Rational choice and the transitivity of betterness

Toby Handfield

April 20, 2013

This is an author's pre-publication draft of a paper to be published in *Philosophy and Phenomenological Research*.

1 Introduction Here is an attractively simple picture about practical rationality, with two elements.

1. (*Judgment*) There is a relationship between states of affairs which we might call *all-things-considered-betterness*. Some states of affairs are better than others. A rational agent aims to form accurate judgments about the relative betterness ranking of available options.[1]

   So some states of affairs are better than others, in this highly inclusive sense of "better". Some states of affairs are equally good, and perhaps yet other states of affairs are unranked by the betterness relation. The betterness relation, we suppose, is transitive and asymmetric. Betterness, then, gives rise to at least a (strict) *partial ordering* over states of affairs.

2. (*Action*) Having formed judgments about which outcomes are better than others, at least some part of rational agency is concerned with acting so as to bring about relatively good outcomes. Many philosophers have thought that this concern is not exhaustive of practical reason or of morality in particular, but it would be an extreme and I daresay fanatical view which denied that seeking desirable consequences was of any importance whatsoever.[2]

   Other things being equal then, a rational agent seeks to bring about states of affairs that are judged *maximal*, given the betterness relation. That is, from a set of possible

1. Of course, in natural language, there are many different senses of the words "good", "better", and "best", but it is tempting to think that we can identify one all encompassing sense that is relevant for practical rationality. This is what is intended by 'all-things-considered-betterness'. Even if you are sceptical that there is such an all-inclusive sense of betterness (Thomson 2008: 25–6), it does seem plausible to think that there is a sense of betterness that is all-inclusive of what is relevant to prudent behaviour: "all things considered, better *for me*". The arguments of this paper can be taken to apply only to that narrower notion, if the reader prefers.
2. See, e.g., John Rawls: "All ethical doctrines worth our attention take consequences into account in judging rightness. One which did not would simply be irrational, crazy" (1971: 30).

states of affairs that the agent can bring about, she will aim to bring about the best. Or if there is no unique best outcome, she will seek to bring about an outcome that is at least *not worse* than any other outcome she could have achieved.

Larry Temkin and Stuart Rachels have each separately, in a number of publications (Rachels 1998, 2001, 2004; Temkin 1996, 1999, 2012), argued that we have serious reason to think that the above picture is false. In particular, Temkin and Rachels claim there is good reason to think that the betterness relation *does not constitute an ordering*, because the relation is not transitive. So there may be cases where although A is better than B, and B is better than C, it is not the case that A is better than C. This is a very surprising suggestion.

Even more surprisingly, Temkin and Rachels argue that the betterness relation may admit cycles, in that there exist states of affairs $\{S_1, \ldots, S_n\}$ such that $S_1 > S_2 > \ldots > S_{n-1} > S_n$, and yet *also* $S_n > S_1$.

The picture I sketched above has a lot of plausibility. The idea that the betterness relation is transitive and acyclic seems extremely plausible. The contrary idea that it admits cycles is regarded by many as crazy. Why do Temkin and Rachels think we should even consider such odd ideas? In his recent book, *Rethinking the Good* (2012), Temkin marshals a range of ingenious examples to illustrate inconsistency between various deeply held views about our practical ideals, both moral and prudential. His arguments appear to show that, if we wish to hold on to these ideals, we will have to abandon the transitivity of betterness. Temkin does not claim that we have decisive reason to reject transitivity as the best response to his arguments, but that it is an option worthy of very serious consideration.

In this paper, I develop an alternative response to a central kind of example that Temkin, in particular, uses to motivate his argument. (I take much of what I say to be relevant to Rachels's position also, but focus on Temkin's presentation as the most recent and most thoroughly developed specimen.) The response is to show that plausible hypotheses about how real agents might behave in situations resembling Temkin's example support alternative hypotheses about the structure of the betterness relation. In particular, I claim that we can imagine hypothetical agents whose behaviour does not look obviously irrational, is consistent with transitivity, but which will suggest non-trivial *incompleteness* in the betterness relation. Another way to put this is to say that the value ordering that is relevant for practical rationality is incomplete. Yet another way to put this is to say that there may be instances of incomparable or incommensurate value that are relevant for practical rationality. This alternative hypothesis – which is consistent with the sketch of practical rationality given above – at least deserves serious consideration as a better account of the nature of

value than Temkin's and Rachels's radical proposals.

2 Temkin's example Temkin presents his central example in the context of four "views" about the nature of betterness. For convenience, I will use four propositions that diverge somewhat from Temkin's original claims (see note), but are nonetheless very similar in plausibility.[3] By using these claims, I can avoid using examples that involve extremely far-fetched scenarios, which I take to be an important methodological advantage. My formulations also enable me to remain neutral on a controversial aspect of Temkin's views, which will be explained below. But if the reader thinks that my modification of Temkin's views renders them significantly less plausible, it is easy enough to see how to reconstruct the ensuing argument so as to target the original claims.

V1: For any unpleasant or negative experience, no matter what the intensity and duration of that experience, it would be better to have that experience than one that was only a little less intense but sufficiently longer in duration.

V2: There is, or could be, a spectrum of unpleasant or negative experiences ranging in intensity, for example, from extreme forms of torture to the mild discomfort of a mosquito bite, such that one could move from the harsh end of the spectrum to the mild end in a finite series of steps, where each step would involve the transformation from one negative experience to another that was only a little less intense than the previous one.

3. Temkin's initial formulation of the four views is as follows (p. 135):

(V1) For any unpleasant or "negative" experience, no matter what the intensity and duration of that experience, it would be better to have that experience than one that was only a "little" less intense but twice (or three or five times) as long.

(V2) There is, or could be, a spectrum of unpleasant or "negative" experiences ranging in intensity, for example, from extreme forms of torture to the mild discomfort of a mosquito bite, such that one could move from the harsh end of the spectrum to the mild end in a finite series of steps, where each step would involve the transformation from one negative experience to another that was only a "little" less intense than the previous one.

(V3) The mild discomfort of a mosquito bite would be better than two years of excruciating torture, no matter how long one lived and no matter how long the discomfort of a mosquito bite persisted.

(V4) "All-things-considered-better-than" is a transitive relation. So, for any three outcomes, A, B, and C, which involve unpleasant experiences of varying intensities and durations, if, all things considered, A is better than B, and B is better than C, then A is better than C.

V3: The mild discomfort of a mosquito bite, endured for 60 years, would be better than enduring one month of excruciating torture.

V4: "All-things-considered-better-than" is a transitive relation.

You may complain that the terms "little" and "sufficient" are *vague*, and it is therefore not clear that we should accept these claims. What we require is a spectrum of experiences such that, on a consistent reading of these vague terms, V1–V4 might all be true. So in that spirit, I suggest the following series of unpleasant experiences, such that each change in intensity is "little", and each change in duration is "sufficient", such that each experience on the list is better than the preceding experience. (If you do not like this example, I invite you to invent your own, for it does seem plausible that there will be *some* spectra and some precisifications of these terms, such that the claims V1 and V2 are true.)

A1. Excruciating torture (1 month)
A2. Vicious torture (2 months)
A3. Serious torture (3 months)
A4. Mild torture (6 months)
A5. Food poisoning plus a sprained ankle (1 year)
A6. Food poisoning plus a bee sting (18 months)
A7. Food poisoning (2 years)
A8. Bad flu (3 years)
A9. Migraine (6 years)
A10. Sprained ankle (10 years)
A11. Bee sting (20 years)
A12. Mosquito bite (60 years)

By V1, A1 is better than A2. Although A1 involves more intense unpleasantness, it is of sufficiently shorter duration that it is better to undergo A1 than to undergo A2. Similarly, A2 is better than A3; A3 is better than A4; and so on, for all adjacent pairs in the list A1–A12.

Given the betterness relations above, and given the transitivity of betterness (V4), it follows that A1 is better than A12.

But by V3, A12 is better than A1. Contradiction.

Temkin says that it is very difficult to give up any of V1, V2, or V3. He suggests that it might be most plausible to give up V4. If so, we can avoid the contradiction, because we can

coherently deny that A1 > A12. The conjunction: A1 > A2 > A3 > A4 ... > A12 is true. But if it is nonetheless false that A1 is better than A12, then it is consistent to say that A12 is better than A1. This is the essence of Temkin's argument against transitivity of betterness: it allows us consistently to endorse V1–V3.

Before discussing Temkin's view further below, it is worth reflecting a little longer on the plausibility of V1–V3.

3 Initial response to the example If one finds that one has a number of inconsistent beliefs, a sensible strategy is to ask which belief can most readily be given up. Which of our four views has the lowest credence, for instance? I will put aside V4 for consideration in the context of Temkin's specific arguments below. Regarding the three other views, I am probably most confident about V3, or something close enough to it.

Note that my version of V3 is in one way a good deal weaker than Temkin's original (see note 3). Temkin claims that the qualitative badness of excruciating torture is so bad that a sufficiently long episode of torture is worse than *any* duration of enduring a trivial discomfort, such as a mosquito bite. In effect, while V1 asserts that quality and quantity of painful experiences can always be traded off against each other, Temkin's formulation of V3 flatly denies this, by asserting that there are some qualitative differences that are so great that they lexically dominate quantitative differences. The lexical dominance claim is very strong, and accordingly has been the subject of independent criticism.[4] I doubt, moreover, that we are well equipped to weigh up the badness of – for instance – 10 million years of enduring a mosquito bite versus the badness of two years of torture. Our imaginative powers regarding large numbers and long durations are too limited to be reliable.[5] But I suggest that we ignore this dispute for present purposes. We can at least agree with Temkin that *many, many* years of enduring a mosquito bite is much better than enduring two years of excruciating torture. Or, to use the cases I offered, 60 years of enduring a mosquito bite is surely much better than 1 month of excruciating torture. So although I am sceptical of Temkin's original claim, I believe that some proposition sufficiently strong to generate the problem should be given high credence.

V2 is also very plausible, though I have slightly lower confidence in it than in V3, be-

4. E.g. Broome 2004: §2.2, Norcross 1997.

5. Though see various passages in Temkin 2012, especially chapter 8, for Temkin's response to concerns of this sort.

cause V2 implies a large number of phenomenological claims which could be empirically disconfirmed. Obviously, the phenomenal characters of a mosquito bite and of torture are very different. But it does seem plausible that there is a possible spectrum of experiences that gets you from one to the other, such that there is only a relatively small phenomenal difference, in terms of unpleasantness, between each adjacent pair.[6]

That leaves us with V1. Because this is a claim about *all* unpleasant experiences that differ only a little in intensity, it is hard to be confident that it will be unrestrictedly true. Indeed, Temkin grants that, strictly speaking, V1 may be false (p. 140). There may be "little" differences in intensity where the trade-off of increased pain for reduced duration does not favour the shorter but more intense experience. But he claims that *there will be some spectra*, which both satisfy V2 and are such that V1 is true of all the unpleasant experiences in the spectrum. That is all he requires to generate the paradox.

How confident should we be of this claim? It does sound plausible, but compared with V2 and V3, I think it has to be the least plausible. Suppose we tried to apply it to the spectrum I came up with above (A1–A12). V1 amounts to the claim that every item in that list is *better* than the following. And while that does not seem a crazy thing to think, it is a "big claim", compared to V3 and V2. (Of course, Temkin does not have to agree with my suggestion that V1 is true of *this* spectrum, but he must be committed to a similarly strong claim about *some* spectrum.) V3 is a claim about the relative betterness of only two things, and the contrast between them is very stark. V2 is a claim about the existence of phenomenological continua which seems plausible enough. But V1 is another evaluative claim, which entails a large conjunction of evaluations that applies to all adjacent pairs of a spectrum. It seems obvious to me that if there is a place to raise doubts, this is it.[7]

So, of V1–V3, I claim that V1 is the most dubious. Temkin, on the other hand, claims that we should seriously consider rejecting V4. How does V4 compare in plausibility to V1?

6. Warren Quinn discusses similar cases in Quinn 1990. Though Quinn's case has the additional difficulty that adjacent points in the spectrum are *indiscriminable*.

7. Notice that the quality of the unpleasant experiences on the spectrum A1–A12 varies significantly. Jakob Hohwy suggested to me that we might think that we could obtain a more reliable case in favour of V1 if we ensured a greater degree of homogeneity across the spectrum. But if we do this, the rewritten version of V3 will be much less plausible. If the experiences at both ends of the spectrum are qualitatively similar, then we will not be as confident that the intense but shorter experience is worse than the longer, less intense experience.

3.1 How can we evaluate structural claims about betterness? It is typical to assume that the betterness relation is asymmetric and transitive. These ideas are seemingly so basic to our understanding of the relation that it is hard to tell how one could obtain evidence that they were incorrect. Inspired by this phenomenon, one might try to argue that, as a matter of conceptual truth, or as a matter of logic, betterness must be transitive.[8] This argument, however, is open to the reply – which Temkin indeed makes – that "you may be right about traditional concepts of betterness, but I am inviting you to use a new concept, which better fits with our other ideals about practical rationality such as V1–V3".[9] I think that pursuing this line of argument further is unlikely to be fruitful.

Another possible response is to try to investigate the role of evaluative beliefs in our ordinary folk psychological theorising. Would a surprising claim about the structure of betterness have implications for how we should think about the behaviour of our friends, colleagues, and ourselves?

Recall the picture I described at the beginning of the paper. Practical rationality involves two elements: *judgment* and *action*. Judgment is coming to a view about the relative value of outcomes, and action involves seeking to bring about better outcomes. If we accept Temkin's proposal regarding the nature of value then we may have to revise some of our judgments, and this will have further implications for how we should act.

One problematic implication of nontransitivity of betterness is that it is compatible with cyclical judgments about betterness. Indeed, Temkin is claiming that for spectra that meet the requirements of V1–V3, betterness relations in fact do form a cycle (see Figure 1 for an illustration). Cyclical betterness judgments, however, may rationally require us to submit to a "money pump".[10] Suppose that you start out in A12. You are offered the opportunity to purchase, at a minimal cost, relief from A12 and to be put in A11 instead. This is an attractive offer, because you regard A11 to be better than A12. You accept. Now you are made a similar offer to exchange A11 for A10. This looks like a good deal also. And so on.

Eventually, you get to A1, having paid out lots of small quantities of money. You are now

8. See, e.g. Broome 2004: §2.1.
9. Temkin 2012: §§1.5, 14.4. Rachels canvasses the possibility of a similar reply, though he does not personally endorse it (Rachels 2001: 218–9).
10. The original money pump idea is presented in Davidson, McKinsey, and Suppes 1955.

    Johan Gustafsson has recently shown how a money-pump can be constructed, even for an agent whose preferences are intransitive but *acyclic* (Gustafsson 2010). It is easier to present the problem as it arises for agents with cyclic preferences, however, so I focus only on that case.

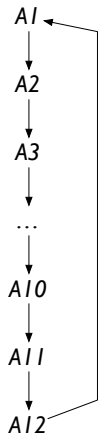

Figure 1: A graph of the alleged betterness relations between the states A1–A12. An arrow from x to y indicates x is better than y. Because Temkin denies transitivity, one cannot infer from arrows running from x to z through y that x is also better than z.

offered A12. Again, you think that A12 is better than A1, so you are rationally required to accept. You have returned to where you begun, while paying out money at every step. But at each step, you believed you were rationally required to accept the trade because you were offered a trade to a better state of affairs.

The money pump argument is offered as part of the traditional justification for why our preferences should be transitive. By giving up transitivity, Temkin seems to be inviting the conclusion that it is rational to be money-pumped in a situation like this.

Note the structure of the argument here:

P1. If betterness is nontransitive, then agents might be rationally required to accept a series of money pump trades.

P2. Accepting a series of money-pump trades is manifestly irrational.

C. Therefore betterness is transitive.

Temkin's response is to deny the first premiss. He claims that only dubious claims about rationality will support money-pumping in the face of nontransitive betterness cycles. In particular, consider the following rule of rational action.

Local maximization: If, at a given time, you face a number of options S, such that one option x in S is better than all other options in S, bring about x.

The local maximizer, because she considers only what is better out of the available options at a given time, will indeed be vulnerable to money pumping, as described above. So

if we accepted the local maximization rule as a complete account of rational action, then the objection would appear to succeed. Temkin quite rightly points out, however, that local maximization is not an entirely plausible theory of rational choice (§6.6). Local maximizers will never be able to make short-term sacrifices for the sake of long-term gains. They presumably will never have reason to quit smoking, to start exercising, or to study hard at school. This sort of behaviour is frequently symptomatic of irrationality, so local maximizing is not a good account of rationality.

So far, so good. But it is incumbent upon Temkin to provide some alternative account of rational choice, and to explain how, conjoined with evaluative judgments that violate transitivity, the agent will avoid manifestly irrational behaviour. This is not a trivial matter. Wlodek Rabinowicz (2000) has shown that even agents who decide using a form of foresight, by employing a backwards induction technique, will be vulnerable to money-pumps if they have cyclic preferences.

A rule of rational choice that might be used to avoid money-pumping is:

Global maximization: If, over an anticipated series of choices, you face a total set of options S, such that by an appropriate series of choices you can end up with any element in S, then make a series of choices which leads to a maximal item in S.[11]

The global maximizer, if able to anticipate the same series of choices as above, will say: "There is no maximal item in S, since for every item in S, some other item is better. So my rule gives me no guidance. All behaviour is equally rational or irrational." The global maximizer, then, might choose A1 over all alternatives – that is the global maximizer might choose the excruciating torture. Although this is choosing an item that is in some sense "less good" than what could have been chosen (e.g. A12), every other choice is "less good" than something else also.

(Note that global maximizing, although an attractive ideal, requires that an agent have abundant information regarding future choices in order to be effective. Real agents may want to behave like a global maximizer, but must do so with less than full information about future choice opportunities. It is for this reason that more short-sighted strategies

11. Temkin mentions these rules, but does not follow through exactly what global maximization appears to entail (§6.6). Temkin is following Elster's influential discussion (Elster 2000). See also Edward McClennen's proposal that an agent may employ resolute choice, which is very similar to the proposal above (McClennen 1990). Rabinowicz (2000), exploring these ideas further, argues that under certain conditions, resolute agents with cyclic preferences can avoid being money-pumped, but expresses doubts about the rationality of being resolute.

like local maximization retain some interest. This matter will be revisited below.)

In itself, global maximizing is a reasonably plausible account of rational choice. But to suppose that someone is behaving rationally by choosing at random from the spectrum A1–A12 is absurd. It also seems extremely far-fetched to suppose that any real individual – rational or otherwise – would have such an attitude. So the conjunction of (*i*) global maximization as an account of action and (*ii*) nontransitivity of betterness as a feature of rational judgment seems to have utterly implausible implications for practical rationality.

Here is what Temkin says to this end:

> [S]trictly speaking all the *money pump* shows is that in certain circumstances it would be irrational to both retain, and *repeatedly act on* one's nontransitive preferences. So, for example, it is assumed that part of the circumstances is that people have an overarching aim that their lives go well in ways that would be precluded if they allowed themselves to go broke by being repeatedly money pumped. If people didn't have such an overarching preference, or didn't allow themselves to *act* on their nontransitive preferences, it is hard to see how the mere possibility of being money pumped in virtue of their having nontransitive preferences would be enough to make people irrational. (Temkin 2012: 186)

Stuart Rachels engages with the money pump objection in similar fashion:

> Suppose we reject Transitivity in favour of the following thesis: for some possibilities, X is hedonically better than Y, Y is better than Z, but X is worse than Z. On a variant of the "money-pump" objection, an informed agent, who holds the thesis and is otherwise rational, would pay a small amount to trade X for Z (since Z is better), then pay a small amount to trade Z for Y (since Y is better), then pay a small amount to trade Y for X (since X is better) – the same X she started with. So, according to this objection, the thesis must be rejected. But the objection fails. The rational agent will not behave like this for exactly the reason why doing so seems irrational; because, from the standpoint of self-interest, one might as well put dollar bills down the garbage disposal. The money-pump objection assumes that a rational agent would always prefer what is better and act on those preferences, no matter what. But that assumption would be rejected along with Transitivity.
>
> (Rachels 2001: 218)

But neither Rachels nor Temkin seems – at least in the passages quoted – to grasp the

full force of the objection.[12] We have adopted a hypothesis about value relations that should make sense of an agent's behaviour, on the assumption that the agent is broadly rational.[13] On two particular hypotheses about rational behaviour, however – global maximization and local maximization – we get predicted behaviour that looks manifestly irrational.[14] Global maximizing looks like a relatively good hypothesis about rationality. If there is something wrong with it, then we should be given some sort of diagnosis of the error. But instead of that, Temkin and Rachels appeal to an ad hoc "overarching" preference not to go broke, or to an unexplained practice of "not acting" on nontransitive preferences. Apart from their inherent implausibility, these moves are inadequate. Consider again our global maximizer who is free to choose any of the outcomes A1–A12, without any associated cost. She says: "There is no best option, so I am indifferent", and chooses at random, ending up with A1: excruciating torture. This seems consistent with having a preference not to be money pumped and consistent with not repeatedly acting on nontransitive preferences, but it still seems irrational. The intuition that Temkin and Rachels rely upon to support V3 is that A1 – a reasonably long period of excruciating torture – is *a very bad outcome*. If that is so, surely a rational agent should take steps to avoid it.

What, then, is a more realistic response, and is there anything that can be said as to how rational that response is?

4 An alternative account If we wanted to know what an agent's beliefs were about the betterness ranking of the items A1–A12 and, in particular, whether they were committed to the transitivity of betterness, a number of different empirical approaches could be used.

12. Characterising Temkin's overall position on this matter is a delicate matter: while he does not explicitly affirm scepticism about practical reason, and attempts to carry out the argument of the book under non-sceptical assumptions, he recognises that his views may provide good reasons for such scepticism. See §14.8 for his most explicit discussion of this.
13. As Davidson writes (Davidson 1984: 153):

> Making sense of the utterances and behaviour of others, even their most aberrant behaviour, requires us to find a great deal of reason and truth in them. To see too much unreason on the part of others is simply to undermine our ability to understand what it is they are so unreasonable about.

14. Another possibility is that rational action involves satisficing, and again, satisficing will have local and global variations. I see no reason, however, to think that this will assist Temkin to explain why choosing at random from the spectrum is not a rational response for the global satisficer.

Putting aside the serious difficulties we would have in obtaining ethics committee approval, two approaches of interest are the following:

1. Binary choice experiments. The agent is endowed with one item from the spectrum, and offered the opportunity to trade it for an adjacent item. So for instance, the agent will be given A5 and then asked if she wishes to trade it for A4. In another experiment the agent will be given A4 and asked if she wishes to trade it for A5. And so on, for all adjacent pairs in the spectrum. We should also ask the agent to make a binary choice between A12 and A1.

   Ideally, the experiment should be arranged in some fashion to ensure that the agent (*i*) is not sure whether any given choice will be her last, leaving her stuck with the most recent choice, and also (*ii*) does not believe that future choices to be offered depend on present choices made. Given these restrictions, the agent will need to rely on relatively "local" evaluations of the two options on offer, rather than be able to use strategic considerations to bring about desirable choice scenarios in future, which are likely to be a confounding factor.

2. Free choice experiments. We tell the agent that she is going to be put in one of the options A1–A12, but that she can choose which. The agent then chooses freely.

   This sort of experiment, provided the agent believes that the choice is final, will allow the agent to adopt a global maximizing strategy over the whole spectrum, and thus gives information about the overall structure of her betterness judgments that complements the binary choice experiments.

In addition to these experimental paradigms, we could employ subtle variations such as binary choices between pairs that are not immediately adjacent; or free choices from subsets of the spectrum. No doubt other protocols will yield useful information also.

To motivate an alternative understanding of the value relations between the states on our spectrum, I ask the reader to indulge me in a rather detailed speculation about how an agent might respond in experiments of these sorts. It is my hope that, even if the imagined behaviour is not obviously rational, it will, after further analysis, appear to be a rational response to choice scenarios of this sort, yet preserve transitivity.

> **Story of Subject S** Our subject S is taken through a series of binary choice experiments. S frequently agrees that items from earlier in the spectrum (shorter duration, more intense experiences) are better than longer duration pains of lesser intensity. So A1 > A2, A2 > A3, etc. Moreover, S agrees that mosquito

bites are better than torture, so if given the choice between A12 and A1, he takes A12.

At two points in the series of experiments, however, S is reluctant to trade. For instance, when considering the trade from A9 to A8, he says: "Gee, a migraine for 6 years is pretty bad. But I do hate having the flu too. And 3 years of having the flu is still a very long time. I'm really not sure what to do here. I think I'll just stop."

S shows a similar reluctance to trade at a second point – suppose it is from A5 to A4.

Upon further experimentation, it becomes apparent that S's reluctance to trade from A9 to A8, for instance, does not reflect a straightforward *preference* for A9 over A8, because when S is endowed with A8 and is offered a trade to A9, he is again reluctant to trade. So for both points in the spectrum where S is reluctant to trade, the reluctance is symmetrical, and S prefers to retain the bundle with which he is currently endowed.

On further investigation, S's reluctance to trade is robust under perturbations that our experimenters call "mild sweetening".[15] For instance, the investigators say that, instead of trading from A9 to A8, S can trade from A9 to an enhanced version of A8, where he will be paid a bonus of $10. S is reluctant to take this trade also. (And similarly, he is reluctant to trade from A8 to a sweetened version of A9.)

Finally, in a free choice experiment, S reports having difficulty in choosing between A5 and A9. Eventually, S chooses A5, but reports that he simply "plumped" for that option on a relatively arbitrary basis. Having chosen, however, he is reluctant to trade A5 for A9, and again this is a reluctance that is insensitive to mild sweetening.

The behaviour of S is, at first inspection, somewhat mysterious. It is not obvious what rational strategy S is following. But nor is it obvious that S is behaving irrationally. Unlike an agent who chooses from the spectrum at random, S appears to have a definite strategy, and that strategy could be an attempt to act rationally, consistently with the account of rationality given at the beginning of the paper.

15. I take this terminology from Caspar Hare (2010).

I suggest that S's behaviour can be interpreted elegantly if we suppose that he has the following value commitments:

- A5 and A4 are *incommensurate* in value, as are A9 and A8.[16]

Incommensurability of value can be understood, in this context, as simply the denial of any other value relations. A and B are incommensurate if and only if A is not better than B, nor is B better than A, nor are they equal in value. Given incommensurability, the graph in Figure 2 is a plausible illustration of the value relations endorsed by S. Note that this interpretation is intended to include the assumption that betterness is transitive.

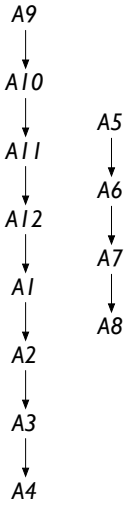

Figure 2: Posited value relations endorsed by subject S. In this graph, it is assumed that betterness is transitive, so additional arrows are implied between, for instance, A5 and A7.

The value relations posited in Figure 2 fit reasonably well with rational choice strategies such as global maximizing in free choice experiments and local maximizing in binary choice experiments. The local maximizer who is started out at A4 will accept trades to A3, to A2, etc., all the way to A9 if possible. Similarly, the local maximizer who commences at A8 will accept trades to each of A7, A6, and A5. The global maximizer, given a free choice, will clearly identify A9 and A5 as maximal choices. Neither is uniquely best, but they are the

16. There is some potential for terminological confusion on this point. Some say "roughly equal", some say "incommensurate" or "incomparable", some say "on a par", and some draw subtle distinctions between these. For my purposes, I will ignore these distinctions, and will say that two goods are incommensurate (or incomparable) if and only if it is not the case that one is better than the other, nor is it the case that they are equal in value.

    See Broome 2004; Chang 1997, 2002; Hsieh 2008; Rabinowicz 2008 for a variety of terminological proposals. See also Parfit 1985: §146.

only choices which *cannot be bettered*. This fits with the supposition that S finds only these two choiceworthy, but has trouble deciding between them.

One aspect of S's behaviour that is not entirely explained, however, is the symmetric reluctance to trade between, for instance, A5 and A4. A local maximizer, choosing between A5 and A4, should arguably be indifferent. Although these two options are not equal in value, nor is it the case that one is better than the other. So one might think that the local maximizing agent should behave similarly to an agent who held them to be equal in value.

Michael Mandler, a theoretical economist, has shown that, for agents with incomplete value functions, in a broad class of choice problems, a strategy of **status quo maintenance** (SQM) will guarantee that the agent avoids sub-optimal outcomes (Mandler 2005). SQM is a strategy which says: when choosing between something you are currently endowed with and an exchange for something else, only exchange if the alternative item is strictly better, otherwise retain your current endowment.

SQM is a heuristic that is widely observed in human decision making and traditionally has been explained by invoking a temporal shift in preferences, such that agents come to prefer retaining what they currently hold more strongly than they originally preferred to obtain that bundle of goods. Mandler argues that the preference shift hypothesis is empirically unfalsifiable, since it is very difficult to assess an agent's preferences for an option without at some point changing the agent's endowments over the course of an experiment. By supposing that agents have incomplete preferences, however, we can retain a stable structure of preferences and explain the behaviour in a plausible fashion (Mandler 2004).

Crudely, the rationale for this heuristic is that it makes the agent less vulnerable to money pumps than she would otherwise be, if she were relatively willing to trade between incomparable pairs. For instance, suppose an agent endorsed the valuations in Figure 2, and were endowed with A5. She is then offered a trade to A9, which she accepts, on the basis that she is willing to trade between incomparable goods. She is then offered another trade, to A6. She accepts this also, and finds that she has ended up with an option that is strictly inferior to one she could have had (A5). Now, it is true that an agent who is an effective global maximizer might be able to anticipate this outcome and intervene in her willingness to trade. But not every agent will be in a position to anticipate this sequence of trades, nor to notice it once it is underway. Effectively, SQM gives an agent who has only local information some of the benefits that the agent would have, if she were able to engage

in global maximization.[17]

SQM is an independently motivated heuristic, then, which allows us to explain subject S's reluctance to trade between pairs that are believed to be incomparable, such as A4 and A5.

So the emerging picture of Subject S is as follows:

- S endorses the truth of V3 and V2.
- S denies V1, because he holds that there are at least two pairs in the spectrum which are incomparable.
- S endorses the transitivity of betterness (V4).
- S uses a strategy of SQM for choice situations involving incomparable options. This leads to a reluctance to trade that looks superficially arbitrary, but which can effectively protect S from a variety of potential money-pumps.

(Of course, the particular value judgments that I ascribed to S, for the purposes of the example, are only one way to satisfy the above schema. Another response would be for an agent to manifest reluctance to trade across many more pairs in the spectrum. The extreme case of this disposition would be an agent who judged A12 better than A1, but all other pairs from the spectrum to be incomparable. Such an agent would reliably pick anything other than A1 in a free choice experiment, but would be reluctant to accept any trades, expect to trade A12 in preference to A1.)

I cannot prove that S's psychology is better attuned to practical reason than Temkin's proposal to endorse V1–V3 and reject transitivity. Nor do I claim that S's psychology is the

17. Erik Carlson proposes a rule of rational decision for agents with cyclic preferences which appears to have some structural similarities to SQM. Carlson's rule, which is specifically intended for cases like Quinn's self-torturer, but which is intended to govern global choices, rather than binary choices, is

> Choose the latest setting which is preferred to each earlier setting. (Carlson 1996: 149)

Here 'setting' refers to a setting on a device which delivers electric shocks of increasing intensity, ranging from no shock (setting 0) to an excruciating pain (setting 1000). He concedes that this rule introduces a seemingly arbitrary element in this decision rule: 'Which option it selects depends on our choice of a "starting point"' (155). He goes on to say, however, that 'the choice of 0 as a starting point can be justified. It is intuitively clear, in the case of the self-torturer, that a reasonable decision rule should select only settings preferred to 0. Apart from 1000, which we know is dispreferred to 0, this intuitive constraint does not hold for any other setting... A decision rule should therefore select a setting which is, loosely speaking, as much better than 0 as possible' (ibid). I confess to finding this passage unclear. I cannot see any reason to ascribe to 0 this special status, except for the fact that it constitutes the status quo. So Carlson's rule either presupposes the reasonableness of SQM, or appeals to some other consideration which is opaque to me.

only conceivable rational response to these sorts of experiments. There may well be other possibilities, and they may not need to rely on incommensurability judgments. My claim is just that the proposed interpretation of S's experimental responses involves a much less radical departure from orthodoxy than Temkin's proposal. In particular, we can explain, on the account given of S, why it would be a disastrous policy to choose from the spectrum at random.

Presumably, if one doubts the rationality of S's response, a crucial point of contention will be whether or not it is plausible that, in any spectrum of the sort described in V2, there will be at least two points where incomparability arises.[18] Temkin may argue that, if we make a sufficiently fine-grained spectrum – much more fine-grained than the simple example I have used – denying V1 will seem extremely implausible. Each and every step in the spectrum will involve a very small change in intensity, and a dramatic change in duration. Surely we should be able to confidently say that badness of a longer duration outweighs the goodness of a lesser intensity? And while we can agree with Temkin that this will seem a very plausible thought for each and every step in the spectrum, we can simply insist that our commitment to V2, V3, and V4 entail that we must find fault with V1 at some point, even if we are not confident which are the correct points at which to baulk.

It is worth noting that, strictly speaking, Temkin never explicitly shows us where in the pain spectrum there is a counterexample to transitivity; viz. three states, A, B, C such that A > B, B > C, yet it is not the case that A > C. Rather, his argument merely shows that V1, V2, and V3 imply that there must be some such trio, and the pain spectrum is intended to illustrate the plausibility of those three premises. But as I mentioned above: V1 is inherently less plausible than V2 and V3; and the larger and more subtle the spectrum to which V1 is intended to apply, the greater the threat to its credibility. We do better, I suggest, to doubt V1 and retain V4. Just as Temkin cannot tell us where the failure of transitivity occurs, I

18. Why must the agent think that there are two points of incomparability? Couldn't an agent be reluctant to trade across only one pair in the spectrum? Of course, this is possible, and it is enough to ensure that the betterness relation is acyclic. But for there to be non-trivial incomparability in a group of objects, at least three objects are needed, such that at least one pair stands in a betterness relation, and at least two pairs do not (Mandler 2004: 268). So for instance, if the agent merely refused to trade between A4 and A5, but was willing to trade for all other pairs, then it would seem that the agent was committed to A5 > A6 > … > A3 > A4. If we wished to endorse transitivity, then, the agent should accept that A5 > A4, and the refusal to trade is seemingly irrational.

cannot confidently say where the incomparability occurs, but to admit incomparability is less damaging than admitting non-transitivity.

A second point of contention will be whether it is possible to motivate this manner of response to the entire range of cases where Temkin shows we are liable to run into similar difficulties. Temkin claims that we have similar reasons to doubt transitivity in cases quite unlike the pain spectrum – especially in cases involving population ethics. Although it is reasonably easy to see how to generalize the strategy I have used in this paper, it is not clear that it will be similarly plausible in all cases. I will have to leave discussion of those cases for another occasion.

5 Vagueness versus incommensurability The account I have offered above is similar in some ways to recent proposals by Christopher Knapp (2007) and Mozaffar Qizilbash (2005), both of whom claim that transitivity may be saved by allowing that the better than relation is affected by *vagueness*. This leads in turn to a sort of incompleteness of value, because for some pairs of states of affairs, it is not *definitely* true that one is better than the other, nor is it *definitely* true that they are equal in value. So as far as what is definite goes, there is an incompleteness in the betterness ranking. Below I briefly describe the idea behind these proposals, before going on to explain why the proposal developed above is at the very least a useful alternative response – and quite likely a more powerful response – to Temkin's case for non-transitivity.

Both Knapp and Qizilbash suppose that although V2 and V3 are true, V1 is false. The thought is that V1 is true when comparing qualitatively similar pains, but where a significant qualitative barrier is crossed, such as the barrier between *awful* pain and non-awful pain, it is no longer true that trade-offs of intensity for duration will satisfy V1. It is the crossing of some such qualitative barrier that is supposed to explain why V3, as Temkin originally formulates it, is true. Recall that Temkin's original argument uses the claim that some pains of sufficient intensity – such as a period of excruciating torture – are worse than *any* duration of a mild pain – such as a mosquito bite.

The burden, then, is to explain why V1 seems so plausible, if it is supposed to be false. The reason we have trouble noticing the falsity of V1, according to this response, is that the relevant qualitative predicates, such as 'awful', are vague. There are pains in the middle of the spectrum that are borderline cases. For such cases, it is not merely difficult but impossible to settle definitely whether they are awful or not.

So suppose that, for the sake of argument, A7 and A8 are both in the borderline region.

Can we definitely say that A7 is better than A8, on grounds that A7 is only a little more intense, but that A8 is of much greater duration? According to Knapp, we cannot, because this implies that it is either definitely true that both A7 and A8 are awful, or it is definitely true that A7 and A8 are not awful.[19] But neither of these is the case, because it is not definite that A7 is awful and it is not definite that A7 is non-awful – that is just what it means to be a borderline case. Exactly the same can be said of A8.

(Personally, I am rather doubtful of this move. It seems entirely possible that we could have two very similar items, both not definitely awful, but one of them is definitely worse than the other. Temkin makes a similar complaint (2012: 536). But I won't press this objection here.)

The vagueness of the relevant qualitative barrier, then, means that not all of the entailments of V1 are definitely true. Although we cannot say of any pair in the spectrum that it is definitely a place where V1 is false, when we enter the borderline region, V1's entailments will be neither definitely true nor definitely false. So in effect, the ranking that corresponds to *definitely true* betterness ascriptions contains gaps.

Provided that the borderline region includes at least two gaps in the ranking, this will have a similar result as incommensurate value. Indeed, it can be seen as a special case of incommensurability: it is incompleteness in what is *definitely better*. Being a special case, this response has more limited dialectical power than my own response, for the following reasons.

First, as I have already noted, Temkin's version of V3, with its implicit commitment to the idea that some qualitative differences between pains lexically dominate quantitative differences, is highly controversial. Consequently, it is desirable that a response to Temkin and Rachels does not *require* that the lexical inferiority claim is true, because the threat to transitivity does not depend on this claim – as shown by the reformulated version of the argument discussed above. But both Qizilbash and Knapp rely upon the lexicality claim in presenting their responses. It might be possible to reformulate their response to avoid this feature, but if so, I suspect their responses will look yet more similar to my own.

Second, we might complain that we know very little about how vagueness in the betterness relation interacts with decision rules. I have offered an account, drawing on Mandler's work, of how incompleteness may justify the SQM choice strategy. But it is not clear that, on all accounts of vagueness, this will be a justified response to incompleteness in the rank-

19. Knapp 2007: 14–15. Qizilbash's analysis is essentially similar, see Qizilbash 2005: 124.

ing of what is *definitely* better. For instance, on an epistemicist reading of vagueness, it may be the case that there is a unique maximal element in the spectrum, but that it is impossible to know what this is. In that case, the appropriate choice strategy might be better understood as involving choice under *uncertainty* as to whether a given item in the spectrum is best. It remains open that this could require quite different choice strategies, whose apparent rationality would require further defence.

In sum, while it is indeed plausible that betterness is a vague concept, it is not clear that vagueness per se does the explanatory work in justifying the retention of transitivity. Rather, it is the appeal to incompleteness in the betterness ranking and consideration of desirable behavioural strategies in light of that, which does most of the explanatory work.

6 Conclusion In conclusion, if we are to take the proposal of nontransitivity seriously, its defenders need to provide us with an account of rational choice that is explicated with a similar degree of rigour as SQM, and which will explain why rational agents would not be required to behave in ways we take to be manifestly irrational. My alternative proposal appears to deliver at least that much. Moreover, my proposal shows that the appeal to vagueness in the betterness relation is merely a special case of a more general strategy, which need not be committed to any controversial claims about the existence of lexical priorities of value. Consequently, I suggest that my proposal is a preferable hypothesis.[20]

REFERENCES

Broome, John. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Carlson, Erik. 1996. "Cyclic Preferences and Rational Choice". *Theoria* 62: 144–60.

Chang, Ruth. 1997. "Introduction". In *Incommensurability, Incomparability, and Practical Reason*, edited by Ruth Chang. Cambridge, Ma.: Harvard University Press, 1–34.

———. 2002. "The Possibility of Parity". *Ethics* 112: 659–88.

Davidson, D., J. McKinsey, and P. Suppes. 1955. "Outlines of a formal theory of value, I". *Philosophy of Science* 22: 140–60.

Davidson, Donald. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

20. Thanks to an anonymous referee, Adam Bales, Stephen Barker, Daniel Cohen, Rohan French, Lloyd Humberstone, Larry Temkin, and Alastair Wilson for comments on earlier drafts of this paper.

Elster, Jon. 2000. *Ulysses Unbound: Studies in rationality, precommitment, and constraints*. Cambridge: Cambridge University Press.

Gustafsson, Johan E. 2010. "A Money-Pump for Acyclic Intransitive Preferences". *Dialectica* 64: 251–7.

Hare, Caspar. 2010. "Take the sugar". *Analysis* 70: 237–47.

Hsieh, Nien-hê. 2008. "Incommensurable Values". In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, fall 2008 edn.

Knapp, Christopher. 2007. "Trading quality for quantity". *Journal of Philosophical Research* 32: 211–34.

Mandler, Michael. 2004. "Status Quo Maintenance Reconsidered: Changing or incomplete preferences?" *The Economic Journal* 114: F518–F535.

———. 2005. "Incomplete Preferences and Rational Intransitivity of Choice". *Games and Economic Behaviour* 50: 255–77.

McClennen, Edward F. 1990. *Rationality and dynamic choice: Foundational explorations*. Cambridge: Cambridge University Press.

Norcross, Alastair. 1997. "Comparing Harms: Headaches and Human Lives". *Philosophy and Public Affairs* 26: 135–67.

Parfit, Derek. 1985. *Reasons and Persons*. Paperback edn. Oxford: Oxford University Press.

Qizilbash, Mozaffar. 2005. "Transitivity and Vagueness". *Economics and Philosophy* 21: 109–31.

Quinn, Warren S. 1990. "The puzzle of the self-torturer". *Philosophical Studies* 59: 79–90.

Rabinowicz, Wlodek. 2000. "Money Pump with Foresight". In *Imperceptible Harms and Benefits*, edited by Michael J. Almeida. Dordrecht: Kluwer, 123–54.

———. 2008. "Value Relations". *Theoria* 74: 18–49.

Rachels, Stuart. 1998. "Counterexamples to the Transitivity of Better Than". *Australasian Journal of Philosophy* 76: 71–83.

———. 2001. "A set of solutions to Parfit's problems". *Noûs* 35: 214–38.

———. 2004. "Repugnance or Intransitivity: A Repugnant but Forced Choice". In *The Repugnant Conclusion: Essays on Population Ethics*, edited by Jesper Ryberg and Torbjörn Tännsjö. Dordrecht: Kluwer, 163–86.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Ma.: Harvard University Press.

Temkin, Larry S. 1996. "A Continuum Argument for Intransitivity". *Philosophy and Public Affairs* 25: 175–8211.

———. 1999. "Intransitivity and the Person-Affecting Principle". *Philosophy and Phenomenological Research* 59: 777–784.

———. 2012. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford: Oxford University Press.

Thomson, Judith Jarvis. 2008. *Normativity*. Chicago: Open Court.