something like this departure from maximising will be needed for any account of the Puzzle as involving genuinely intransitive preferences: if I determinately prefer the next stage, then standard decision theory will tell me to advance to the next stage.

But the Bankrupt Shepherd shows us why filtered series work on a sorites model *without* such radical surgery. Even though the slope of each 1-trade is indeterminate, at longer ranges there's a determinate downward slope. Filtered series don't imaginatively restructure the problem, they focus on one aspect of it—the longer-range transitive preferences. By filtering we remove stages, which means we remove places for sharpenings to disagree, which means we remove indeterminacy in our choices.

Finally, here's an objection that my accommodation of Nonsegmentation—by making it indeterminate how adjacent stages compare—doesn't go far enough. Isn't it implausible that ST's preferences between adjacent stages are indeterminate? How could (in isolation) $10,000 for a barely-perceptible increase in pain *ever* fail to be preferable? This objection undergirds the intuitive appeal of intransitivity.

[12] Quinn (1990), pp. 85-6.
[13] Quinn (1990), p. 86.
[14] Quinn (1990), p. 87. Emphasis in original.

But I think it fails to visualise the comparison in sufficient detail, especially past the near part of the torture series. Consider what life at stage 700 (for example) is like. When offered $10,000 to advance to stage 701, ST already has a lot of money and a lot of pain. She might be in tears from the agony and the knowledge that it's permanent. It's not obvious that an extra $10,000—which she won't really be able to enjoy in any case—is worth making the pain even worse, albeit slightly so.

Iain M Banks offers a particularly vivid picture of such suffering in *Surface Detail*. A billion souls are trapped in a simulated Hell, and their lives are not worth living. But they can be saved by Chay, who can permanently kill one soul a day, at the cost of a small amount of extra pain. There are many volunteers, but more people in the Hell than she could ever feasibly 'rescue' in this way, which she takes herself to have moral reason to do:[15]

> With every death she took on a little more pain... an aching tooth here, a stabbing feeling in her gut there, a persistent headache, what felt like a trapped nerve in one hip ... a thousand almost infinitesimal little pangs and stings and sprains and strains and ulcers and chafings, either adding incrementallly to some established hurt or starting a fresh site. ... No single ache dominated, and even when taken together the sum of them was not utterly debilitating, but they all nagged, all had their effect, filling her days... (Banks 2011, 471–72)

Whilst here Chay is tempted to 'turn the dial' for moral reasons or compassion rather than money, her case shows that it's not obvious that the Self-Torturer should always turn the dial in later stages. A little extra pain could take Chay across some qualitative thresholds she strongly disprefers.

I am convinced that this objection (that 'the next stage is always clearly preferable') has been dealt with. But if you aren't convinced, the vagueness view another resource. It's 'clear' that adding 1mm to a short man gives you a short man, and yet denying this intutive judgement is the price we pay to escape the sorites. The correct theory of vagueness should explain why such claims are compelling but false, and extending that explanation to ST's case is preferable to (for example) vindicating intransitive preferences.

But isn't this all a bit overwrought? At the start of the self-torture series, couldn't we simply form a plan to stop at stage (say) 70, and a preference to stick to that plan? Quinn defends such resolute choice: he thought that the Self-Torturer must set an initial strategy and stick to it, to cope with intransitivity.[16]

I think planning is a good *implementation strategy* for dealing with the practical sorites we face, but doesn't address the underlying puzzle. That puzzle recurs at the stage of plan-formation: wouldn't it be better to plan to go just a little further, to move 11 instead of 10 stones? Moreover, we'd like a solution that also helps an agent who starts turning the dial without a specific plan (as in Sally's Surprise, she may not know that the situation calls for one). Intuitively, even if failing

[15] Banks (2011), pp. 394–7.

[16] See especially Quinn (1990), pp. 86–87.

to form a plan is sometimes irrational, it's not fatally so: rationality should still offer her some guidance, to avoid turning the dial all the way to the end.

Compound locates the irrationality of continuing to turn the dial not in violating a plan or a past decision, but in—together with your past *actions*—completing a determinately impermissible n-trade, and a plan can help one avoid doing that. The irrationality is located in actions, not in plans. In the next chapter I turn to environmental damage, where planning and implementation are separated, and the situation becomes more difficult as a result.

5 Conclusion

I have always thought that the Puzzle of the Self-Torturer simply *must* be some kind of sorites paradox. There are too many similarities between the puzzle and ordinary sorites for it not to be. The challenge has always been to model it as vagueness whilst reflecting all of its features. I believe that I've done that here, at least if I can help myself to Compound.

References

Andreou, Chrisoula. 2023. *Choosing Well: The Good, the Bad, and the Trivial*. 1st ed. New York: Oxford University Press. https://doi.org/10.1093/oso/9780197584132.001.0001.

Asgeirsson, Hrafn. 2019. "The Sorites Paradox in Practical Philosophy." In *The Sorites Paradox*, edited by Sergi Oms and Elia Zardini, 229–45. Classic Philosophical Arguments. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316683064.013.

Banks, Iain M. 2011. *Surface Detail*. New York: Orbit.

Elson, Luke. 2016. "Tenenbaum and Raffman on Vague Projects, the Self-Torturer, and the Sorites." *Ethics* 126 (2): 474–88. https://doi.org/10.1086/683533.

Quinn, Warren S. 1990. "The Puzzle of the Self-Torturer." *Philosophical Studies* 59 (1): 79–90. https://doi.org/10.1007/BF00368392.

Tenenbaum, Sergio. 2020. *Rational Powers in Action: Instrumental Rationality and Extended Agency*. First edition. Oxford: Oxford University Press.

Tenenbaum, Sergio, and Diana Raffman. 2012. "Vague Projects and the Puzzle of the Self-Torturer." *Ethics* 123 (1): 86–112. https://doi.org/10.1086/667836.

Tuck, Richard. 1979. "Is There a Free-Rider Problem and If so What Is It?" In *Rational Action*, edited by Ross Harrison, 147–56. Cambridge University Press.

———. 2008. *Free Riding*. Cambridge (Mass.): Harvard university press.