A Defense of Transitivity

by

Carson Calloway

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved April 2015 by the
Graduate Supervisory Committee:

Brad Armendt, Chair
Doug Portmore
Nestor A. Pinillos

ARIZONA STATE UNIVERSITY

May 2015
ABSTRACT

This thesis seeks to defend transitivity as a rational constraint on preferences against two putative counterexamples to transitivity. This thesis is divided into three sections. In the first section, I consider two famous and popular arguments in defense of transitivity and argue they are insufficient to adequately defend transitivity. I then outline a desiderata for successful arguments in defense of transitivity and identify some basic assumptions I will be making throughout the thesis. In section two, I consider the first putative counterexample to transitivity: Quinn’s Puzzle of the Self-Torturer. I offer two plausible interpretations of Quinn’s puzzle and argue that both fail. One because it does not genuinely induce intransitive preferences, and the other because the situation it requires is logically impossible. I conclude this section by defending my arguments against known objections in the literature. Finally, in the third section, I consider a counterexample to transitivity from Larry Temkin that has received little attention in the literature. I argue that while the initial counterexample is unpersuasive it can be augmented and made into a more forceful argument. I then argue that this improved counterexample fails due to some erroneous assumptions prevalent in the literature on incomparability. I conclude the thesis with a brief summary and some closing remarks.
TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................ iii

SECTION ONE ..............................................................................................................1

The Semantic Argument ............................................................................................2
The Money-Pump ..........................................................................................................4
Objection One ...........................................................................................................10
Objection Two ..........................................................................................................12
The Right Kind of Reasons ......................................................................................18

SECTION TWO ...........................................................................................................22

Quinn and the Puzzle of the Self-Torturer ................................................................23

Two Alternative Interpretations ...............................................................................25
The First Interpretation ..............................................................................................27
The Second Interpretation and the Phenomenal Sorites .........................................32
Fara’s Argument for the Transitivity of ‘Feels the Same As’ ..................................34
Objections to Fara ....................................................................................................42

SECTION THREE .....................................................................................................49

Temkin’s Counterexample .......................................................................................50
Temkin’s Counterexample Improved .......................................................................55
Improved Counterexample Rejected .......................................................................57

CONCLUSION ..........................................................................................................62

REFERENCES ..........................................................................................................60
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>60</td>
</tr>
</tbody>
</table>

SECTION ONE

A thesis which set out to argue that intransitive preferences are irrational would be boring at best if there are already well-known, cogent arguments in defense of the same conclusion. Thus, if my thesis is to be at all interesting I must provide reasons for thinking that the standard treatment of the rationality of intransitive preferences is insufficient. I think there are such reasons. The purpose of this section is to present two standard arguments for the irrationality of intransitive preference sets—the semantic argument and the famous money pump argument—and argue these treatments are in some form or another, deficient. The consequence of this discussion will be that it is an open question, so to speak, as to whether or not transitivity is a constraint on rational preferences, and further, if we wish to maintain that intransitive preferences are irrational in the face of arguments to the contrary, we are in need of different arguments. I will begin this section with the semantic argument as it has been formulated by Donald Davidson and John Broome, and offer my reasons for rejecting this approach. I will then consider the famous money-pump and offer two competing interpretations of this argument. I will argue that neither of these interpretations are sufficient to show that intransitive preferences are necessarily irrational. The shortcomings of these arguments for intransitivity will illuminate a desiderata for any argument against the rationality of intransitive preference orderings and provide a context for remainder of the thesis. Throughout this section I will also identify certain assumptions that I will be taking as basic, taking care to note when I am making a substantial departure from the traditional debate concerning preferences.
**The Semantic Argument**

The semantic argument for the transitivity of strict preference was first formulated by Donald Davidson, who argued:

> The theory in each case is so powerful and simple and so constitutive of concepts assumed by further satisfactory theory (physical or linguistic) that we must strain to fit our findings, or our interpretations to preserve the theory. If length is not transitive, what does it mean to measure length at all? We could find or invent an answer but unless or until we do we must strive to interpret ‘longer than’ so that it comes out transitive. Similarly for ‘preferred to’.  

This line of reasoning suggests that transitivity is just part and parcel of what it means to have preferences, and this is supposed to be analogous in some way with the measurement with length. It seems plausible that transitivity might be a “property embedded in the meaning of the relation” ‘longer than’; however it is unclear as to why ‘longer than’ is analogous to ‘preferred to’ in the relevant sense.  

Davidson did not expound on what he meant, but Jon Broome has given a similar argument in defense of the transitivity of ‘better than’ and perhaps this is what Davidson has in mind:

> Take any monadic predicate such as ‘dangerous’ [and] designate it with the letter ‘F’. We can often form from F a dyadic predicate, or relation, designated by ‘more F than’.... Call this the ‘comparative relation’ of F.... A comparative relation is necessarily transitive. This is an analytic feature of the operator ‘more … than’: the meaning of ‘more … than’ implies that ‘more F than’ is transitive.

---

1 Davidson, Donald (1980). *Essays on Actions and Events*. Oxford University Press. 273

2 John Broome has offered a similar argument for the transitivity of “all things considered better than” in Broome, John (2004). *Weighing Lives*. Oxford University Press.

Broome concludes that since ‘better than’ just means ‘more good than’, better than is a transitive relation. On the assumption that ‘preferred to’ is synonymous with ‘more preferable than’ Broome’s argument will generalize to the transitivity of ‘preferred to’.

If this is Davidson’s argument, I do not find it convincing. At least one relevant difference between length and preference is that while I cannot even imagine a collection of strings such that string A is longer than string B, string B is longer than string C, and String C is longer than string A; I can imagine a person whose preferences are intransitive and often people at least behave as if their preferences are intransitive notably in cases of procrastination.

What should we make of these people who exhibit intransitive preferences and even claim to have intransitive preferences? While I take it to be an open question as to whether or not they are rational, it does not seem at all plausible that they are misusing the word ‘preference’, and even granting that they are misusing the word, the interesting philosophical question is what we should make of their preferential attitudes whatever we decide to call them. I side with Larry Temkin that the semantic argument for the transitivity of better than is “at best a hollow victory.” “The crucial question is whether a

---


5 Chrisoula Andreou argues convincingly that “intransitivity, although it has been neglected in the literature on procrastination, may be, in one form or another, at the heart of all cases of procrastination.” Andreou, Chrisoula (2007). Understanding procrastination. *Journal for the Theory of Social Behaviour* 37 (2):183–193. p. 184
central element of moral and practical reasoning involves an intransitive relation … it matters not what we call this element."67

I have just concluded what little I have to say about the semantic argument for transitivity, I will now consider what I take to be the more forceful argument for transitivity: the money-pump argument.

**The Money Pump**

The classic money pump involves Mr. S, a job candidate, who has been offered his choice of three positions:8

a) Full Professor at a salary of 5,000

b) Associate professor at a salary of 5,500

c) Assistant professor at a salary of 6,000

Weighing these options, Mr. S prefers option a to option b, because he feels the better rank offsets the small difference in salary. Mr. S prefers option b to option c for the same reason; however, because the difference in salary between option c and option a is sufficiently large so as to outweigh the difference in rank, Mr. S prefers option c to option a. Mr. S is left with the following intransitive preference ordering: \( a > b > c > a \).

Now imagine Mr. S is offered a series of trades:

---


7 It bares noting that Temkin and Broome are discussing the transitivity of 'better than' and hence are not directly discussing the transitivity of 'preferred to'. However, I take these issues to be closely related and Temkin's comments are relevant for the debate on preferences as well.

The department head, advised of Mr. S's preferences, says, 'I see you prefer b to c, so I will let you have the associate professorship—for a small consideration. The difference must be worth something to you.' Mr. S. agrees to slip the department head $25 to get the preferred alternative. Now the department head says, 'Since you prefer a to b, I'm prepared—if you will pay me a little for my trouble—to let you have the full professorship.' Mr. S. hands over another $25 and starts to walk away, well satisfied, we may suppose. 'Hold on,' says the department head, 'I just realized you'd rather have c than a. And I can arrange that—provided . . .'

Mr. S has been turned into a money pump: as we pump water from a water pump, we pump money from Mr. S; indeed, the cynical department head could continue to offer Mr. S trades until Mr. S has been financially ruined and all because Mr. S’s preferences are intransitive.  

I am to infer on the basis that Mr. S’s preferences have led him to be a money pump that he is irrational, but is not clear how this inference is supposed to go. Although the money pump argument has been the standard argument in defense of transitivity since it was introduced by Davidson et al. what exactly the argument is supposed to be has not been universally agreed upon. There seem to be two schools of thought prevalent in the

---


10 Technically this version of the argument only applies to agents with cyclical preferences which are a subset of intransitive preferences. Suppose an agent’s preferences are such that \( a \succ b \succ c \succ a \). This agent’s preferences are intransitive: \( a \) is preferred to \( b \) and \( b \) is preferred to \( c \) but \( a \) is not preferred to \( c \), here the standard money pump won’t work because the agent need not trade \( a \) for \( c \) to satisfy her preferences. Rabinowicz argues that in the case I have described the money pump can be salvaged by offering an incentive, small enough that the agent will suffer and overall loss but large enough that agent now prefers \( c \) to \( a \), but it is not clear that this solution will generalize satisfactorily to other cases of intransitivity like instances where \( a \succ b \succ c \) and \( c \) is on a par with \( a \), \( c \) is incomparable with \( a \). See footnote 7 in Rabinowicz, W.: 2008, “Pragmatic Arguments for Rationality Constraints”, in: M.C. Galavotti, R. Scazzieri, and P. Suppes (eds.) Reasoning, Rationality, and Probability, Stanford: CSLI Publications, pp. 139-163 for the discussion of money pumps for intransitive acyclical preferences and Chang, Ruth (2002). The possibility of parity. Ethics 112 (4):659-688 for the discussion of parity.
literature. One school sees the argument as a logical continuation of Frank Ramsey’s Dutch-Book argument.\(^{11}\) This understanding of the Money Pump argument goes something like this:

1. No rational set of preferences would lead an agent to take a series of trades that leaves her having lost money and better off in no respect.
2. If an agents preferences are intransitive, then that agent could be made to take a series of trades which leave them worse off in one respect and better off in no other.
3. So, intransitive preferences are irrational.

This appears to be the understanding Frederick Schick has in mind when he argued that an agent could avoid being pumped if only he planned accordingly:

Does a person with cyclical preferences have no grounds for declining offers? Let him look back and see the arrangements he has already paid for. He may then come to see which way the wind is blowing, that if he accepts the current offer, he will then get another, and then another, and still another, every cycle bringing him back to where he was at the start, only poorer. Seeing what is in store for him, he may well reject the offer and thus stop the pump.\(^ {12}\)

Schick’s argument that an agent with cyclical preferences need not be exploited is only relevant on the assumption that the exploitability of the agent is what makes the intransitive irrational. In a similar vein, Edward McClennen argues that the agent could avoid being exploited if only the agent chose resolutely.\(^ {13}\) Wlodek Rabinowicz and Tom Dougherty have argued that Schick and McClennen’s arguments are not successful, yet

\(^{11}\) In addition to the philosophers here this appears to be the interpretation of the money-pump Hedden, Brian (2015). Options and Diachronic Tragedy. *Philosophy and Phenomenological Research* 90 (2):423-451 though Hedden argues that money-pump arguments are not successful, he agrees that this is how they should be understood.


they seem to agree on this framing of the money-pump argument. Rabinowicz and Dougherty agree that were Schick correct that an agent with cyclical preferences could avoid being turned into a money-pump by anticipating future offers and refusing to trades, the money-pump argument would fail to show that intransitive preferences are irrational; however, they argue that an otherwise rational agent with cyclical preferences would be forced to accept the trades either because backwards induction would compel an agent with cyclical preferences to take the trades,14 or because dominance reasoning compels the agent to take the trades regardless of how they anticipate acting in the future.15

The important feature of this understanding of the argument is that it is because the agent has preferences that leave her exploitable that she is irrational.

Against this interpretation, the second school understands the money pump argument as showing that the agent’s preferences violate the non-dominated choice principle: “A rational choice is one which selects an alternative to which none is preferred.”16 Understood this way the money pump argument looks like this:

1. If an agent’s preferences are such that the agent cannot make a choice without violating the non-dominated choice principle then those preferences are irrational.
2. If an agent’s preferences are such that they could be made to take a series of trades which leave them having lost money and better off in no respect, then that agent’s preferences are such that the agent cannot make a choice without violating the non-dominated choice principle.

---


16 This particular formulation comes from Gustafsson, Johan E. (2013). The Irrelevance of the Diachronic Money-Pump Argument for Acyclicity. Journal of Philosophy 110 (8):460–464. But it can be found in more or less the same terms in Davidson et al. “Outlines of a formal theory of value, I” 145.
3. The agent $S$’s preferences are such that they could be made to take a series of trades which leave them having lost money and better off in no respect.
4. Agent $S$’s preferences are irrational.

Here, being vulnerable to a money pump does not itself make a preference ordering irrational, but is a symptom of a violation of the non-dominated choice principle, and it is the violation of this principle that makes the agent’s preferences irrational. This understanding of the money pump has recently been defended by Johan Gustafsson who suggests this is closer the original interpretation of the significance of the money pump argument given by Davidson et al.\textsuperscript{17} Gustafsson argues that if we follow this understanding of the money pump, many of the standard objections to the money pump can be obviated by a synchronic version of the argument.

As with the traditional diachronic money pump, Johan Gustafsson’s synchronic money pump begins with an agent who has intransitive preferences, but in the synchronic version the agent is not offered any trades. Instead, the agent is given their choice of the options. In the case of Mr. S, imagine the department chair is less cynical and offers Mr. S whichever position he should like. Assuming Mr. S would prefer to have any of the offered positions to having none of them, Mr. S would then find himself in a pickle: Mr. S can’t rationally choose option $a$, because he prefers $c$ to $a$, and he can’t rationally choose option $c$ because he prefers $b$ to $c$, nor can he choose option $b$ because he prefers

\textsuperscript{17} Gustafsson, Johan E. (2013). The Irrelevance of the Diachronic Money-Pump Argument for Acyclicity. Journal of Philosophy 110 (8):460–464. This does seem to be closer to Davidson et al.’s intended argument. In discussing why Mr. $S$’s preferences are irrational they write “the obvious principle would appear to be this... a rational choice is one which selects an alternative to which none is preferred. But it is clear that the set of Mr. $S$'s preferences makes a rational choice impossible, for whichever alternative he chooses there will be another alternative which is preferred to it. We may imagine a scene in which the point becomes obvious.” And which point they introduce the money pump.
to $b$; because his preferences are intransitive, rational choice is impossible and hence, this set of preferences is irrational. This version of the money pump argument doesn’t rely on sequential choice, and therefore objections that appeal to foresight and resolute choices will be of no use.

Though ultimately I will argue that the money pump argument fails under either interpretation, I find the second interpretation much more plausible than the first. I will shortly offer a criticism that strikes against both interpretations, but first I will give some considerations that suggest that we should prefer the second interpretation.

My preference for the second interpretation is motivated by an intuition concerning the following two sets of preferences:

**Set One:**
1. Prefer full professor at a salary of 5,000 to associate professor at a salary of 5,500.
2. Prefer associate professor at a salary of 5,500 to Assistant professor at a salary of 6,000.
3. Prefer assistant professor at a salary of 6,000 to full professor at a salary of 5,000.

**Set Two:**
1. Prefer full professor at a salary of 5,000 to associate professor at a salary of 5,500.
2. Prefer associate professor at a salary of 5,500 to Assistant professor at a salary of 6,000.
3. Prefer assistant professor at a salary of 6,000 to full professor at a salary of 5,000.
4. Prefer to never make more than two trades with the same person.\(^\text{18}\)

\(^{18}\) I suspect many will want to say that this preference is irrational; however, it need not be. Perhaps the agent made a promise to his very conservative father that he would not make more than two trades. So although he himself does not have anything against trading, he prefers honoring the memory of his father to any benefit that might be secured by trading. Such a preference would, I take it, be abnormal, but could be subjectively rational.
My intuition about the above sets of preferences is that if \textbf{Set One} is irrational, so is \textbf{Set Two}, and if these preference sets are irrational, they are irrational for the same reason. However, if exploitability is what makes intransitive preferences irrational, an agent with \textbf{Set Two} preferences must not be irrational as an agent with \textbf{Set Two} preferences will not take enough trades to be exploited.\footnote{At least not for that reason.} Therefore, the first interpretation of the money pump forces us to conclude that either \textbf{Set One} preferences are irrational while \textbf{Set Two} preferences are rational, or that if they are both irrational they are not irrational for the same reason. This is highly implausible.

With these two candidate interpretations of the money pump argument in place, I will now turn to my objections to both arguments.

\textbf{Objection One}

The above discussion of the money-pump argument has centered on preferences over outcomes: the outcome of which job the subject will receive, but preferences do not necessarily concern outcomes; preferences can range over outcomes, actions, and alternative states of affairs. I have used the terminology of “acts, states, and outcomes” to mark an important departure I take from the decision theory of Leonard Savage. For Savage, preferences are exclusively over acts and outcomes and Savage is quite clear that he understands preferences as essentially tied to action. “I think it of great importance that preference, and indifference … be determined, at least in principle, by decisions between acts and not by response to introspective questions.”\footnote{Savage, Leonard J. (1954). \textit{The Foundations of Statistics}. Wiley Publications in Statistics.} I depart from Savage in
that I think preferences can range over more than just acts and outcomes. I can and often do have preferences over alternatives that I know couldn’t possibly be a consequence of my actions. I would prefer it if *Heaven was for real* which is to say that I prefer heaven existing to heaven not existing; however, I have no illusion that I could choose whether heaven existed or not, or that heaven’s existence could possibly be a consequence of my actions. This poses a problem for the money pump argument. I have already defined **Set One** and **Set Two** preferences, now let me introduce **Set Three**:

**Set Three:**

1. Prefer the Christian God existing to Zeus existing.
2. Prefer Zeus existing to no God existing
3. Prefer no God existing to the Christian God existing.  

Neither interpretation of the money-pump will apply to these intransitive preferences. I could not be offered a series of trades on the basis of these preferences that would leave me exploitable, and this includes Gustafsson’s synchronic money-pump as I could not possibly trade God’s existence for Zeus’s existence. This is a problem for the money-pump argument, because presumably if transitivity is a rational constraint on **Set One** and **Set Two** preferences, it is a constraint on **Set Three** preferences and presumably transitivity applies to **Set One** and **Set Two** and **Set Three** preferences for the same reason. Since the money-pump argument only applies to **Set One** and **Set Two**

---

21 I could just as comfortably used Jeffrey’s formulation of preferences for news items in which case the first preference of **Set Three** could be equivalently expressed as the preferring to receive the news that the Christian God exists the receiving the news that Zeus exists. Of course, I am assuming here that to have a preference is to have a particular kind of mental state—that is that preferences are in some sense psychologically real. This denies the behaviorist idea of preferences as revealed by behavior which Savage seems to indorse. Jeffrey, Richard (1983). *The Logic of Decision*. University of Chicago Press. p 83
it cannot be the reason that rational preferences must be transitive. Therefore the money-pump argument fails.

I have just concluded my first objection to the money-pump argument. I will now advance my second objection.

**Objection Two**

As I have already argued, the dispute between the first interpretation of the money pump argument and the second is on where to place the locus of irrationality. The first interpretation places the locus of blame on the exploitability of the agent who has intransitive preferences. The second sees the exploitability as merely a symptom of the true source of irrationality: the violation of the non-dominated choice principle. Here I will argue that both interpretations are mistaken as both place the locus of blame for the irrationality on the wrong kind of source—that is, neither the exploitability of the agent nor the violation of the non-dominated choice principle are the kinds of reasons for which an agent’s preferences can be irrational. To defend this conclusion I will repurpose Kavka’s famous toxin puzzle.\(^{22}\)

In his now famous 1983 paper “The Toxin Puzzle” Gregory Kavka invites us to imagine ourselves in the following scenario:

You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects…. The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed….. All you have to do is sign the agreement and then intend at midnight tonight to drink the stuff tomorrow afternoon. You are

perfectly free to change your mind after receiving the money and not drink the toxin. (The presence or absence of the intention is to be determined by the latest 'mind-reading' brain scanner and computing device designed by the great Doctor X. As a cognitive scientist, materialist, and faithful former student of Doctor X, you have no doubt that the machine will correctly detect the presence or absence of the relevant intention.)

Kavka suspects that a rational agent will not be able to form the intention to drink the toxin and this despite having a reason to form the intention to drink the toxin. If this is so, the agent will lose out on the million dollars. One might imagine someone criticizing the subject for failing to form the relevant intention and losing the money, but it seems that the subject would have a good defense against this criticism. The subject could argue “look, I know I had a reason to form the relevant intention, the trouble is, it’s not the right kind of reason for forming intentions, and hence it is not a reason for which the rationality of my intentions can be criticized.”

Now imagine the same situation as before, but for a minor difference: the subject is paid the million dollars not if they can form an intention to drink the toxin, but if, at midnight, they can prefer drinking the toxin to not drinking the toxin, tomorrow afternoon. My suspicion is that this difference changes very little in the puzzle. One is not able to rationally prefer drinking the toxin tomorrow afternoon because one knows that by tomorrow morning, one will either have the million dollars or not. If one has the million dollars, they prefer not to drink the toxin. If they don’t have the million, they prefer not to drink the toxin. Hence, although one would certainly prefer to prefer to drink the toxin, one is unable to prefer drinking the toxin to not drinking the toxin. As

before, one has a reason to form some attitude, this time a preference, yet one is unable to form this attitude, because one does not have the *right kind* of reason.

Suppose we changed the puzzle a little more. Suppose that instead of being positively rewarded for forming the right preference, the subject is to be negatively punished: if the subject is unable to form a preference for drinking the toxin tomorrow, they will suffer a swift kick in the gut. I suspect that this doesn’t significantly impact the rationality of forming a preference for drinking the toxin. The conclusion to draw from this is that just because having a certain preference leads to good or bad consequences is not a reason to think that a given preference is rational or irrational, because the consequences of the preferences are not the sort of reason for which one can form a preference.

Again we might imagine someone criticizing the subject for failing to form the relevant preference, but just as before, it seems perfectly appropriate to rejoin that just because having a preference leads to good or bad consequences is not itself a reason to have or not to have a preference, and hence it is not the kind of reason for which the subject’s rationality can be rightly criticized.

This cuts directly against the first interpretation of the money pump argument. Recall that the first interpretation of the money pump argument looked like this:

1. No rational set of preferences would lead an agent to take a series of trades that leaves her having lost money and better off in no respect.
2. If an agent’s preferences are intransitive, then that agent could be made to take a series of trades which leave them having lost money and better off in no respect.
3. So, intransitive preferences are irrational.

But if what I have argued is correct, we have reason to be suspicious of premise one as premise one assumes that the negative consequences of a set of preferences can make a
set of preferences rational or irrational; however, the lesson of the toxin puzzle is that this is not the case.

Does the second interpretation of the money pump argument fare better? Recall that the second interpretation went like this:

1. If an agent’s preferences are such that the agent cannot make a choice without violating the non-dominated choice principle then those preferences are irrational.
2. If an agent’s preferences are such that they could be made to take a series of trades which leave them having lost money and better off in no other, then that agent’s preferences are such that the agent cannot make a choice without violating the non-dominated choice principle.
3. The agent S’s preferences are such that they could be made to take a series of trades which leave them worse off in some respect and better off in no other.
4. Agent S’s preferences are irrational.

This version of the money pump argument does not assume that the loss of money is the locus of irrationality, but rather that intransitive preferences cause you to violate a principle of rational choice: the non-dominated choice principle, and this is why intransitive preferences are irrational. Is this enough to skirt my objection?

No. The violation of the non-dominated choice principle is just another kind of bad consequence. It is certainly nice to have preferences that permit a choice that does not violate the non-dominated choice principle, just as it is nice to have a million dollars, but just as a million dollars is not itself a reason to change one’s preferences, not violating the non-dominated choice principle is not itself a reason to change one’s preferences.

Suppose that instead of taking advantage of Mr. S, the department chair pointed out that Mr. S’s preferences precluded rational choice because they caused him to violate the non-dominated choice principle no matter what he chooses. It seems Mr S. could reasonably respond “I understand that in light of my preferences I cannot choose an
option to which none is preferred, but that is not itself a reason to think that my preferences are irrational. Perhaps, in light of what I value, there is no best option. Why is that irrational? Additionally, I cannot simply decide to prefer option \( a \) to option \( c \) so that my preferences straighten out, I don’t prefer option \( a \) to option \( c \) because I think difference in salary is sufficiently great so as to outweigh the difference in rank.

Similarly, I have reasons for all of my preferences.”

Both accounts of the money pump argument fail and they fail for the same reason: both identify the wrong kind of reason as the locus of irrationality for intransitive preferences. Therefore, if intransitive preferences are irrational it is not for concerns of money-pumps or violations of non-dominated choice principles. Why then, should we believe intransitive preferences are necessarily irrational? Why posit a principle like transitivity in the first place?

Although I do not find the semantic argument or the two interpretations of the money-pump argument to be sound, I do not think the failures of these arguments should make us completely agnostic about whether or not strict preference is a transitive relation. One reason for thinking that ‘preferred to’ is a transitive relation is that it is intuitive, it just seems right. Furthermore I think there is room for a third interpretation of the money-pump which suggests intransitive preferences are irrational. This interpretation is as follows:

1. Intransitive preferences can sometimes make a person into a money pump.
2. Money-pump behavior does not seem like the kind of activity rational people would engage in.
3. So intransitive preferences are likely irrational.
This version of the money-pump is not sensitive to the objections I have raised as it does not posit that intransitive preferences are irrational \textit{for the reason} that they leave subjects exploitable. Instead this version suggests that the money-pump-exploitability is evidence that there is something irrational in having intransitive preferences. This version of the money-pump does not offer a reason \textit{why} intransitive preferences are irrational, it merely argues \textit{that} intransitive preferences are irrational. Ultimately to make this argument \textit{prove} that intransitive preferences are irrational we would have to demonstrate that there is no rational ranking of alternatives in terms of preference that is intransitive. However, even if this argument is not a proof, surely it is \textit{evidence} that intransitive preferences are irrational.

I think that the strength of the money pump is that it shows that intransitive preferences can sometimes lead to behavior that seems manifestly irrational. While this behavior is not itself the reason that intransitive preferences are irrational, it is a reason \textit{to think that} intransitive preferences are irrational. Therefore, if there are no compelling objections to transitivity—no reasons to think intransitive preferences \textit{can} be rational we should trust transitivity. However, there are numerous arguments that have purported to show that rational preferences can be intransitive.\footnote{I will not be able to address all of the important counterexamples to transitivity that have been argued for in the literature. The two most important arguments I will not be addressing are Larry Temkin’s continuum argument for the intransitivity of ‘all things considered better than’ which he argues for in Temkin, Larry S. (1996). \textit{A Continuum Argument for Intransitivity}. \textit{Philosophy and Public Affairs} 25 (3):175–210. For important criticisms of Temkin’s arguments I recommend Voorhoeve’s review of Temkin book \textit{Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning} in Voorhoeve, Alex (2013). \textit{Vaulting Intuition: Temkin’s Critique of Transitivity}. \textit{Economics and Philosophy} 29 (3):409-425. Stuart Rachels has also argued both for the intransitivity of ‘better than’ and ‘preferred to’ along similar lines as Temkin. I would suggest Rachels, Stuart (2005). \textit{Counterexamples to the Transitivity of ‘Better Than’}. In Toni Rønnow-Rasmussen & Michael J. Zimmerman (eds.), \textit{Recent Work on Intrinsic Value}. Springer. 249—263 as a good place to start.}
thesis will be dedicated to showing that two important putative counterexamples to transitivity fail to substantiate their conclusion, and I will turn to this task in the following two sections. However, before turning to these putative counterexamples I will spend the remainder of this section explaining what I take the *right kind of reasons* for which preferences can be irrational. I turn to this discussion presently.

**The Right Kind of Reasons**

I have argued that one reason the money pump argument fails is that it places the locus of rationality on the wrong kind of reasons for preferences. Here I sketch what I take the right kinds of reasons to be, and hence, the right kinds of reasons for which intransitive preferences are irrational.

The first distinction I would like to suggest is between being mistaken about one’s preferences, one’s preferences being mistaken, and one’s preferences being irrational. To motivate this distinction consider the following case:

**The Restaurant Case:**

Alice is visiting friends in a town she has not visited in quite some time. She and her friends have elected to go to a local Mexican food restaurant for lunch. Alice remembers dining at this restaurant once before and she remembers quite clearly that she did not enjoy her meal. She does not remember, however, what she ordered, she only remembers that it was most unpalatable. Alice also recalls that the meal her companion ordered last time looked quite delicious. She does not recall, however, what her friend ordered. Alice would prefer ordering what her friend ordered last time to ordering what she ordered before. Alice looks over the menu which has two options *Combo #1* and *Combo #2*. She reads over the description of both combos and decides for good reasons that her friend definitely ordered *Combo #2*. When the waitress arrives and asks for her order Alice orders *Combo #2*. Some time passes and the waitress brings Alice *Combo #2* as she ordered and Alice instantly recognizes *Combo #2* as what she ordered last time.
My intuition about this case is that although the waitress did bring Alice what she ordered, she did not bring Alice what she preferred. Alice was mistaken about her preferences. She thought she preferred **Combo #2** to **Combo #1** when in fact she preferred just the opposite.

Compare the restaurant case to the following:

**The Genie Case:**

Stephany was walking along the beach when a peculiar glimmer caught her eye. Upon investigating, Stephany discovered a magic lantern out of which sprung a magic genie. Naturally, Stephany was quite excited and asked the genie if she would be allowed to wish for whatever she desired. The genie, of course, rolled his eyes frustrated with the misinformation about his kind propagated by children’s stories. “No” the genie corrected, “we don’t do that, we’ve never done that, you get to choose between the same two options everyone else does: would you like a million dollars or ten thousand dollars?” Stephany considers her options carefully and decides to choose the million dollars over the ten thousand reasoning that she prefers the option with the greater amount of money. The genie, surprised, gives Stephany her million dollars and remarks “I would have thought you would have chosen the option that would have brought you the most happiness” before disappearing back into the lantern.

I have two intuitions about this case. First, that, unlike the restaurant case, the genie **did** give Stephany what she preferred: the million dollars. Second, on the assumption that Stephany only cares about money in so far as she thinks money will make her happy, and that the genie is correct and the ten thousand dollars would bring Stephany more happiness in her life, her preferences were mistaken. Her preferences were mistaken because given what she values, she did not prefer the best option.

The first point I would like to make in light of the restaurant case and the genie case is that only preferences that are in some way mistaken can be irrational. The second point is that not all preferences that are mistaken need be irrational. Arguably, in light of the evidence Stephany had available to her, it was rational for her to prefer the million
dollars to the ten thousand dollars. This leaves two final questions: When are mistaken preferences irrational? Why are intransitive preferences necessarily irrational?

Our preferences follow from two more basic features of our psychology: our desires and our beliefs. By desires I mean the things we find to be worthwhile or valuable, and sometimes I will use the word ‘value’ in place of ‘desire’; these need not be the same for everyone, but the standard list tends to include things like pleasure, love, friendship, autonomy, truth, and knowledge. There may or may not be external rational constraints on what an agent desires—that is, there may not be a list of things the agent must value in order to count as rational—I need not take a side on this issue for my arguments to go through. Though an agent may find two things to be intrinsically desirable, for instance love and knowledge, she need not desire them equally to be rational. Perhaps she values love and knowledge intrinsically, she may desire love more than knowledge and if presented with her choice between a life which exemplifies love and one that exemplifies knowledge, she may rationally desire the life of love more than the life of knowledge. In addition to desires, there are beliefs. The beliefs I have in mind here are the beliefs about what the options are how the various options under consideration will satisfy or dissatisfy our desires.

For an example of how desires and beliefs come together to form preferences, consider an agent who prefers pie A to pie B. This preference follows from two considerations. One, she desires pleasure. Two she believe that apple pies are better tasting than blueberry pies (and hence more pleasurable to eat) and she believes that pie A is an apple pie and Pie B is a blueberry pie. My assumption is that all preferences will have a similar relationship between beliefs and desires.
We now have the appropriate tools to answer the questions with which this talk of desires and preferences began: in light of what are preferences rational or irrational? Irrational preferences are not irrational because they lead to bad consequences, contrary to what the money-pump argument has supposed, but because they are the product of inconsistent or incoherent evaluations of options which arise from either incoherent desires or incoherent beliefs.

Why are intransitive preferences necessarily irrational? Intransitive preferences are irrational because they are the result of inconsistent evaluations of alternatives— that is, no consistent evaluations of alternatives could result in an intransitive preference ordering. My suspicion is that this could be proven with a formal model, but I confess that I don’t, as of yet, have such a model to give. Instead, for the remainder of this thesis I will look at some influential putative counterexamples to transitivity and argue that they are not successful. If the best cases for rational intransitive preferences fail, either because it can be shown the agent is not evaluating their options in light of their desires and beliefs consistently or that the situation giving rise to the counterexample is impossible, then we will have good reason to suppose that intransitive preferences are irrational. In the following sections I will address putative counterexamples of intransitivity.
SECTION TWO

In the previous section I argued that the standard argument against the rationality of intransitive preferences fails because it attends to the wrong kind of reasons for evaluating preferences. Preferences are not irrational because they lead to bad consequences, they are irrational in so far as they are the result of some inconsistency or mistake in the evaluation of alternatives in light of the agent’s desires and beliefs. If we are to maintain the intuitive conclusion that intransitive preferences are irrational in the face of putative counterexamples we need arguments that focus on these evaluations.

In the following two sections I will look at two arguments that have purported to show that rational preferences, preferences that follow from a reasonable and consistent evaluation of alternatives in light of desires and beliefs, can be intransitive. These arguments can be broken up into two distinct strategies. One strategy involves a sort of continuum argument, presenting a series of alternatives with very slight differences between consecutive alternatives which culminate with vastly distinct endpoints and then arguing that a rational agent’s preferences over these alternatives might rationally be intransitive. Warren Quinn’s puzzle of the self-torturer uses a variation of this strategy, and will be discussed in this section. The second strategy involves evaluating alternatives along a plurality of values and desires. This strategy elicits an apparently rational intransitive preference ordering amongst options which are so different that they cannot all be judged along the same values. I will address this strategy in the next section. I will tackle both strategies separately as I believe they require different treatments, beginning with the first strategy. I will start my analysis of the first strategy by presenting Quinn’s puzzle of the self-torturer.
Quinn and the Puzzle of the Self Torturer

Warren Quinn’s famous puzzle of the self-torturer has been a persistent challenge to transitivity, and has sparked a heated debate in the literature. I will primarily focus on Quinn’s original argument using the more recent discussion to clarify the problem raised by Quinn and evaluate purported solutions. In his 1990 paper “The Puzzle of the Self-Torturer” Quinn invites us to imagine a medical device, with settings ranging from 0 to 1000, capable of administering an electric current to the body in very precise increments; so precise, in fact, that the difference in electrical current between any two adjacent settings is so slight that adjacent settings feel no different—at setting 1, the amount of electricity in your body is so minute that it feels no different from setting 0 at which there is no current at all.

Now imagine an agent, whom Quinn pessimistically dubs “the self-torturer”, agrees to have the device attached to him permanently. He begins at setting 0 and every week he is given the option to either advance a setting or to remain at his current setting. He is given a trial period at the start of each week to experiment with different settings before the device resets to its previous setting. Should he advance a setting, he will be paid a sum of 10,000 dollars. However, he can never go back a setting once he has advanced, and he cannot advance more than one setting a week.

Although the self-torturer finds himself in unusual circumstances, he is otherwise really very much like us. The self-torturer values money and disvalues pain,

---

and although there are some amounts of pain he would tolerate for large amounts of money, there is no amount of money that would make a life of agony preferable.

In light of these values, the self-torturer prefers setting 1 to setting 0 because setting 1 is worth an additional 10,000 dollars but feels no more painful than setting 0. The self-torturer prefers setting 2 to setting 1 for the same reason, as with setting 3 to setting 2, setting 4 to setting 3 … and setting 1000 to setting 999. However, while the difference in electrical current between adjacent settings is slight enough that the self-torturer cannot feel the difference, setting 1000 and setting 0 are sufficiently far apart that the self-torturer can feel the difference—where setting 0 is pain free, setting 1000 is intolerable. Because the self-torturer prefers a life of penury to a life of agony, the self-torturer prefers setting 0 to setting 1000 and would even trade the fortune he has amassed to return to setting 0. Quinn suspects that although the agent’s preferences are clearly intransitive, it is not at all clear that the agent is irrational and hence we are left with a puzzle.\(^{26}\)

In this section I will argue that Quinn’s puzzle is not a genuine counterexample to transitivity. I will argue that there are three options in responding to the puzzle and the most plausible of which involves denying that Quinn’s puzzle is coherent; however before I discuss what these options are and which option ought to preferred, I wish to

\(^{26}\) It’s worth noting that the self-torturer’s preferences aren’t merely intransitive, they are cyclical and hence are prime candidates for a money pump argument, though virtually no one argues on the basis of the money-pump that the self-torturer is irrational. I take this to be further evidence that the money-pump is not a convincing argument when the values and beliefs of the agent are known and seem rational. Frank Arntzenius and David McCarthy are a potential exception, arguing “One indication that there is something wrong with Harry’s reasoning is this: he can be turned into a money pump in a way he could easily foresee. . . . This makes it even clearer that there is a real problem about his sequence of apparently rational decisions.” Arntzenius, Frank, and David McCarthy. "Self Torturer and Group Beneficence." Erkenntnis 47, no. 1 (1997): 129-44
clarify exactly what the counterexample to transitivity is, and how Quinn’s puzzle is supposed to work. I will present two alternative interpretations of the puzzle, and argue that only one poses a genuine counterexample to transitivity.

**Two Alternative Interpretations**

Quinn claims that the self-torturer “cannot feel any difference in comfort between adjacent settings,” but this could mean one of two things. It could mean that adjacent settings feel no different—adjacent settings feel the same—or it could mean that while adjacent settings do feel different, this difference is so slight that the self-torturer doesn’t, or can’t, notice the difference—adjacent settings are indiscriminable.

Quinn offers some clarification of his puzzle in responding to three anticipated objections regarding the phenomenal difference between consecutive settings. The first objection is that we might be able to discern a difference in comfort between adjacent settings by attending to behavioral evidence. Perhaps the self-torturer cannot introspect a difference between adjacent settings, but there might be other evidence: the self-torturer might be more irritable at the later setting, or look more uncomfortable. The second objection is that the self-torturer may be able to discern a difference between adjacent settings by triangulating with some third setting. Setting 1 may be indiscernible from

---

27 Alex Voorhoeve and Ken Binmore have pushed a version of this objection. Imagine the self-torturer is experimenting with different settings recording whether or not the settings feels comfortable with reports like “very comfortable” “slightly uncomfortable” or “very uncomfortable”. The self-torturer might be able to discern a difference between consecutive settings by attending to the discrepancies between reports. If setting S is described as “very uncomfortable” 49 percent of the time when following setting 0 and setting S + 1 is described as “very uncomfortable” 51 percent of the time when following setting 0, then the self-torturer has evidence that setting S + 1 is slightly more painful than setting S. Voorhoeve, Alex & Binmore, Ken (2006). Transitivity, the Sorites Paradox, and Similarity-based Decision-making. *Erkenntnis* 64 (1):101-114
setting 2 and setting 2 from setting 3, but if setting 1 is discernibly different from setting 3 then there must be some difference in how setting 2 and setting 3 feel. Thus the self-torturer has reason to think that his comfort does decline in moving from setting 2 to setting 3.\footnote{Quinn p.81 Donald Regan develops this triangulation objection more fully in Regan, D.: 2000, 'Perceiving Imperceptible Harms: With Other Thoughts on Transitivity, Cumulative Effects, and Consequentialism', in M. Almeida (ed.), \textit{Imperceptible Harms and Benefits}, Kluwer Academic Publishers, Dordrecht, pp. 49-73. Erik Carlson has also pushed Quinn on this point arguing that it is impossible that settings 1000 and settings 1 are perceptually distinguishable while there is no phenomenal difference detectible by triangulation. I won’t review these arguments here as nothing in my objections to Quinn rest on this point. Carlson, Erik. "Cyclical Preferences and Rational Choice." Theoria 62 (1996): 144-60. Especially pages 152-154.}

Quinn counters that although triangulation and behavioral evidence might be sufficient to detect the phenomenal differences between very similar stimuli, he is imagining a case in which the objects of comparison are so similar that there is no phenomenal difference detectible by introspection, triangulation, or behavioral evidence. Quinn writes:

\begin{quote}
If there are increments of voltage just small enough to be directly undetectable, it seems there might be even smaller increments that cannot be detected by triangulation. And I want such a case. . . . What naturally matters [to the self-torturer] is that the comfort status of \( s \) and \( s+1 \) are, introspectively and behaviorally, no different – either in direct comparison to each other or in oblique consideration with any third setting. It is enough for him that the empirical data give him no reason to suppose that his comfort declines either directly or relative to some fixed point, in any step.\footnote{Quinn 83}
\end{quote}

Although these passages clarify that Quinn has in mind a case in which adjacent settings are indiscriminable, it is still open for interpretation whether or not the adjacent settings feel the same.
We are left then with two alternative interpretations of the puzzle. On one interpretation, although there may be slight differences in how some of the adjacent settings feel, this difference is so slight that the self-torturer does not notice this difference. On the other, adjacent settings are pairwise indiscriminable because the difference in electricity between adjacent settings is so slight that adjacent settings feel the same. I think the second interpretation alone presents a plausible counterexample to transitivity, but I will briefly consider and reject the first interpretation.

**The First Interpretation**

Discriminability has to do with the agent’s ability to tell the difference between the relevant objects or experiences. Often, two experiences are indiscriminable in some relevant respect because these experiences look or feel the same; however these can come apart. Suppose that setting 1 does feel slightly different than setting 0, it could still be the case that setting 1 and setting 0 are indiscriminable. Perhaps this is due to the self-torturer’s imperfect memory – the self-torturer never feels setting 1 and setting 0 at the same time, so when the self-torturer compares setting 1 and setting 0 he has to rely on his memory. If the self-torturer’s memory is sufficiently limited he may simply forget

---

30 Sergio Tenenbaum and Diana Raffman have defended a version of the self-torturer puzzle in which it is not assumed that adjacent settings feel the same. They argue that “the puzzle doesn’t require adjacent settings of the dial to be indiscriminable. It seems equally rational to prefer large sums of money over nearly imperceptible, or even just slight, differences in pain, and yet prefer abject poverty over sustained agony; again, these seem to be the preferences of most ordinary agents.” It is not clear, that this is right. By hypothesis there are certain amounts of pain that the self-torture would prefer to avoid at any monetary cost. Presumably, then, the self-torturer would not prefer to advance to setting that had that amount of pain, even if it is only slightly different then the setting he is currently at, for any amount of money. Tenenbaum, Sergio & Raffman, Diana (2012). Vague Projects and the Puzzle of the Self-Torturer. *Ethics* 123 (1):86-112.
the slight difference in how the earlier setting felt. \textsuperscript{31} If memory is not the issue, perhaps the self-torturer just doesn’t notice the difference in how setting 1 and setting 0 feel because the difference are really quite small. To make the difference between feeling the same and being indiscriminably different salient, we could suppose that some warlock snuck into the self-torturer’s bedroom at night, stole his arm, and planted false memories that he had never had a right arm to begin with. I imagine that although in the morning the self-torturer would not be able to discriminate any difference from the way he was the night before, he would in fact feel different. This would be an instance in which there is a relevant phenomenal difference between two phenomenal states but one that is indiscriminable.

This distinction is relevant, as presumably the self-torturer cares about how different settings feel, and only cares about his ability to discern the differences between adjacent settings in so far as this is his means for determining whether adjacent settings feel the same. If this is so, then Quinn might need something stronger than mere indiscriminability to generate the intransitive preference orderings.

One could, under the first interpretation, develop an argument against transitivity along the following lines: Though there are in fact slight differences in comfort between some adjacent settings, the self-torturer cannot detect a difference, in any way, between adjacent settings. It seems to him that adjacent settings feel the same. Hence the self-

\textsuperscript{31} Alex Voorhoeve and Ken Binmore have given a version of this argument “it is conceivable that when Alice is presented with two similar stimuli in succession, in her memory the first stimulus always becomes assimilated to the second, so that she finds them introspectively indistinguishable.” Sergio Tenenbaum and Diana Raffman dispute this point in footnote 17 of Tenenbaum, Sergio & Raffman, Diana (2012). Vague Projects and the Puzzle of the Self-Torturer. Ethics 123 (1):86-112.
torturer rationally comes to believe that should he advance a setting, his comfort will not decline, and he will be paid handsomely for the advance. Such a belief is in perfect accordance with his evidence. Therefore he rationally forms a preference on the basis of this belief to advance to the next setting. This reasoning will apply to all adjacent settings, so the self-torturer will soon find himself at setting 1000, but setting 1000 is tortuous and he can certainly feel the difference between setting 0 and setting 1000 and will gladly give up his fortune to return to setting 0, which he prefers to setting 1000. The self-torturer’s preferences are intransitive, but perfectly rational, given his evidence.

I have two objections to this formulation of the self-torturer argument. The first is that it is not clear that the self-torturer genuinely prefers to advance at each step. The second is that on the assumption that the self-torture is not mistaken about his preferences it is not clear that his preferences are rational as the beliefs generating this preference ordering is incoherent. I will begin with the first objection.

I think that in this interpretation of the puzzle, the self-torturer is mistaken about his preferences for at least one advance. Recall that to be mistaken about one’s preference is to think that one’s preferences are other than they in fact are. In my restaurant example, Alice thought that she preferred Combo #2 to Combo #1 but she was mistaken about this, because what she thought to be Combo #2 was in fact Combo #1. I think the self-torturer is similarly mistaken for at least one pair of settings. Ex-hypothesi at least one of the settings, setting 1000, is so painful that no amount of money could compensate for a life at that setting. It follows that there are some settings for which no amount of money can compensate, call these the intolerable settings. So, assuming that there are slight differences in how adjacent settings feel (as this
interpretation assumes) there is a setting for which the advance to the next setting is not worth the 10,000 dollars—the setting which takes the self-torturer from a tolerable setting to one of the intolerable settings. For convenience I will suppose that these settings are setting 900 and setting 901 where setting 901 is the first intolerable setting.

Now under this formulation of first interpretation, the self-torturer arrives at setting 900. As he has done before, he experiments with setting 901 to see if he can discern any difference between this setting and setting 900 and just as before, he finds any difference between these settings to be indiscriminable. He takes the fact that he can’t discriminate a difference to be good evidence that there is no difference and concludes that setting 901 feels the same as setting 900. He then comes to believe that should he advance he would be at a setting which feels the same as his current setting, but receive an additional 10,000 dollars. He decides to advance.

The question to ask is whether the self-torturer genuinely preferred to advance to setting 901 or did he merely think he preferred to advance? Given the structural similarity to the restaurant case I gave in the previous section, it is plausible that the self-torturer did not genuinely prefer to advance. The self-torturer thought he preferred to advance to setting 901 because he prefers a setting which feels the same as setting 900 but worth 10,000 more, and he thought setting 901 was one such setting. However, the self-torturer is mistaken about this, and setting 901 is not the setting he prefers, because

---

32 Duncan MacIntosh gives a variant of this argument in MacIntosh, Duncan (2010). Intransitive Preferences, Vagueness, and the Structure of Procrastination. In Chrisoula Andreou & Mark D. White (eds.), The Thief of Time. Oxford University Press. Here he argues that the self-torturer could begin at 1000 which is intolerable and scroll back the settings until he finds the first tolerable setting.
setting 901 feels differently than setting 900 and the difference is not worth 10,000 dollars. This concludes my first objection to this interpretation.

My second objection is that on the assumption he genuinely prefers to advance all the way to 1000 only to prefer 0 to 1000, it is not clear that his preferences are rational. My first objection grants the assumption that the belief of the self-torturer that consecutive settings feel the same was rational given the pairwise indiscriminability. My second objection questions this assumption.

The reason I am suspicious of the rationality of the self-torturer’s belief is that I doubt it is coherent to believe that each setting feels the same as the setting before it while also believing that the last setting is more painful than the first setting. The argument against transitivity from the first interpretation takes as a premise that the indiscriminability of the differences between settings gives the self-torture evidence that consecutive settings feel the same as each other, but the fact that the self-torturer begins to feel pain in later settings constitutes good evidence that consecutive settings do not feel the same. Where, after all, is all of this pain coming from? This objection is tied to whether or not it is possible to experience phenomenal continuum in which ‘looks the same as’ is not transitive. As this objection will stand or fall on the success of my objections to the second interpretation I will leave this consideration here for now.

The first interpretation fails because it is not rational for the self-torturer to prefer settings for the fact that they are indiscriminably different as the self-torturer only cares about how settings feel. Furthermore even if the self-torturer rationally takes the indiscriminability of setting pairs as evidence that consecutive pairs feel the same this does not generate an intransitive preference ordering, as this merely leads the self-
torturer to be mistaken about his preferences. And finally the beliefs generating these preferences may not be coherent in the first place. In either case, the self-torturer’s actual preferences will be transitive or he will be irrational, and therefore this interpretation of Quinn’s puzzle is no threat to transitivity.

**The Second Interpretation and the Phenomenal Sorites**

The first interpretation will not generate a counterexample to transitivity for the reasons I have argued; however these arguments will not apply to the second interpretation under which adjacent settings are indiscriminable because they actually feel the same. This interpretation would need to invoke the kind of phenomenal continuum presumed in phenomenal sorites arguments.

There are three options for responding to the second interpretation of Quinn’s puzzle: One, we could accept the counterintuitive conclusion that intransitive preferences can be rational. Two, we could argue that although it seems as if the agent is evaluating his options rationally when forming his preferences, we are mistaken in this, the agent is actually irrational (this is option I took in responding to the first interpretation). Three, we could deny that the puzzle Quinn has presented is logically possible—deny that there could be a device where all adjacent settings feel the same, but the first setting feels different from the last. Ultimately I will argue that the third avenue is the route we should take, however I will first consider the second avenue and argue that it is not a plausible way out of this interpretation of the puzzle.

If we are to take the second avenue, we must somehow argue that contrary to appearances, the self-torturer is not rational. This would amount to showing that either the self-torturer’s desires or values are incoherent or the self-torturer is operating under
false beliefs as these are the reasons for which preferences can be rational or irrational.

What are the relevant desires? The self-torturer’s desires are for more money and less pain all else being equal and to avoid a life of agony at any financial cost. What are the beliefs? The self-torturer believes that his comfort does not decline in moving from one setting to the next adjacent setting and that he will be paid ten thousand dollars for each advance. These beliefs in conjunction with the self-torturer’s desires give rise to the cyclical preferences. Where might the irrationality be? The desires seem straightforwardly rational. There doesn’t seem to be anything irrational in desiring more money all things being equal, and it doesn’t seem to be anything irrational in desiring less pain either. Perhaps, then, the self-torturer’s beliefs are the problem; after all, how can it be that the self-torturer’s comfort does not decline in any advancement from one setting to the next and yet he begins in relative comfort and ends up in sheer agony?

If Quinn’s puzzle is coherent, then there is no phenomenal difference between adjacent settings, and since comfort is purely a matter of how things feel, the self-torturer’s comfort really does not decline in any step. Therefore, so long as the self-torturer really is be paid for his advances, the self-torturer’s beliefs seem rational.

The second avenue for making sense of Quinn’s puzzle is a dead end, so long as everything Quinn has told us is correct, the self-torturer is rational in forming his intransitive preferences. Thus if rational preferences are necessarily transitive it must be the case that not everything Quinn has told us is correct. This takes us to the third avenue for which I will argue presently.

The third avenue for interpreting Quinn’s puzzle is to deny that the situation generating the puzzle is possible. Specifically I will argue that the phenomenal
continuum that Quinn is presupposing is logically impossible because ‘looks the same as’ is a transitive relation. The argument I will be advancing against the possibility of Quinn’s continuum is borrowed from Delia Graff-Fara and her 2001 paper *Phenomenal Continua and the Sorites.* I will first present her argument and show why it is a problem for Quinn’s puzzle. I will then consider some recent objections to Fara’s arguments, and defend Fara’s account against these objections. I will now turn to Fara’s argument.

**Fara’s Argument for the Transitivity of ‘Feels the Same As’**

Quinn is committed to denying the transitivity of ‘feels the same as’, otherwise his puzzle will be incoherent: If ‘feels the same as’ is transitive then we could reason as follows: setting 0 is pain free and setting 1 feels the same as setting 0 setting 1 must be pain free. Likewise if setting 1 is pain free and setting 2 feels the same as setting 1 then setting 2 must be pain free. Likewise … setting 1000 must be pain free, but by hypothesis setting 100 is tortuous, so it must be that ‘feels the same as’ is intransitive.

Fara’s argument against the possibility of the kind of phenomenal continuum Quinn is imagining is developed within the context of the phenomenal sorites puzzle. The logical form of the phenomenal sorites puzzle looks something like this:

**PHENOMENAL SORITES ARGUMENT:**

- Premise 1: $F_a$
- Premise 2: If Ryx then if Fx, Fy
- Premise 3: there is a series $a_1 \ldots a_n$ such that $Ra_{i+1}a_i$ for all $i$.
- Conclusion: $Fa_n$.

---


If I were to derive a sorites argument from Quinn’s example, $F$ would stand for the property ‘feels comfortable’ and $Ryx$ would stand for the relation ‘feels the same as’. The series $a_1 \ldots a_n$ would be the consecutive settings with the corresponding electrical voltage.

Given a phenomenal continuum such as the gradual rise of pain in Quinn’s example or the more traditional example of a series of color patches gradually moving from red to yellow, premises one through three all appear to be true, but, supposing that $n$ is a suitably large number, the conclusion is false. As the only inferences are modus ponens, and the conclusion is obviously false, there remains only one way out of the puzzle: we have to deny a premise.

The traditional route has been to deny the second premise which amounts to denying the transitivity of the ‘looks the same as’ relation, and this appears to be the option Quinn has taken; however, Fara’s strategy is to first push the plausibility of the second premise, that ‘looks the same as’ really is a transitive relation, and then undercut the motivation both for thinking that the third premise is true (that we do experience sorites-style phenomenal continua) and for rejecting the transitivity of ‘looks the same as’. If Fara’s argument is successful then Quinn’s puzzle is impossible as there will have

---

35 Because Quinn is using electrical current and pain, the $R$ stands for the “feels the same as” relation, but of course this is not meant to suggest that sorites puzzles always concern pain. The typical relation would be “looks the same as”, but it is equally possible to have a sorites puzzle involving the “sounds the same as” relation or any other sense modality.

36 This isn’t exactly right. There are responses to the sorites puzzle that do not involve denying a premise—one could deny that logic even applied to sorites paradoxes; Frege, Russell and Quine all pushed some version of this thesis, but denying a premise is the most common route and it is the one I will focus on in this paper. Hyde, Dominic, "Sorites Paradox", The Stanford Encyclopedia of Philosophy (Spring 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2014/entries/sorites-paradox/>.
to be a perceptible difference between at least one pair of adjacent settings and an
intransitive preference ordering will not be generated.

Fara’s argument in support of the transitivity of ‘looks the same as’ relations is so
simple it is hard to frame it in a way that does not look question begging. Fara’s
argument is that if ‘looks the same as’ is not a transitive relation, the same-appearance
claims of the form “if one paint patch looks the same as another then if one looks red, so
does the other” are not true, but they are true, so looks the same as is transitive. This is
likely to appear question begging because the very reason we denied transitivity in the
first place is that we had to deny the truth of these same-appearance claims to avoid the
paradoxical conclusions of the sorites puzzle. All the same, Fara’s argument is not
question begging when taken in light of her overall strategy. Fara is trying to evaluate, in
light of the sorites puzzle, which is really the most plausible premise to deny, the truth of
the same-appearance claims or the existence of the phenomenal continua. At this stage in
her argument she is merely stressing the intuitive appeal of same-appearance claims.
After all, imagine the self-torturer were to say ‘Setting 10 feels the same as setting 11,
but setting 11 feels painful but setting 10 does not.’ Could we even make sense of that?
The same-appearance claims Fara is defending don’t merely seem true, they seem
trivially true—its truth seems to follow from the meaning of the terms used in the claim.
Given that two objects look the same it is hard to see how they could differ in any
respect relevant for the applicability of the predicate “feels painful”.

There are two considerations motivating the traditional move of rejection of the
second premise over the third. The first is that we do seem to have phenomenal continua
of the sort Quinn is imagining. Consider a bucket of red paint. It seems entirely plausible
that if a single drop of yellow paint was mixed into the bucket, the paint would look to be the same color. Similarly for the next drop, and the next, until the bucket looked yellow. Because these experiences seem perfectly possible, ‘looks the same as’ must not be transitive.

The second consideration is that a similar move is taken in escaping the original non-phenomenal sorites paradox. In the non-phenomenal sorites paradox our premises look just the same:

**NON-PHENOMENAL SORITES ARGUMENT:**

Premise 1: $F_{a_1}$
Premise 2: If $Ryx$ then if $Fx$, $Fy$
Premise 3: there is a series $a_1 \ldots a_n$ such that $Ra_{i+1}a_i$ for all $i$.
Conclusion: $F_{a_n}$.

Except here, $F$ will stand for a non-phenomenal predicate like ‘is tall’ and $a_1$ will be a seven foot tall man, $a_n$ will be some four foot tall man, and $R$ will stand for ‘is one-hundredth of an inch shorter than.’

Here it seems perfectly appropriate to deny the second premise and absurd to deny the third. Denying the third would amount to saying that there could never be a series of men each, one hundredth of an inch shorter than the last, but clearly this is possible.

Fara’s argument against the second consideration is fairly straightforward. The difference of a hundredth of an inch is relevant to the applicability of “tall”. But in the

---

37 Notice that in this sorites paradox, nothing treads on how tall the man appears to be.

38 As Fara notes, that the second premise is to blame should not entail that there is wide spread agreement as to why it is to blame. This is a source of contention.
phenomenal sorites argument it is stipulated that there is no difference in how the two objects in consideration look, and if this is so, then there is no difference relevant for the application of any visual predicate. Or, in the self-torturer’s case, no difference relevant for the applicability of ‘feels painful’.

The first consideration is more difficult. Against this consideration Fara must argue that we do not, and cannot, experience the kinds of phenomenal continuums that Quinn is imagining: continuums in which red paint gradually becomes yellow, or comfort gradually gives way to pain in a way that there is no difference in how narrow enough section of the continuum look or feel. The problem is further complicated because it is standardly assumed that denying that we have these kinds of phenomenal continua implies that our perceptual systems are perfectly accurate; therefore Fara must also show that denying we experience sorites style phenomenal continua does not implausibly imply that our perceptual systems are perfectly sensitive to changes in stimulus.

Fara argues that the most plausible account of our perceptual limitation is that our perception is finite in the following sense:

**Perceptual Limitation Thesis:**

For some sufficiently slight amount of change (in color, sound, or position, etc.), we cannot perceive an object as having changed by less than that amount unless we perceive it as not having changed at all (as having changed by zero amount).

Fara grants that our perceptual systems may be limited in this way, but stresses that non-transitivity does not automatically follow from this thesis. That our perceptual system is

---

39 Fara does not call this theses by this name. Fara initially presents two possible accounts called “a” and “b” and the one I have quoted here is her “b”. I have omitted A as Fara dismissed it as implausible and it is not a point Phillips contends.
limited in this way only entails that when we perceive something as changing it must change by at least some set amount— that we cannot have an experience of change less than that amount. But this does not entail that we could have a phenomenal continuum wherein adjacent color patches all looked the same but the end points did not, because our perceptual system being limited in this way is compatible with our experiencing change discretely: objects looking to take small, but sudden changes in color or position. Fara argues that to get non-transitivity we must also assert that we experience continuous changes, for then non-transitivity would follow.

Fara argues that it is not obvious whether we experience continuous or discrete changes. She invites us to consider two possible experiences of a mouse cursor moving across a computer screen. First imagine the visual experience of a mouse cursor looking to move discontinuously across the screen, and then an experience of a mouse cursor looking to move continuously—but-jerkily across a screen. Though both experiences would clearly be different it is difficult to say what experience you are having as you move your mouse cursor across the screen. Therefore, the fact that we seem to experience continuous movement and often characterize our experience as being experiences of continuous movement that does not mean that we actually have these experiences. It is possible that if we were having discontinuous experiences of a particular sort, we would still categorize them as continuous.

Furthermore, even if we do experience a phenomenal continuum, it does not automatically follow that ‘feels the same as’ is not a transitive relation. All that is required of a spectrum to be continuous is that for any subsection of that spectrum representing a positive change in color, there is a narrower width of that subsection
wherein the color looks to change by less than the change of the wider subsection. This does not entail that we perceive narrow enough subsections of continuous changes as homogenous, which would entail that ‘looks the same as’ is intransitive. Fara considers one way we might motivate such a ‘homogeneity thesis’ on these grounds: It could be that when confronted by a phenomenal continuum—say, that of the slow motion of an hour hand of a clock moving around the clock face, we are not able to visually discriminate certain very close together positions. That is why the hour hand does not appear to move over a period of ten seconds. While the hour hand has, of course, moved, the position it has moved to is not far enough away from where it was for use to discriminate the distance. So it looks as if it has stood still. After a longer period of time, the hour hand has moved a great enough distance to be perceivable.

Fara thinks this explanation is implausible as it should leave us wondering why then all movement isn’t perceived as slow motion. The second hand sweeping around the clock face appears to be moving constantly, but if our above account of slow motion is correct there are certain changes in distance which should be perceived as no change at all. So for the sufficiently short enough distance wherein the second hand has not changed in position enough for that change to be perceived it ought to look as though it is standing still. Of course, it never looks to be standing still. So we ought to think that the homogeneity thesis, although possible, is likely false.40

40 Ian Phillips has raised some problems for this account. He argues that Fara has smuggled in an implausible “zoetrope” conception of the perception of time. This objection may successfully undercut Fara’s motivation for thinking that the homogeneity thesis is false, it does not provide any motivation for thinking the homogeneity thesis is true either. I will pay more attention to Phillip’s objections in a later section. Phillips, Ian (2011). Indiscriminability and experience of change. Philosophical Quarterly 61 (245)
If Fara rejects both the homogeneity thesis and the existence of phenomenal continua, how can she account for the very plausible intuition that we could be presented with a color spectrum wherein adjacent sections of the spectrum look the same and yet the end points do not?

Here Fara introduces a distinction between two stimuli looking different and noticing that two stimuli look different. Perhaps it is not the case, as our original phenomenal sorites argument suggested, that there could be a series $a_1 \ldots a_n$ such that $Ra_{i+1}a_i$ for all $i$. The reason we think that there could be such a series is that when confronted with a gradual change in color we often fail to notice the difference between how adjacent colors look. For the hour hand example this would amount to saying that although the hour hand does look to be in a different position after ten seconds, we fail to notice this difference and hence report that the position of the hour hand has not changed.

Fara’s overall argument could be tersely summarized as follows: We have good intuitive grounds to think that ‘looks the same as’ is a transitive relation because we have good reasons to think same-appearance claims are necessarily true in virtue of the meaning of the terms. We do not, however, have as strong a ground to think that the phenomenal continua needed to deny the transitivity of ‘looks the same as’ are possible. This is because one of the most common arguments in favor of such phenomenon: that our perceptual systems are finite, does not entail that phenomenal continua are possible, and even if phenomenal continua are possible, that does not itself entail that ‘looks the same as’ relations are non-transitive because it would not entail the homogeneity thesis. The homogeneity thesis is suspect because it seems to entail that we cannot have experiences as of constant motion. The intuitive plausibility that we do experience...
phenomenal continua can be accounted for by appealing to a looks/notices distinction which is compatible with the transitivity of ‘looks the same as’ relations. Therefore we have reason to favor the transitivity of ‘looks the same as’ over the possibility of phenomenal continuums of the sorites variety.

If Fara’s argument is successful then we have good reason to doubt that the medical device of Quinn is actually possible. If the situation Quinn has imagined is not possible then there is no reason to think that the puzzles these situations generate constitute genuine examples of rational intransitive preferences. I will conclude this section by addressing some of the influential objections to Fara’s argument.

**Objections to Fara**

There have been a number of objections to Fara’s arguments, some of which are worth addressing. I will begin with objections raised by Ian Phillips and Phillipe Chuard who both raise a similar worry for Fara’s account.41

Phillips argues that Fara makes a mistake in failing to draw a distinction between failing to notice a noticeable difference and there being no noticeable difference. Fara’s own example of failing to notice a friend’s haircut is an example of failing to notice a difference that is noticeable, however in the case of phenomenal continuums we typically take the differences in color between adjacent slices of a color spectrum to be differences that are not noticeable. Phillips argues that Fara’s account entails both that an unnoticeable object or event is perceived even though it is not noticeable, which is

---

controversial enough as it is, and that there is no limit to the visual discriminations we can make. Phillips argues “if this [Fara’s] argument convinces us that a change in position as small as $1/6^\circ$ is visible but not noticeable, we must also allow that changes of $1/6,000,000^\circ$ are visible but not noticeable.”

Philippe Chuard also argues that Fara’s argument entails that our powers of visual perception are infallible. Chuard first invites us to consider that our powers of perception are limited in the following way:

\[(\text{LCP})\] for any two coloured objects $\chi$ and $y$ where the chromatic differences between $\chi$ and $y$ is less than $L$ but more than zero, veridical visual experiences of $\chi$ and $y$ represent $\chi$ and $y$ in the same way with respect to their colour.\(^{42}\)

We are then asked to imagine three green objects $o_1$, $o_2$, and $o_3$ whose spectral values are $o_1 = 526-528$ nm, $o_2 = 525-527$ nm, and $o_3 = 524-526$ nm.\(^{43}\)

Chuard’s argument then proceeds as follows:

1. There are limitations $L$ on the way in which the human visual system represents colours such that, for any two coloured objects $\chi$ and $y$, where the chromatic differences between $\chi$ and $y$ is less than $L$ but more than zero, veridical visual experiences of $\chi$ and $y$ represent $\chi$ and $y$ in the same way with respect to their colour.
2. $\chi$ and $y$ look chromatically different to $S$ if and only if $S$'s experience of $\chi$ and $y$ (if it is veridical) represents $\chi$ and $y$ differently with respect to their colour.
3. Since the chromatic difference between $o_1$ and $o_2$ is below $L$, veridical experiences of $o_1$ and $o_2$ represent the colour of $o_1$ and the colour of $o_2$ in the same way.
4. Since the chromatic difference between $o_2$ and $o_3$ is below $L$, veridical experiences of $o_2$ and $o_3$ represent the colour of $o_2$ and the colour of $o_3$ in the same way.

\(^{42}\) Ibid. 179

\(^{43}\) The exact differences in value are not all that important for his argument so long as they are small enough to fall below $L$. 

43
5. Since the chromatic difference between o1 and o3 is above L, veridical experiences of o1 and o3 represent the colour of o1 and the colour of o3 different.
6. o1 and o2 look chromatically the same & o2 and o3 look chromatically the same & o1 and o3 look chromatically different.
7. (ntl) is true: there are coloured patches such that the relation of looking chromatically the same as between such patches is non-transitive.44

Chuard argues that if one seeks to defend the transitivity of looks the same as, the only plausible premise to deny is the first premise, however Chuard argues that denying the first premise entails that we have no perceptual limitations on wavelength discriminations. “The argument shows that looks-fallibilists are forced to deny the existence of perceptual limitations on wavelength discrimination.”45

Both Phillips and Chuard argue to the effect that if we take what Fara has argued to be true we are committed to denying that our powers of perception are limited in a certain respect even though they clearly do seem to be limited in that respect. I will now argue that Fara’s position does not imply these implausible conclusions.

The first point I wish to make in defending Fara is against an objection of Phillips: “Fara’s noticing-based account must insist that an object or event can be perceived even though it is unnoticeable. This is not something we naturally admit in other cases, and it is far from clear that it is coherent.”46 Against this I would first ask what is it that Fara is claiming we might not notice? Fara need not claim that we can perceive something that is unnoticeable. Fara is out to defend the transitivity of ‘looks the

44 Ibid. 181
45 Ibid. 190
same as’, so what is at issue here is our ability to notice the difference between the relevant aspects of phenomenal states. The question then is, is it possible for a subject to experience two phenomenal states that are different in some respect without being able to notice what the difference is? As I have already argued in my example of the Warlock and the self-torturer, these can come apart. We could further imagine a being who is like us but for the fact that the being has no memory whatsoever, including sensory memory. I take it that though the phenomenal states of that being would no doubt be different than our own, we could plausibly assume, that this being would still have phenomenal states, throughout the life of this being its phenomenal states would change, and the being would not be able to notice any difference between these phenomenal states; however, this is perfectly compatible with the transitivity of ‘looks the same as’.

What about the more substantive objection that Fara’s argument entails that we have perfect powers of discrimination? Clearly if the correct way to cash out our perceptual limitations is Chuard’s LCP, then the intransitivity of ‘looks the same as’ is a short step away as his argument demonstrates. Furthermore this argument is perfectly compatible with the distinction between two phenomenal states being different and not noticing this difference. Against his objection I will argue that Chuard’s principle is not the only to cash out our perceptual limitations, nor is it the best way.

To see why Phillips’ and Chuard’s objections fail it is important to review why Fara posits the looks/notices distinction in the first place. Fara is trying to give an account of our reputed experiences as of homogeneity between small enough regions of supposed phenomenal continua. The looks/notices argument is supposed to explain why it is that we might report that adjacent regions of a supposed continuum look the same even if they
do not look the same. The only reason this argument would entail that there is no limit to the visual discriminations we can make is if one held that if our powers of discrimination are finite, then sorites-type phenomenal continua are possible, but this is a claim Fara explicitly rejects when she introduces the perceptual limitation thesis. Fara argues persuasively that the perceptual limitation thesis does not entail that we experience phenomenal continua. Recall that the perceptual limitation thesis held:

**Perceptual Limitation Thesis:**
For some sufficiently slight amount of change (in color, sound, or position, etc.), we cannot perceive an object as having changed by less than that amount unless we perceive it as not having changed at all (as having changed by zero amount).

All that this thesis entails is that “when we experience an object as changing we experience it as changing by at least some given amount.”

One way we might cash out the perceptual limitations with respect to color perception consistent with Fara’s perceptual limitation thesis is as follows: First, the total number of distinct shades we are capable of perceptually representing is less than the number of possible wavelengths of light relevant for color. Second, this limitation is manifest in our perceptual experiences in the following way: under constant circumstances and instances of veridical perception, for each shade of color our visual system can potentially represent, there is a range of wavelengths such that objects reflecting light within that range will look to have the same shade of color. But this is

---


48 I specify constant circumstances and veridical experiences to obviate worries that how color is represented is sensitive not only to differences in the wavelength of reflected light, but also to other factors, like whether the subject is depressed or under the influence of drugs, and also to rule out cases of hallucination or illusions.
perfectly compatible with the denial of Chuard’s LCP and compatible with the transitivity of ‘looks the same as’. This entails that there are differences in wavelength that we cannot perceptually represent, not because the differences are too small, but because the wavelengths fall within the range of wavelengths that are all represented as one particular shade of color. This also entails the denial of LCP as whenever the wavelength of light changes from one range to another it is represented as being a different shade of color even if this difference is extremely small.

I take it that the picture I have sketched is possible and the only question that remains is what picture is in fact representative of our perceptual limitations. This is ultimately an empirical question. I don’t know for certain which picture is empirically correct, but I can offer the following argument in favor of my picture. My picture is compatible with (1): the plausible assumption that our powers of perception are limited. It is compatible with (2): the plausible assumption that we often seem to have experiences as of sorites style phenomenal continuums—because although there are small differences in the objects look to us, we don’t always notice this difference, hence we report that adjacent color patches look the same even though they actually don’t. Finally, it is compatible with (3) the truth of same appearance claims like “If A looks the same as B then if A looks red, so does B”. Alternatively, Chuard and Phillips’ position does no better in explaining why we take (1) and (2) to be true, and entails the denial of (3), so we ought to prefer my picture.

Fara has given a cogent argument that we do not in fact experience sorites style phenomenal continua and given the plausibility of Fara’s perceptual limitation thesis, neither Phillips’ nor Chuard’s objections were successful. Although the context of Fara’s
argument concerned our ability to perceive slight differences in color, the argument generalizes to our ability to feel pain. We therefore have good evidence to suppose that contrary to what Quinn has supposed, there could not be a series of adjacent settings such that there was no phenomenal difference between adjacent settings yet the first and last settings feel radically different. If such a situation is not possible, we need not worry that it poses a threat to the transitivity of preferences.

49 Alex Voorhoeve and Ken Binmore give a nice argument along these lines: “For consider a device that registers charge only in whole kilovolts. If we hooked this device to a machine that administered a current of varying voltage, started with the dial at 0 and kept increasing the charge by small increments, then at some point the device will change from registering ’0 kV to registering ’1 kV. This implies that there are no just-noticeable differences in the sense under discussion for this device, even though its capacities of discrimination are limited.” This is, in effect, what I have argued of our perception of color applied to the perception of electricity.
SECTION THREE

The second strategy against transitivity I will be considering has not received as much attention as Quinn’s puzzle yet it is potentially more problematic. This strategy elicits an intransitive preference ordering amongst options which are compared with respect to a plurality of values and considerations. In this section I consider a candidate counterexample to transitivity advanced by Larry Temkin in *Weighing Goods: Some Questions and Comments* which employs this strategy and argue it is unsuccessful. Temkin’s argument involves the comparison of three alternatives via multiple values and factors; however, because each alternative under consideration is sufficiently dissimilar from the others, relevant factors in comparing some alternatives do not apply when comparing others. From this Temkin argues an intransitive preference ordering can rationally result. After explain Temkin’s argument in detail I will argue that as it stands, it is not clear that a rational intransitive preference ordering will result from his example. I will argue that this ambiguity is due to the vagueness with which Temkin has presented his argument. I will first attempt to repair Temkin’s argument, applying some intuitive distinctions employed in the literature concerning incomparability to make

---

50 Temkin, Larry S. (1994). *Weighing goods: Some questions and comments*. *Philosophy and Public Affairs* 23 (4):350-380. Kenneth May has given a similar argument. Where Temkin argues that intransitivity results because some values or factors apply to some comparisons but not others, because the objects under consideration are sufficiently different, May argues that intransitivity could result from a kind of intrapersonal Condorcet paradox. While Temkin argues that his intransitivity is rational, May is concerned solely with a descriptive account of intransitivity and does not argue that the agent is rational; however William A. Edmundson has argued that the agent in May’s example could be rational in forming her intransitive preferences. I have not included this discussion here as I think Temkin’s argument is stronger and in any case the final reasons I give for rejecting Temkin’s argument will also cover Edmundson’s argument. Kenneth O. May, 1954, “Intransitivity, Utility, and the aggregation of preference patterns,” *Econometrica* 1-23. Edmundson, William A. (2009). *Pluralism, Intransitivity, Incoherence*. In Mark White (ed.), *Theoretical Foundations of Law and Economics*. Cambridge University Press
Temkin’s argument more precise. Finally, I will argue that even this more precise version of Temkin’s argument fails.

**Temkin’s Counterexample**

Temkin’s argument against transitivity begins with a plausible judgement:

> How two outcomes compare *all things considered* depends on how they compare in terms of the relevant factors for making that comparison. Thus, if *different* factors were relevant for comparing different outcomes, it could be true that even if A were better than B, and B better than C in terms of the relevant factors for comparing *those* alternatives, A might *not* be better than C, in terms of the relevant factors for making *that* comparison… For example, suppose that factors x and y are each relevant when comparing A,B,and C, but that another factor, z, is *only* relevant when comparing A and C. Then it might be that, all things considered, A is better than B (regarding x and Y), and B is better than C (regarding x and y), but C is better than A. This could be so if C was sufficiently better than A regarding z, to outweigh the extent to which it was worse than A regarding x and y. ⁵¹

Temkin provides an example of the kind of situation he is imagining. We are asked to consider three job applicants Sue, a white woman, Maria, a Hispanic woman, and Ella, a black woman. The person evaluating the job applicants is charged with picking the best candidate in light of their overall qualifications and an affirmative action policy. The person in charge of hiring reasons as follows. Sue is a better candidate than Maria, because of her qualifications and experience. Maria is a better candidate than Ella because of her qualifications and experience. However, Ella is a better candidate than Sue, because the affirmative action policy trumps their difference in qualifications and experience. An intransitive preference ordering results.

⁵¹ Ibid. 361-62 Once again Temkin is arguing for the intransitivity of “better than” but this generates an intransitive preference ordering so long as we are operating under the plausible principle that rationality requires that one prefers what they believe to be the better option.
The obvious place to balk seems the invocation of the affirmative action policy in comparing Sue and Ella. Temkin’s justification is that while blacks were not enslaved by Hispanics, they were once enslaved by whites, and Hispanics were neither enslaved by whites, nor by blacks. This justifies giving greater preference to blacks over whites than whites over Hispanics or blacks over Hispanics. Of course it is of little importance whether this is actually the reason for affirmative action policies in the real world or if the historical details are accurate, Temkin just needs a semi-plausible example. But is it semi-plausible? I have some doubts.

What I have argued so far is that preferences are rational in so far as the reasons for the preferences are rational. This affirmative action policy does not seem rational as there does not seem to be a good reason to only apply this policy when comparing a black person with a white person. If, as Temkin suggests, the justification is that black people should get preferential treatment because they have been oppressed in ways whites and Hispanics have not, why wouldn’t this apply when comparing blacks and Hispanics? If the reasoning is that whites should be punished for oppressing blacks, why wouldn’t this apply to when comparing whites and Hispanics? What is needed is a rational justification of an affirmative action policy that would only apply to the comparison of blacks and whites.\textsuperscript{52}

\textsuperscript{52} Someone might respond that this line of objection misses the forest for the trees. The point is that the person evaluating the job candidates has a rule which only “kicks in” under certain conditions and the following of this rule generates the intransitive preference ordering the justification for that rule is only minimally important. If this is how the argument is supposed to work then why not just make a rule that says, if A is better than B and B is better than C, then consider C to be better than A”? Surely such a rule doesn’t answer the question as to whether intransitive preferences really are rational.
Temkin has not yet offered a plausible counterexample to transitivity. Temkin’s argument is that an intransitive preference ordering could result when the relevant considerations for forming preferences differ according to which alternatives are being compared, but Temkin has not provided an example in which this is clearly the case, nor has Temkin provided independent philosophical motivations for thinking that this is true. However, Ruth Chang has provided motivation for thinking that the relevant considerations for comparing alternatives depend on the alternatives under consideration. In what follows I will borrow from Chang’s work on parity to construct a more precise counterexample to transitivity. After constructing this counterexample I will argue that this version also fails to undermine transitivity.

In her introduction to *Incommensurability, Incomparability, and Practical Reason* Ruth Chang argues that comparisons must proceed in terms of a ‘value’, where a value is “any consideration with respect to which a meaningful evaluative comparison must be made. Call such a consideration a ‘covering value’ of that comparison.”53 The idea is that one cannot compare two things simpliciter, they must be compared with respect to some value. For example, it is often said that one cannot compare apples and oranges, but according to Chang, all that is needed is an appropriate covering value and once one is in mind, apples and oranges are perfectly comparable. If we wish to compare apples and oranges with respect to pedagogical usefulness for teaching the color orange, oranges are

---

53 Chang, Ruth (ed.) (1997). *Incommensurability, Incomparability and Practical Reason*. Harvard University Press. 5 Here Chang uses 'covering value' though in her later publications on parity she uses 'covering consideration'. I will use both interchangeably.
better than apples. If we wish to know which is better for keeping away medical professionals, apples are better than oranges.

A covering value can have many ‘contributory values’. For example if I am trying to decide what hospital would be best for my family, the covering value ‘suitability for my family’ will have contributory values like professionalism of the staff, cleanliness, proximity to my home, cost, and others. Chang calls the degree to which an item does well according to its covering value its ‘merit’. Chang argues that many putative examples of incomparability are really just cases where no covering value has been specified.

Chang uses the assumption that all comparisons must proceed with respect a covering value to motivate a distinction between incomparability and noncomparability. Not all covering values apply to every pair of options. I can’t compare the number 9 and the number 7 with respect to gustatory pleasure, nor can I compare a cup of tea and a gingerbread cookie with respect to musical talent. In order for two options to be comparable with respect to a covering value, the covering value must ‘cover’ the options under consideration. Chang specifies that “for each two-place comparability predicate, there is a domain of pairs of items to which the predicate can apply.”

54 Chang argues that unless a covering value has both items under consideration in its domain then they are neither comparable nor incomparable with respect to that covering value. This is a formal

54 Ibid. 28
failure of comparison, and the items are *non*comparable with respect to that covering value.\textsuperscript{55}

Importantly for our purposes, for two options to be comparable with respect to a covering value, the covering value must cover both of the options. If the covering value covers one of the options, but doesn’t cover the other, the options are noncomparable with respect to that covering value.

Chang does not think noncomparability is a threat to practical reason:

Practical reason never confronts agents with comparisons that could formally fail…. Practical reason does not require us to compare noncomparables; as rational agents, we will never be confronted… with a choice between a lamp and a window for prime minister. Indeed no choice could ever have as its justification or its justifying force a comparison of the alternatives with respect to a value that does not cover them.\textsuperscript{56}

I am less optimistic that noncomparability is not a threat to practical reasoning. In fact I think we can build from Chang’s noncomparability an intransitive preference ordering in Temkin-fashion.

In order to generate my counterexample to transitivity, I need one more assumption in addition to what Chang has argued so far. My assumption is that if a covering value can fail to cover both items of a comparison, so can a contributing value of that cover value. With this we can generate our counterexample to transitivity.

\textsuperscript{55} Two options would be incomparable if two options are covered by the same covering value, but the trichotomy thesis fails. That is, the objects under consideration are neither better, nor worse, nor equal to each other. Michelangelo and Mozart are within the domain of ‘creative geniuses’ but it might not be possible to compare the two in terms of who is the better creative genius. Chang calls this a substantive failure of comparison.

\textsuperscript{56} Ibid. 29
**Temkin’s Counterexample Improved**

Tiffany is at home considering what to do with her Friday night when she fortuitously receives three emails. The first email is from Chuck, a good friend. Chuck is getting a group of six friends together this Friday for a taco party at Alonzo’s Taqueria and wants to know if Tiffany would like to come. The second email is from Brittany, an equally good friend, who has four tickets to go to the ballet this Friday and would like to invite Tiffany and some mutual friends. The third email is a congratulatory email from her work, Tiffany has won a ticket to a renowned food art festival in town this Friday night. Tiffany has three options to compare A, Alonzo’s, B, the ballet, and C, the congratulatory food art festival. The covering value with respect to which she will make this comparison is ‘better use of a Friday night.’ For Tiffany this covering value has three relevant contributory values: friendship, artistic beauty, and quality food.

Tiffany consider whether it would be a better use of a Friday night to go to Alonzo’s to spend time with friends or to go to the ballet with friends. Tiffany is evaluating these options in light of the three contributory values, food, friendship building, and art. She judges that she cannot compare A to B with respect to food, as there is no food at the ballet. Neither can she compare A to B with respect to art as Alonzo’s is not a place of art. She can compare A and B with respect to friendship, and because Alonzo’s has more friends than the ballet, she prefers A to B with respect to the covering value: better use of a Friday night.

Tiffany compares B to C. Tiffany cannot compare B to C with respect to friendship as the food art festival is a solitary endeavor. She cannot compare B to C with respect to food, as once again there is no food at the ballet. She can, however, compare B
to C with respect to art as they are both places of artistic expression. Tiffany judges B to be better a use of a Friday night than C as she prefers ballet as a form of art to food art.

Finally Tiffany compares C to A. She cannot compare C to A in terms of friendship as the food art festival is a solitary outing. She can’t compare C to A in terms of Art as Alonzo’s is not an establishment of Art. She can however compare C to A in terms of food. The food art festival will be home to some of the greatest edible art in the world, and Alonzo’s is good, but not spectacular. Therefore, Tiffany judges C to be a better use of a Friday night than A as it has better food than A.

Tiffany’s preferences are intransitive. She prefers A to B because of the contributory values that cover both A and B, A is better than B. She prefers B to C for the same reason, as with C and A. Her reasoning seems perfectly rational in light of Ruth Chang’s plausible assumption regarding the comparability of alternatives.

I take it that this counterexample is along the lines of what Temkin had in mind when he claimed that the factors relevant for comparison differ depending on the alternatives, and that this could generate an intransitive preference ordering. However, whereas Temkin’s argument was unclear as to why the relevant factors did not apply to certain comparisons, my method is not as I have utilized Chang’s plausible distinction between comparability and noncomparability. I take it that the counterexample I have proposed is stronger than Temkin’s, the question remains as to whether or not it is persuasive. I will now argue that it is not.

---

57 One might object that surely there are more contributory values than the three I have considered: financial cost, health, productivity, etc. Just assumed that while there are other contributory values that apply to all three, Tiffany is indifferent: she judges A, B and C to be equal in regards to all other contributory values.
The counterexample to transitivity I have been developing is not persuasive because it is based on a false premise. The false premise is that a covering value must cover *both* items for those items to be comparable with respect to that value. Instead I will argue that a covering value need only cover one of the two items for those items to be comparable with respect to that value. I will begin by trying to motivate this objection with a thought experiment.

**Improved Counterexample Rejected**

**Regrettable Rick:**

Imagine Rick has been tasked with picking up lunch for his boss. His boss has written him a note with clear instructions “go to the sandwich shop and order for me the best tasting thing available.” Rick, who is very obedient, goes to the sandwich shop and asks the attendant for their available options. The attendant responds that they only have two options today: the turkey club or the special, which is an mp3 file of Mozart’s greatest works. Rick becomes confused. He knows he is supposed to order the best tasting thing, but he finds Mozart’s greatest hits and the turkey club to be noncomparable with respect to taste as Mozart’s greatest hits does not fall within the domain of ‘better with respect to taste’. Rick returns to his manager empty handed having been unable to choose the best amongst noncomparable options. He is promptly fired.

I expect that Rick’s boss was right for firing Rick and that she was displeased for good reason: Rick should have ordered the turkey club. It was obviously the better choice, but where did Rick go wrong? Rick’s mistake was thinking that because Mozart’s music does not fall within the domain of ‘better with respect to taste’ that it is non comparable with the alternative that is within the domain of ‘better with respect to taste’. This is where I take Chang to be mistaken. I will now offer my alternative picture.
Items are comparable with respect to a covering value so long as one of the items falls within the domain of the covering value. When both items are within the relevant domain, items are compared in terms of merit—that is, they are compared in terms of which better satisfies the relevant covering value. When only one item falls within the domain of the relevant covering value, the alternatives are compared in terms of which satisfies the value at all (where satisfying the value means it has some positive merit relative to the covering consideration). If the item that is within the domain of the covering value satisfies that covering value, it is de facto better than the alternative with respect to that covering value. If it dissatisfies the value (negative merit), then the agent ought to be indifferent between the alternatives with respect to the covering value. If the covering value does not cover either of the alternative under consideration, the alternatives are noncomparable just as in Chang’s picture.

I will try to clarify my alternative picture in light of the Tiffany example. When Tiffany compares Alonzo’s with the ballet she is right to conclude that Alonzo’s does not fall within the domain of ‘better with respect to Art’ just as she is right to conclude that the ballet does not fall within the domain of ‘better with respect to food’. However, she is wrong to infer from this that A and B are noncomparable with respect to these contributory values. This is because, according to my picture, only one of the options needs to be covered by the covering value for the options to be comparable with respect to that covering value. Alonzo’s is covered by the value of ‘better in terms of food’ so Alonzo’s can be compared with the ballet even though the ballet is not.

---

58 This is, of course, a point of departure from Chang, however I think it is correct.
According to the picture I am advancing, when only one of the items is covered by the relevant covering value, the alternatives are compared in terms of whether the alternatives satisfy the value. ‘Satisfying the value’ means that the option has some positive merit in light of that value. Alonzo’s has good food and hence has positive merit in light of the covering value food. The ballet has no food and hence is not positive (though not negative) with respect to food. Hence Alonzo’s is better than the ballet in terms of food and, for the same reasons, the ballet is better than Alonzo’s in terms of art.

What if only one alternative is covered by the relevant value, but the covered option is negative in light of the relevant value? Suppose Rick had gone to the sandwich job just as before, but in this scenario the options were either the Mp3 file or a two week old meatball marinara sub which has begun to turn a shade of green. The meatball marinara does fall within the domain of tastiness, while the Mp3 file does not. According to the picture I am proposing they are comparable with respect to tastiness. When the covered item is negative in light of that value the option equal to the non-positive alternative. That is, in the meatball version of Regrettable Rick Rick ought to be indifferent between the moldy meatball sandwich and the Mp3 file as neither satisfy the covering value.  

In light of these considerations we can dismantle my improved counterexample to transitivity. Recall that the options under consideration are tacos with friends at Alonzo’s,

---

59 I could be wrong about this. I could see ranking the meatball sandwich as being judged less tasty than the Mp3 file for the reason that while the Mp3 file does not satisfy the value, because it has no taste, the meatball sandwich dissatisfies the value. This is interesting and worth considering, but doesn’t affect the strength of my argument.
ballet with fewer friends, and food art festival with no friends. Under the assumption that a covering value had to cover both items for those items to be comparable, Tiffany generated an intransitive preference ordering. Under my alternative picture, however, no intransitivity results. Consider the following table representing the merit of each alternative with respect to the relevant covering value:

<table>
<thead>
<tr>
<th></th>
<th>Friendship</th>
<th>Art</th>
<th>Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alonzo’s</td>
<td>2</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Ballet</td>
<td>1</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Food Art</td>
<td>-</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Because Alonzo’s involves more friends than the ballet it has more merit than the ballet with respect to friendship. The food art is not covered by the covering value of friendship so it is no positive value with respect to friendship but is not negative either. The ballet is better in terms of art than the food art festival and hence has one more merit than the food art festival. Finally the Food art festival is better than Alonzo’s in terms of food and received one more merit than Alonzo’s. The result is that each option is better in terms of one value, worse, but positive, in terms of another, and not positive in terms of another. Assuming Tiffany values friendship, art, and food equally, that is to say that neither is more important than the other, then Tiffany ought to be indifferent amongst the alternatives. If one value is more important than the others than the option with the most merits should be preferred.

60 Obviously the values I have put in to the table are arbitrarily chosen. However the main point is that the options with the most merits are to be preferred, and as ‘greater than’ is necessarily transitive, the resulting preference ordering will be transitive.
merit with respect to that value will be the most preferred, and the option with the least positive merit will be least preferred. Either way, Transitivity is preserved.

I don’t imagine that I have offered the final word on counterexamples to transitivity that adopt Temkin’s strategy; however, the arguments I have given at least show that it is far from obvious that there is a genuine counterexample to transitivity to be derived from pluralism about value. Therefore the burden of proof is with those who wish to reject transitivity.
CONCLUSION

This thesis set out to accomplish two main objectives. The first was to show that there is a genuine debate to be had concerning transitivity of preferences. Two of the most prominent arguments for transitivity were considered and found wanting. While these arguments may suffice to put the burden of proof on those who think intransitive preferences can be rational, they are far from the final word on the topic. In light of this I considered two putative counterexamples to transitivity which employ two very different strategies. The first counterexample I considered was Quinn’s puzzle to the self-torturer. I considered two interpretations of this puzzle and argued that neither posed a genuine counter example to transitivity. I then considered a counterexample from Larry Temkin which argued that intransitive preference could result when the options under consideration were sufficiently different that the relevant factors for comparing some of the options were irrelevant in comparing others. I considered one reason that this could be true and ultimately rejected that it could generate an intransitive preference ordering.

The implication of my arguments is that we should continue to think that transitivity is a rational constraint of preferences, but it is important to note that I have far from proven that rational preferences are necessarily transitive. Furthermore there are other counterexamples to transitivity that I have not considered here. My final conclusion is that while there is some good reasons for thinking that transitivity is a constraint on preferences, it is ultimately a live question worthy of further philosophical debate.
REFERENCES


