Preference Cycles and the Requirements of Instrumental Rationality

April 27, 2020

Abstract

Many decision theorists take standard normative decision theory to be a theory of instrumental rationality, and its core requirements to be requirements of instrumental rationality. This paper argues that at least one such core requirement, namely the acyclicity of preference, cannot be defended as a general requirement of instrumental rationality. The standard instrumentalist defence of the requirement to have acyclical preferences, the Money Pump Argument, relies on a fatal equivocation about the standard of instrumental rationality. If the fundamental conative attitudes against which our actions are evaluated are attitudes to the fully described outcomes of our actions, we cannot show that being money-pumped is instrumentally irrational. If, however, the standard of instrumental rationality is attitudes to features of the outcomes of our actions, we cannot show that instrumentally rational agents must adopt acyclical preferences to avoid being money-pumped.

1 Introduction

Instrumental rationality requires agents to take the best means to the ends they desire, but is silent on what ends agents ought to have. Many decision theorists assume that standard normative decision theory is concerned with instrumental rationality alone. For the most part, this instrumentalist, or 'Humean' interpretation of decision theory is so entrenched, it is rarely stated or explicitly argued for.[1]

To defend standard decision theory as a theory of instrumental rationality, we need to show that its requirements are requirements of instrumental rationality. Some requirements of standard decision theory — such as the requirement that agents should maximize with regard to their preferences — have appeared to many to be obvious requirements of instrumental rationality. Others, such as the requirement to have acyclical preferences, have been defended as requirements of instrumental rationality by appealing to various instrumentalist arguments. What these arguments typically have in common is that they point out that agents who violate those requirements are prone to making a sure loss in some choice scenarios. Importantly, those arguments typically take the intuitive plausibility of maximization for granted. According to the Money Pump Argument, which will be the main focus of this paper, agents with preferences that form a strict cycle can be offered series of choices where they will end up paying for something they could have had for free, if they maximize with regard to their binary preferences at each point in time. Since being money-pumped in this way is taken to be instrumentally irrational, it is then argued that instrumental rationality requires agents to have acyclical preferences.

This paper aims to establish that at the heart of this instrumentalist defence of acyclicity lies an equivocation about what I will call the the *standard of instrumental rationality*. Before we can even address the question of instrumentalist justification of core requirements of decision theory, we must answer a question of interpretation. Most importantly, which of the agent's attitudes are taken to pick out her ends, and how do these relate to the elements of standard decision theory? This question is important, because those attitudes will serve as the standard against which the agent's choices and other attitudes will be evaluated.

It is a pervasive assumption in philosophical decision theory that the agent's preferences

[1]The position that instrumental rationality is all there is to practical rationality is often described as a Humean notion of rationality, following Williams (1979), who traces it back to Hume's (1739) *Treatise of Human Nature*. While this instrumental notion of practical rationality has been popular, it is not uncontroversial that Hume himself actually held it. See Hampton (1995) for arguments that he did not. In any case, Humeanism about decision theory is weaker, since it is silent on whether there are non-instrumental rationality requirements that are beyond the scope of decision theory. For an explicit statement of this instrumentalist interpretation of decision theory, see, e.g., the opening of Joyce's (1999) *The Foundations of Causal Decision Theory*, p.9. Lewis (1988) defends Humeanism about decision theory by arguing that one major form of non-Humeanism is not compatible with decision theory.

over the objects of choice form this standard of instrumental rationality against which the agent's actions and other attitudes are evaluated. In the case of choice under certainty, we can consider those objects of choice to be outcomes. The requirement that our choices ought to be guided by our preferences over outcomes, as captured, for instance, by the standard requirement to maximize, is plausible according to this outcome-based notion of instrumental rationality. However, I will argue that the Money Pump Argument fails according to this standard, because we cannot establish that being money-pumped is instrumentally irrational. The intuition that it is instrumentally irrational to be money-pumped in fact relies on a different understanding of the standard of instrumental rationality, one that relates to underlying attitudes to *features* of outcomes. But according to this feature-based standard, the requirement to be guided by one's preferences in action is no longer justifiable in the context of the Money Pump Argument.

I thus show that the Money Pump Argument fails according to both ways of understanding the standard of instrumental rationality. If attitudes to outcomes are the standard of instrumental rationality, being money-pumped can't be shown to be instrumentally irrational. If attitudes to features of outcomes are the standard of instrumental rationality, agents need not adopt acyclical preferences to avoid being money-pumped. They can simply act against their preferences at the right points in time. Either way, the instrumentalist case in favour of *acyclicity* fails. Still, I will argue that the second way of understanding the standard of instrumental rationality is in fact more plausible. Moreover, I will conclude by showing that the feature-based standard of instrumental rationality also makes it possible, at least, to provide an instrumental justification for acyclicity that is conditional on the agent having certain desires. This, I argue, is the best instrumentalist case we can make for acyclicity.

2 Preference Cycles

According to our ordinary understanding of preference, a preference is a relational conative attitude. When I say that I prefer one shade of yellow to another, or one brand of cereal to all others, that is usually taken to mean that I like that shade of yellow or that brand of cereal more than the other(s), or that I have a stronger desire for it. Most philosophical decision theorists follow this common usage,[2] and this is the sense of preference I want to start out with. Standard decision theory moreover assumes that preferences are *binary* relations that hold between exactly two objects. Here I am concerned only with decision theory in the context of certainty, where each possible action an agent might take is associated with exactly one outcome. We can thus think of her as directly choosing between outcomes.

[2]See Sobel (1997) for evidence in support of this claim.

Outcomes, in turn, are usually taken to be complete descriptions of anything the agent cares about in the circumstances the action brings about.

Preference cycles over outcomes are both wide-spread, and often appear sensible.[3] Two types of cases are commonly used in order to motivate the idea that cyclical preferences may sometimes not be irrational, or may even be called for. In one kind of case, the possible outcomes of the different actions available to the agent differ in various different dimensions the agent cares about. And in the other kind of case, the outcomes of the actions available to the agent are in some respects seemingly indistinguishable to her. To start with the first kind of case, suppose I am looking for an apartment. Three apartments are available for the same rent, which I can afford. They differ only in terms of their size, their views, and the length of the commute I would have if I lived in the apartment. All three of these are factors that I care about.

Apartment A: 40 m^2 large; view onto a garden; 5 minute commute.

Apartment B: 70 m^2 large; view onto the skyline, lake and woods; 60 minute commute.

Apartment C: 100 m^2 large; view onto the brick wall of the building next door; 30 minute commute.

When it comes to choosing where to live, my pair-wise preferences over the outcomes of living in each of these apartments (denoted by 'Apartment A', 'Apartment B', 'Apartment C') may well be cyclical. Suppose I have the following preferences over outcomes, where $\prec$ represents strict preference:

$$\text{Apartment A} \prec \text{Apartment B} \prec \text{Apartment C} \prec \text{Apartment A}$$

What may make these preferences seem defensible is that I can explain them in the following way: I prefer Apartment B over Apartment A because Apartment B is larger and has such a lovely view, and this outweighs the fact that it has a longer commute. I prefer Apartment C over Apartment B because Apartment C is even larger, and has a shorter commute, and this outweighs the fact that it does not have a good view. And I prefer Apartment A over Apartment C, because it has an even shorter commute, and a better view, and this outweighs the fact that it is smaller.

The second kind of case is best illustrated with the Puzzle of the Self-Torturer, first introduced by Quinn (1990). Suppose an evil scientist straps a device to your arm that

[3]For instance, González-Vallejo (2002) discusses evidence of cyclical preferences in the context of choices that are similar along some dimension. Korhonen et al. (1990) provide evidence of cyclicality in the context of multi-criteria decision-making.

causes you pain with electric shocks. The device has 1,000 different settings. At the first setting, it causes you no pain. At the highest, the pain is excruciating. However, adjacent settings differ so little in their electric current that you cannot distinguish them by the pain you feel when experienced subsequently.[4] Now suppose you are offered $10,000 in exchange for each setting you are willing to go up. And so each setting of the device is associated with an amount of money. Let $S_1, S_2, S_3, ...S_{1000}$ be the outcomes of ending up with the level of pain and amount of money associated with the 1,000 different settings. Now it seems reasonable to have the following, cyclical preferences over these outcomes:

$$S_1 \prec S_2 \prec S_3 \prec ... \prec S_{1000} \prec S_1$$

Out of two adjacent settings, you always prefer the higher one. After all, you cannot detect a difference in pain between them when experienced subsequently, and $10,000 is a substantial amount of money. However, when you consider the highest setting, you find the amount of pain so unbearable that you would gladly forego the fortune associated with it in order to be pain-free at the lowest setting.

According to all well-known decision theories, the preferences we described in both of these cases are irrational, since they violate the requirement of *acyclicity*. Let X be the set of outcomes the agent's actions may bring about. Let $\succcurlyeq$ represent weak preference between outcomes: $x \succcurlyeq y$ if and only if the agent either strictly prefers x to y, or is indifferent between x and y. *Acyclicity* requires the following:

Acyclicity: For all x_1, x_2, ... $x_n \in X$, $x_1 \succ x_2$, $x_2 \succ x_3$, ... $x_{n-1} \succ x_n$ implies that $x_1 \succcurlyeq x_n$.

What I want to investigate in the following is whether we can provide an instrumental justification for *acyclicity*, as it seems we would need to, were we to defend standard decision theory as a theory of instrumental rationality alone.

[4]Voorhoeve and Binmore (2006) and Arntzenius and McCarthy (1997) argue that subsequent settings cannot all be indistinguishable to the agent all things considered. All that matters for us, however, is that even if they are right, the self-torturer's preferences seem intuitively reasonable. This could be because two adjacent outcomes are still subjectively indistinguishable when directly compared, or because any difference in pain will be expected to be tiny. And so, even if these authors are right, we are in need of an argument for why the self-torturer should not have cyclical preferences. Moreover, as Aldred (2007) shows, similar preference structures can arise when preferences are dependent on the application of vague predicates, the permissibility of which is arguably less controversial.

3 The Money Pump Argument

On the face of it, *acyclicity* looks like a non-instrumental requirement of rationality. It is a requirement on what kinds of preferences an agent may hold. But instrumental rationality was supposed to be silent on what ends an agent may have. In fact, Hampton (1994) rejects the Humean interpretation of standard decision theory on that basis, and Dreier (1996) extends the Humean interpretation to include requirements on the structure of an agent's preferences. There are, however, instrumentalist defences of various standard requirements on an agent's preferences. These typically point out that an agent with the preferences in question is prone to making a sure loss. In the case of *acyclicity*, this instrumentalist defence comes in the form of the Money Pump Argument.

The Money Pump Argument was first formulated by Davidson et al. (1955), but goes back to ideas in Ramsey (1928/1950). We can apply it to our examples. To start with the first, suppose a rental agency gives me the chance to choose between Apartment A and Apartment B. Choosing according to my preference between these two apartments, I go with Apartment B. But then the agency offers me the opportunity to switch to Apartment C instead. Choosing in accordance with my preference between Apartments B and C, I choose C. Now suppose I get offered the chance to switch to Apartment A, in exchange for a small fee, say \$25. In my current circumstances, I prefer more money to less, other things being equal. But seeing that I have a strict preference for A over C, I probably still prefer A, even when I have to pay \$25 for it. If not, there will be a small enough positive amount of money ϵ that I will be willing to pay. My preferences are thus:

$$\text{Apartment A} \prec \text{Apartment B} \prec \text{Apartment C} \prec \text{Apartment A - } \epsilon$$

If I choose in accordance with my binary preferences at every point in time, I will end up with Apartment A having lost \$25. But I could have had Apartment A without losing that money, if I had only chosen Apartment A right away, and refused further trades. Ending up with Apartment A having payed \$25 seems instrumentally criticizable. Moreover, the rental agency could potentially repeat offering me this series of swaps, effectively turning me into a 'money pump'.

The same fate could meet you in the second example. Suppose you are offered the chance to go up by one setting every week, in exchange for the \$10,000. Going with your binary preference between two adjacent settings, you should always go up by one setting, all the way to the highest setting. This way, you would turn yourself into the eponymous 'self-torturer'. Now suppose somebody offers you the chance to go back to the lowest setting, in exchange for giving up your entire fortune, plus an additional \$25 (or a small enough amount of money ϵ). You gladly accept. But again, you could have been pain-free for less money,

making you apparently instrumentally criticizable. Moreover, the cycle could be repeated, turning the self-torturer into a money pump.

Note that, in the examples as we have described them, foresight could help our agents. Debates in dynamic choice theory have shown, however, that foresight doesn't always help. Schick (1986) originally proposed that if an agent can foresee she is going to be offered the series of trades we just described, she will stop trading early on. He argued that rational agents will use a process of backward induction, whereby they consider how they will choose in the last choice, assuming they will act in accordance with their binary preference over the two outcomes available then. They then take their prediction as given when considering the second to last choice, and so on. This approach to dynamic choice is an instance of a more general choice rule that has come to be known as 'sophisticated choice'.[5] It consists in the continued application of a norm to be guided by one's preferences over the options available at the time of action, combined with an expectation of future abidance by the norm. Rabinowicz (2000) shows that sophisticated agents can still be money-pumped, however. All we need is persistence on the side of the money pumper, such that agents are offered further trades even after they have refused one.

Susceptibility to being money-pumped is widely held to provide a good justification for the requirement to have acyclical preferences. The goal of instrumentalist arguments for some purported rational requirement is to show that agents who violate that requirement can't be the kinds of agents who take the best means to their ends. Accordingly, one rough way of cashing out the Money Pump Argument is as follows.

P1 Agents with cyclical preferences can be placed in, or find themselves in situations where they can't rationally avoid being money-pumped while retaining their cyclical preferences.

P2 Being money-pumped is an instance of instrumental irrationality: Agents who end up money-pumped are not serving their ends well.

C Therefore, agents with cyclical preferences can't be instrumentally rational. *Acyclicity* is a requirement of instrumental rationality.

In addition, it is often taken to follow that, insofar as their preferences are under their control, the Money Pump Argument provides agents with cyclical preferences with a reason, grounded in their ends, to adopt acyclical preferences instead.

As stated, it is unclear why the conclusion should follow, given that some agents may never face the kinds of situations where there is a threat of being money-pumped. And

[5]See McClennen (1990) for a formal treatment.

presumably, to some extent, agents can actively avoid facing such situations. Defenders of the Money Pump Argument could either respond that what matters in establishing instrumental irrationality is just pointing out that agents with cyclical preferences are thereby ill equipped for some hypothetical choice scenarios, and that this alone suffices in establishing a rational deficiency in their preference stucture. Or they could point out that it is never entirely in an agent's control whether they face a potential money pump. Thus, adopting acyclical preferences is the only way of *insuring against* the instrumental irrationality involved in being money-pumped, and one might think this is enough to establish a rational requirement.

The following will not depend on how one fills in the details here. Instead, I am concerned with P1 and P2, which I take it will feature in any plausible reconstruction of the Money Pump Argument. What I wish to argue in this paper is that there is no conception of the standard of instrumental rationality that allows us to establish both.

4 Outcome-Based Instrumental Rationality and the Money Pump Argument

At this point, let us return to the fundamental question of interpretation the instrumentalist about decision theory faces, which I posed in the introduction: Which of the agent's attitudes are taken to pick out her ends, and how do these relate to the elements of standard decision theory? That is, which attitudes will serve as the standard of instrumental rationality? The standard response appears to be that this role is played by the preferences over outcomes that decision theory ascribes to agents. In the following, I will refer to this conception of the standard of instrumental rationality as 'outcome-based instrumental rationality'. Outcome-based instrumental rationality appears to be implicitly assumed, in the context of certainty, by most decision theorists, at least within philosophy.[6]

In one respect, outcome-based instrumental rationality seems to help us make the Money Pump Argument. It provides the best explanation of why instrumental rationality should require agents to act in accordance with their preferences in binary choice contexts. As we have seen, the Money Pump Argument relies on agents choosing in such a way. In fact, outcome-based instrumental rationality seems to justify a more general requirement of *preference-guidance*, a requirement to fulfil, and to avoid frustrating one's preferences as much as possible in each of our choices. If instrumental rationality is about doing well by our preferences over outcomes, the best way of doing so would seem to be to act in accordance with those preferences in this sense.

[6]See [redacted] for evidence for this claim.

Ultimately, however, the Money Pump Argument fails under outcome-based instrumental rationality, because proponents of outcome-based instrumental rationality cannot establish the truth of P2. Arguably the most common way of defending P2 is as follows: Agents who are money-pumped fail to maximize with regard to their preferences. But maximization is a requirement of instrumental rationality, which agents who are money-pumped thus violate. One way of cashing out this requirement is as follows:

Maximization: Agents ought to choose an outcome such that no other available outcome is strictly preferred to it.

Maximization is indeed a standard requirement of orthodox decision theory, and one way of spelling out the more general requirement of *preference-guidance*. But first, note that the agents in our money pump scenarios do not actually violate a requirement to maximize on any individual choice they face. As we said, they always choose in accordance with their binary strict preference between the two options they are facing. Money-pumped agents do fail to maximize *over time*, however, and thus only violate a diachronic version of the alleged requirement to maximize. Gustafsson (2013) notes that the diachronic version of *maximization* may not be as uncontroversial as the synchronic version. However, he shows that we can give an agent with cyclical preferences a single choice that will force her to violate *maximization* synchronically. Namely, we can confront her with a single choice between all the outcomes over which she has cyclical preferences. No matter what she chooses, she will end up with an outcome to which another available outcome would have been preferred. Thus, she is bound to violate *maximization*. Gustafsson claims that this shows that the cyclical preferences are irrational. And the argument that cyclical preferences are irrational because they make it impossible for agents to maximize can also be found in earlier literature. Levi (2002) makes the same argument, and so do Davidson et al. (1955), just before they present their Money Pump Argument.

Does the failure to maximize explain the irrationality of being money-pumped? Recently, Andreou (2016) has compellingly argued that on this way of justifying P2, the Money Pump Argument is question-begging. The problem is that those who don't already believe that cyclical preferences are irrational may simply respond that maximization is not a plausible choice rule for agents with cyclical preferences, given it is impossible for them to follow the rule. This turns out to be exactly the right response within the framework of outcome-based rationality if we don't have prior reasons to take cyclical preferences to be irrational.

What Gustafsson, Levi and Davidson et al. seem to ignore is that the requirement to maximize is itself a principle that needs to be instrumentally justified, if decision theory is supposed to be a theory of instrumental rationality. It needs to be shown that agents need to abide by it in order to serve their ends well. Under outcome-based instrumental rationality, the agent's ends are picked out by her preferences over outcomes. When the

agent's preferences are cyclical, the question is thus what it would take for the agent to serve her cyclical preferences well. Unlike in the case of agents within acyclical preferences, the requirement to maximize does not qualify as a good general principle of choice for such agents, because it may be impossible for those agents to maximize. But that need not mean that instrumentally rational choice is impossible for them.

In fact, choice rules that extend to cyclical preferences have been proposed that seem to capture *preference-guidance* for agents with cyclical preferences. For instance, according to a rule proposed in Schwartz (1972), an agent should choose a member of a subset of the available outcomes such that (i) no outcome outside of the subset is strictly preferred to any member of the subset, and (ii) no proper subset of this subset fulfils condition (i). In our examples, if the agent is given a single choice between the outcomes over which she has cyclical preferences, according to this rule she is permitted to choose any of the options. At the same time, in each of the binary choices, the agent is required to choose the outcome she prefers. Schwartz's rule is still in the spirit of outcome-based instrumental rationality. An agent with cyclical preferences may have to frustrate some of her preferences. But the rule identifies a set of outcomes that seems to ensure no preferences are frustrated unnecessarily. Her preferences can in this way still serve as the standard of instrumental rationality. Given the availability of plausible alternative choice rules for cyclical preferences, the violation of maximization can't be what makes being money-pumped irrational.

There is another way in which one could argue in favour of P2. Rather than argue that agents with cyclical preferences violate some other requirement of instrumental rationality, namely maximization, one could try to establish directly that, when they are money-pumped, agents with cyclical preferences are not serving their ends well. This is suggested, for instance, in McClennen's (1990) explanation of what is supposedly instrumentally irrational about being money-pumped: "[A] principle of choice is valid if failure to adhere to it would result in choice of means insufficient to desired ends – in the agent pursuing his objectives less effectively than he could have under the circumstances in question." (p.4)[7] What is implied on this way of understanding the irrationality of being money-pumped is that, if the agent did not have cyclical preferences, then her objectives would be better served. On this construal, the Money Pump Argument could be thought of as providing the agent with reasons, grounded in her ends, to adopt acyclical preferences. Note that on the interpretation of the argument we previously considered, the Money Pump Argument does not imply this. On that interpretation, the problem was simply that agents who have cyclical preferences fail to maximize. If maximization is itself required by instrumental rationality, then the instrumental irrationality of being money-pumped just consists in violation of that principle, whether the agent's ends permit of being better served or not.

[7]Note, however, that McClennen switches back and fourth between this, and the previous way of understanding the argument.

How might this alternative way of justifying P2 work within outcome-based instrumental rationality? According to outcome-based instrumental rationality, instrumental rationality consists in doing well by one's preferences over outcomes. *Acyclicity* is a principle about what preferences an agent may hold. Essentially, what the argument would now need to establish is that having different preferences would serve your preferences better. When could having different preferences over outcomes serve your preferences over outcomes better? This appears to be only so if having different preferences comes with autonomous benefits, in terms of the preferences you currently hold. For instance, if an evil demon were to severely punish you for continuing to have the preferences you currently have, and you want to avoid such punishment at all cost, then it serves your preferences as they are now to have different preferences.

Is susceptibility to being money-pumped like this? To say so, we would have to show that being money-pumped is bad in terms of the agent's cyclical preferences over outcomes. That is, we have to show that having cyclical preferences leads to an outcome that is bad in terms of those cyclical preferences, and that adopting different, acyclical preferences would lead to a better outcome according to those cyclical preferences. The alternative justification of P2 fails within the outcome-based framework because outcome-based instrumental rationality does not allow us to say so. First, many critics of cyclical preferences claim that cyclical preferences mean that there is no outcome that it would be rational for the agent to choose precisely because cyclical preferences do not pick out any outcome as 'best'. In fact this is exactly what the first interpretation of the argument, which we just considered, relies on. But if we believe that, it would also seem like those preferences can not act as a standard of what alternative preference relation may serve the agent better.

Second, note that the Money Pump Argument exploits the following fact about the preferences of agents with cyclical preferences who prefer more money to less. If an agent displays a preference cycle over some outcomes, then she also has cyclical preferences over a set of outcomes that includes one of the original outcomes with some small amount of money deducted. In the apartment example, if I have these cyclical preferences:

$$\text{Apartment A} \prec \text{Apartment B} \prec \text{Apartment C} \prec \text{Apartment A}$$

I also have these cyclical preferences:

$$\text{Apartment A} \prec \text{Apartment B} \prec \text{Apartment C} \prec \text{Apartment A - } \epsilon \prec \text{Apartment A}$$

Intuitively, if preferences over outcomes are all we can go by, when I am offered a choice between A, B and C, I am permitted to choose any. If there is any outcome that it would be rational in terms of my cyclical preferences to end up with, then each of these choices should

be permitted. If we want to use only facts about my preferences over the outcomes available to determine which outcomes it would be rational for me to end up with, as outcome-based instrumental rationality demands, then we cannot treat outcomes A, B, and C differently.

At the same time, to make the instrumentalist argument we are considering here, it needs to be instrumentally irrational, in terms of the agent's cyclical preferences, to end up with A - ϵ. There thus needs to be a difference between the cyclical preferences over A, B, and C, and the set of cyclical preferences over A, B, C and A - ϵ. Intuitively, we in fact see the preference of A over A - ϵ as one that may not rationally be frustrated, in contrast to the other preferences in the cycle. It is this intuition that the Money Pump Argument relies on. But outcome-based instrumental rationality cannot accommodate this intuition, at least not in the general terms in which we hold it. According to outcome-based instrumental rationality, my preferences over outcomes are basic, and my actions are judged by how well they serve my preferences. A, B, C and A - ϵ in our example are all shorthand for different outcomes that involve me having some apartment and a particular amount of money. If all we can go by in judging the instrumental rationality of an action are preferences over outcomes, we cannot treat our preference of A over A - ϵ as different in kind from our preference of A over C, such that the former can never be rationally frustrated, while the latter sometimes can. For instance, Schwartz's rule, which is formulated entirely in terms of preferences over outcomes, does not make this distinction, and does not rule out choosing A - ϵ out of the set of A, B, C, and A - ϵ.[8] The best explanation of why the preference of A

[8]There are some criteria appealing, at least explicitly, only to preferences over outcomes that one could use to rule out A - ϵ but not A, B, and C in many money pump scenarios, once we take into account plausible additional preferences. For instance, plausibly, A - ϵ is "covered" by A in the sense first introduced by Miller (1980) in the context of tournament theory: A is preferred to it, and it ranks no higher with regard to any other available option. This would be so if B is preferred to A - ϵ. It has been proposed that rational choice in the context of cyclical preferences rules out choosing a covered option [redacted]. However, ultimately it is not clear whether this criterion can be given a purely outcome-based rationale. And moreover, it cannot always rule out agents being money-pumped, while intuitively, being money-pumped is always irrational. To start with worries that the criterion cannot be given an outcome-based rationale, imagine we replace A - ϵ with some other outcome D that stands in no special relation to A — it is simply a fourth apartment. Yet, the preference relations are the same as just described, and D is covered by A. On an outcome-based view, why should D be ruled out? After all, it is preferred to another option, C, which is not ruled out. Moreover, ending up with D involves frustrating precisely two preferences among the available options, which is no more than the number of preferences frustrated by ending up with C, which is not covered and hence allowed by Miller's rule. And so the outcome-based rationale for Miller's rule is at least not clear. More importantly, there are potential cases of agents being money-pumped that Miller's criterion does not rule out, and outcome-based instrumental rationality cannot rule out for any other reasons, but that nevertheless strike us as irrational. In our case, it is intuitively irrational to end up with A - ϵ even if the agent's preferences are such that A - ϵ is not covered. For instance, this would be so if A - ϵ was preferred to B. Now one might say that given we have assumed ϵ is something the agent values, and she prefers B to A, it would be irrational for her to prefer A - ϵ to B. But this would be appealing to what, below, I will call the feature-based account of instrumental rationality, as it is ultimately information beyond outcome-preferences that makes it so plausible that the agent would prefer B to A - ϵ, given she prefers B to A.The fundamental problem is that the sense in which ϵ is an unambiguous loss for the agent relates to the features of the options involved, and can only be made directly rationally relevant with feature-based standards, as they are introduced below.

over A - ϵ is one that may not be rationally frustrated in fact comes from the feature-based approach introduced in the next section.

I thus conclude that the proponent of outcome-based instrumental rationality cannot establish P2, that is, the instrumental irrationality of being money-pumped. Nevertheless, something does indeed seem to go wrong if an agent is money-pumped: An agent ends up paying to get something that she could have had for free. And so while we have shown that outcome-based instrumental rationality cannot establish the irrationality of being money-pumped, we have not done away with the intuitive irrationality of being money-pumped. The next section will argue in favour of an alternative conception of instrumental rationality that can account for the intuitive irrationality of being money-pumped. Ultimately, however, I will argue that the Money Pump Argument still fails to provide us with an unconditional justification for *acyclicity* according to this alternative account.

5 Feature-Based Instrumental Rationality and the Cost of Being Money-Pumped

We have seen that outcome-based instrumental rationality cannot establish P2, despite the intuitive irrationality of being money-pumped. Outcome-based instrumental rationality is in fact implausible on independent grounds. I here want to motivate an alternative, before showing that while it can establish P2, it fails to support P1.

Outcome-based instrumental rationality takes preferences over outcomes to form the standard against which actions are judged. Outcomes, in turn, are descriptions of all the circumstances an action may lead to that the agent may care about. In our first example, at a minimum, the outcome Apartment A would consist in a description of the size of the apartment, the length of my daily commute, and the views from the apartment. In the self-torturer case, the outcomes would include descriptions of both the level of pain I will feel, and the amount of money I will have. One might think that appealing to 'what the agent may care about' here already presupposes a notion of caring about something other than outcomes. Yet, decision theorists have implemented this idea in a way that again appeals to only preferences over outcomes. Joyce (1999, p.52) cashes out the rule for specifying outcomes as follows: Whenever there is some circumstance such that an agent would strictly prefer an outcome in the presence of that circumstance to the same outcome in the absence of that circumstance, the outcome has been underspecified. Clearly, this rule for specifying outcomes will lead to outcomes being very detailed descriptions of states of affairs that may come about as the result of my actions. In fact, moving to more detailed descriptions of outcomes is a common move made in order to accommodate apparent violations of the

standard axioms of decision theory, including *acyclicity*. And so those who want to defend *acyclicity* need to embrace very fine-grained outcomes as the object of choice. But it is the fine-grained nature of these outcomes that makes them implausible candidates for the object of the conative attitude that should form the standard of instrumental rationality.

Let me first highlight that we do not ordinarily think of such detailed descriptions of states of affairs as the objects of our desires, even if they are the object of choice under certainty. What we claim to desire in ordinary discourse are simpler states of affairs. For instance, I might say, 'I desire to drink a glass of fizzy water right now', 'I desire to have more time to practise viola', or 'I desire to sail the Inside Passage'. In the last case, the object of my desire is not a complete description of one course that my life could take in which I sail the Inside Passage - complete with the description of what flavour ice-cream I will have after dinner tonight. The object of my desire is simply my sailing the Inside Passage. We also often speak of preferences with regard to such simpler states of affairs. I may say that I prefer sailing the Inside Passage to sailing the Northwest Passage, for instance. The objects of this preference are not fully specified outcomes. In fact, there are many outcomes involving me sailing the Northwest Passage that I would prefer to many outcomes involving me sailing the Inside Passage. I will prefer an outcome involving me sailing the Northwest Passage if that outcome also involves you giving me a million dollars by the end of it. But this does not make it any less true that I prefer sailing the Inside Passage to sailing the Northwest Passage in the ordinary sense just described. My desire to also have a million dollars is irrelevant to this preference.

The alternative account of the standard of instrumental rationality I want to propose here evaluates the agent's actions in terms of her attitudes, be they desires or preferences, over simple states of affairs, like having another cup of coffee. Outcomes, as full descriptions of everything the agent may care about, comprise many simpler states of affairs. Since such simple states of affairs are thus features of fully described outcomes, I will hence refer to this notion of instrumental rationality as 'feature-based'. This alternative notion of instrumental rationality is not only more plausible than outcome-based instrumental rationality. As we will see, it can also express what is intuitively instrumentally irrational about being money-pumped.

One reason to err towards such a feature-based notion of instrumental rationality is that it intuitively seems like desires for (or preferences over) simpler states of affairs are more basic, and explain preferences we may have over outcomes. If I prefer Apartment A to Apartment C, it is because I desire to have a beautiful view and a short commute, and this outweighs my desire to have a large living space. Pettit (1991) goes so far as to refer to this idea as a 'platitude of desiderative structure'.[9]

[9]See my [redacted] for arguments that preferences over outcomes as they feature in decision theories in fact do not offer satisfactory folk psychological explanations on their own, without appealing to attitudes to

Reflecting on the way we make decisions in the context of conflicting desires provides further support for this 'platitude'. In these contexts, we do not generally come readily equipped with preferences over full outcomes. Trying to decide on where to live when moving to a new city can be very hard, even if I know all the relevant facts about the various options, as we are assuming in the context of certainty. Instead of consulting preferences over the relevant outcomes directly, I consult all the relevant attitudes I have regarding features of those outcomes. Any preferences over outcomes I form will be the result of weighing many conflicting considerations regarding the features of those outcomes. This seems to support the claim that in our reasoning processes, at least, attitudes to features of outcomes are more basic than preferences over outcomes.

Feature-based instrumental rationality goes further than this explanatory claim and says that our actions are ultimately rationally answerable to our attitudes to features of outcomes, not to our preferences over outcomes. This again seems intuitively plausible. The kind of reasoning we are engaged in when forming preferences over outcomes on the basis of our underlying attitudes to outcome-features appears to be instrumental reasoning. If this reasoning goes wrong, and I act on my outcome-preferences, this seems to be an instrumental failure. In fact, it is a kind of rational failure that is familiar, especially when forming preferences over outcomes involves weighing up many different kinds of consideration. Yet, outcome-based instrumental rationality cannot cast this as an instrumental failure.

Assuming that this alternative account of the standard of instrumental rationality is correct, much more needs to be said on exactly how an agent's preferences over outcomes and her actions should be based on her attitudes to features of outcomes. In fact, there is a growing literature in decision theory on how all-things-considered preferences are formed from more basic values.[10] Luckily, to account for the irrationality of being money-pumped, we only need to appeal to one minimal requirement of feature-based instrumental rationality.

We argued above that outcome-based instrumental rationality cannot establish that being money-pumped is irrational. To privilege the preference of A over A - ϵ as one that may not rationally be frustrated, and to give rational significance to the fact that A - ϵ is the same as A, except that the agent has less of some good she desires, we need to make features of outcomes instrumentally relevant. One plausible minimal principle of feature-based instrumental rationality is that one should not frustrate one of one's desires for features of outcomes unnecessarily, that is, without in return better satisfying another desire to a

features of outcomes.

[10]See, for instance, Dietrich and List (2013). One difficulty with this research project, within a Humean framework, is that we will have to impose a great deal of structure on the agent's feature-attitudes in order to say anything general about the process of preference formation. The Humean cannot take this structure to identify rational consistency requirements, in which case their status is unclear. Here, we are only concerned with whether the Money Pump Argument in favour of *acyclicity* works under feature-based instrumental rationality.

greater extent. This is the principle the money-pumped agent violates: She is deprived of money without receiving anything else she values in return, and her desire for money is unnecessarily frustrated.

More formally, I propose the following feature-based principle to capture the irrationality of being money-pumped:

Q: It is irrational to make a series of choices when

(a) it leaves one with an outcome y, when there would have been an alternative series of choices which leaves one with outcome x, which is identical to y, except that it has one additional feature one desires under the circumstances, or except that it differs to y with regard to only one feature such that x's feature is preferred to y's under the circumstances.[11] and

(b) there is an alternative series of choices of which (a) is not true, and within which each individual choice is permissible given one's feature-attitudes concerning the outcomes still available at the time of action.

This principle claims that it would be irrational for an agent to be money-pumped if she could have avoided being money-pumped by taking a series of actions each of which is itself unproblematic in terms of feature-based instrumental rationality. Note that it is a diachronic principle, as it has to be in order to apply to the diachronic money pump scenarios. [12]

Note that, if the Money Pump Argument is to be successful in the way under discussion now, then (b) must be true of money-pumped agents. Suppose it was not, and there is no alternative series of choices which does not leave an agent money-pumped, and which is itself permissible given the agent's feature-attitudes. In that case the Money Pump Argument as we just cashed it out would be hopeless. Proponents of the argument must believe that there is some acyclical preference relation that it would be rational for the agent to adopt. After all, we are trying to show that agents with cyclical preferences are not serving their ends as

[11] I mean features of outcomes to be picked out such that they themselves are not desired or preferred by the agent in one respect, while she is averse to them in another respect. For instance, if there is some respect in which I am averse to having more money, I do not violate principle Q when I am money-pumped. I will assume for the sake of argument that money is not like that for our agent, and that she unambiguously desires to have more money under her present circumstances. If this is too far fetched, we can reformulate the Money Pump Argument with appeal to a more basic good.

[12] This principle is similar to the diachronic part of Andreou's (2016, p. 1454) principle P, adapted to the feature-based account of instrumental rationality developed here, and weakened to have less problematic implications for cases where an agent's ends (here conceived of in terms of feature-attitudes) change over time, as in what are sometimes called 'temptation cases' (e.g. Gauthier 1996). In such cases, avoiding a sure loss in terms of one's feature-attitudes over time may require taking an action that is not endorsed by the attitudes the agent has at the time of action regarding features of the outcomes still available to her then. This might seem in itself instrumentally problematic. Q is silent on such cases.

well as they could, and have a reason, grounded in their ends, to adopt acyclical ones. If (b) was false, then no acyclical preference relation would be permissible given the agent's feature-attitudes, which we are now treating as the standard of instrumental rationality. The agent's underlying attitudes would be such that they can't be represented by an acyclical preference relation over outcomes. And so, if there is hope for this kind of Money Pump Argument, which, for the sake of argument, we will grant, money-pumped agents are irrational according to Q. The principle provides us with a plausible feature-based account of what is irrational about being money-pumped, and thus with support for P2.

6 Feature-Based Instrumental Rationality and Alternative Ways to Avoid Being Money-Pumped

We said above that the success of the Money Pump Argument relies on us being able to establish both P2, that being money-pumped is indeed instrumentally irrational, and P1, that agents with cyclical preferences can find themselves in situations where they can't rationally avoid being money-pumped while retaining their cyclical preferences. P2 implies that there is no alternative, rationally permissible way for agents with cyclical preferences to avoid being money-pumped other than adopting acyclical preferences. We have argued that outcome-based instrumental rationality cannot establish P2 for agents with cyclical preferences. The feature-based alternative we sketched in the last section, however, arguably can. Moreover, we argued that this is the more plausible account of the standard of instrumental rationality for independent reasons. The important question now is whether we can establish P1 according to this alternative account of the standard of instrumental rationality. I will argue that we cannot.

It has been proposed that ways of deciding in dynamic choice problems other than the sophisticated strategy described above may help an agent with cyclical preferences to never be money-pumped. One such choice strategy is what McClennen (1990) calls 'resolute choice'. Resolute agents choose a sequence of actions in the beginning of a series of choices, in accordance with their preferences then, and then simply go through with that sequence of actions. We can understand this either as involving counter-preferential choice, or a temporary adjustment to the agent's preferences within the context of the dynamic choice problem only. In our money pump scenarios, a resolute agent would choose as she would in a synchronic choice over all the outcomes she may reach in the series of trades she will be offered. Feature-based instrumental rationality would require her not to choose to be money-pumped in such a synchronic choice. Resolute choice, if it is rationally defensible, would thus keep an agent with cyclical preferences from being money-pumped. In fact, being resolute is not the only way of avoiding being money-pumped while generally keeping one's cyclical

preferences. Any way of stopping the trading early, either by acting counter-preferentially, or by temporarily adjusting one's preferences, will do so.[13]

Standard worries about these alternative choice rules are usually expressed in terms of outcome-based instrumental rationality, and thus, I think, miss the point in the context of our discussion. Counter-preferential choice may be obviously instrumentally irrational if preferences themselves form the standard of instrumental rationality. Likewise, perhaps temporary changes to one's preferences cannot be rationally required on an instrumentalist notion of rationality if those preferences themselves are what fix the agent's ends.[14] But, as I have argued above, outcome-based instrumental rationality makes it impossible to make the Money Pump Argument anyway. What defenders of the Money Pump Argument need to show is that alternative ways of avoiding being money-pumped are incompatible with instrumental rationality according to the feature-based alternative we described in the last section. But the irrationality of counter-preferential choice or temporary adjustments to one's preferences is not so obvious on the feature-based picture. I will argue in the following that indeed, feature-based instrumental rationality allows for alternative ways of avoiding being money-pumped.

According to feature-based instrumental rationality, attitudes to features of outcomes are the fundamental conative attitudes that pick out the agent's ends, and that instrumental rationality requires her to serve effectively. This raises the question of how the preferences over outcomes that feature in formal decision theories relate to those attitudes. As long as we continue to think of preferences over outcomes as conative attitudes in their own right, as is standard, the most natural way to understand our preferences over outcomes is as expressions of all of our feature-attitudes that are relevant for the outcomes under consideration taken together, as a result of the difficult weighting process alluded to above. Our preferences could then be understood as all-things-considered comparative evaluations of outcomes on the basis of our attitudes over all the states of affairs the outcomes comprise, that is, as a kind of summary attitude.[15]

Does feature-based instrumental rationality rule out alternative ways of avoiding being money-pumped when outcome-preferences are understood in this way? It would rule out counter-preferential choice if we could justify *preference-guidance* as a requirement of feature-based instrumental rationality. But we can immediately note one caveat regarding the requirement of *preference-guidance* under feature-based instrumental rationality. Unless we take outcome-preferences to be infallible as all-things-considered expressions of the

[13]See, for instance, Rabinowicz (2014), and Ahmed (2016) for alternative dynamic choice strategies that allow agents to avoid being money-pumped.

[14]One caveat here is that having different preferences can serve an agent's preferences as they are in cases of 'autonomous benefit'. However, we argued above that the money pump scenarios cannot be construed as such cases of autonomous benefit under the assumption of outcome-based instrumental rationality.

[15]See Hausman (2012) for a comprehensive defence of such an interpretation of preference.

agent's feature-attitudes, it can happen that the agent's preferences misrepresent her underlying concerns. It is not implausible to think that this happens, for instance, in some cases of temptation. At least in some such cases, it must be rationally permissible, in terms of feature-based instrumental rationality, to act counter-preferentially. And so *maximization*, or any other norm of preference-guidance, can be required at best conditionally: Agents ought to maximize if their outcome-preferences represent their feature-attitudes correctly.

In its conditional form, *preference-guidance* seems, at first sight, well-supported by feature-based instrumental rationality. As a matter of fact, the consequences of our choices are full outcomes: I choose to live in an apartment, along with all that that implies. I do not only choose a beautiful view. In fact many different desires seem to be relevant for my choice of apartment. A conative attitude over full outcomes thus seems more directly applicable to my choice than attitudes to only individual features of that outcome. And if my outcome-preferences correctly capture all the different attitudes to the states of affairs the outcome comprises, then acting on the preference means that my action is still ultimately based on those feature-attitudes. This seems to license the claim that feature-based instrumental rationality demands an agent's actions being in some sense guided by her preferences, if those preferences are indeed correct representations of her feature-attitudes.

Unfortunately, *preference-guidance* cannot be given such a simple justification if we want any hope of the Money Pump Argument being successful. The core problem is that letting one's preferences guide one's actions is rationally demanded by feature-based instrumental rationality only if there is only one preference relation over outcomes that is admissible given the agent's underlying feature-attitudes. Let me call this condition *uniqueness*. If it holds, then for each pair of outcomes, the agent's attitudes to the features of those outcomes uniquely determine whether the agent should prefer one over the other or be indifferent. In that case, it seems like feature-based instrumental rationality indeed requires *preference-guidance*. But suppose *uniqueness* fails: Several different preference relations over outcomes would equally well express the agent's attitudes to features of outcomes. And suppose the agent only adopts one of these permissible preference relations.[16] Feature-based instrumental rationality now no longer requires that the agent is guided by the preferences she adopted. If she chooses in accordance with preferences she does not have, but that would have been admissible given her attitudes to features of outcomes then she is not instrumen-

[16]Another possibility would be that in these kinds of cases, an agent should adopt a family of preference relations, expressing an 'imprecise preference'. However, in that case, we would already have given up on *preference-guidance* in the form of *maximization* or Schwartz's rule, since we would have to formulate new choice rules for families of preference relations. How we could then give an instrumental justification for each preference relation in the family being acyclical is unclear. Most plausibly, a choice rule for imprecise preference would be fairly permissive, and allow the agent to act in a way that is licensed by at least one of the permissible preference relations. Principle Q would justify us saying that the agent should choose such that she is not money-pumped, insofar as this is compatible with this permissive choice rule. But clearly, to conform to principle Q and such a permissive choice rule, the agent's individual preference relations need not each be acyclical.

tally criticizable according to feature-based instrumental rationality. She would be serving her feature-attitudes no worse than had she acted on her actual preferences.[17]

I here want to argue that proponents of the Money Pump Argument are not entitled to assume *uniqueness*, in which case they cannot establish *preference-guidance* and rule out alternative ways of avoiding being money-pumped. On the feature-based picture, the aim of the instrumentalist justification of acyclicity would be to show that agents are required to form acyclical preference in order to serve their feature-attitudes well. For there to be any hope for this endeavour, for any agent, there must actually be at least one acyclical preference relation that correctly captures her underlying feature-attitudes. As mentioned above, I will grant this here. But now the question of *uniqueness* arises: Is one such acyclical preference relation the one unique preference relation over outcomes that expresses the agent's feature-attitudes correctly?

This can't in fact be so if we want to deliver a Money Pump Argument that does any justificatory work. Suppose that for some agent with cyclical preferences, *uniqueness* holds with regard to some acyclical preference relation. That is, this acyclical preference relation is the only preference relation that correctly captures her underlying attitudes to features of outcomes – even though, as a matter of fact, she has cyclical preferences. In that case, her actual cyclical preferences must be a mistaken expression of her underlying feature-attitudes. Now, either the Money Pump Argument establishes this, or we know this independently. If we know this independently, then the Money Pump Argument does no work. We then already know that the agent's cyclical preferences are not those she rationally ought to have. But the intuitive plausibility of cyclical preferences in our examples in fact speaks against us always having independent reason to think the preferences are mistaken. Moreover, proponents of the Money Pump Argument presumably think it actually helps to justify *acycliclity*.

It must thus be the Money Pump Argument that establishes the agent's cyclical preferences are mistaken representations of her underlying concerns. The reasoning could be that if the agent's preferences over outcomes capture her feature-attitudes fully, then actions guided by the preferences should not frustrate those very attitudes. But this argument takes for granted that agents should always be guided by their preferences in action, which begs the question. We may be able to justify *preference-guidance* if *uniqueness* holds, and if the agent's preferences in fact correctly express her feature-attitudes. But the argument we just made assumes that the agent would also be guided by her preferences in action if she had cyclical preferences instead of the supposedly uniquely correct acyclical preferences. And we cannot take that for granted. There is, moreover, no feature-based instrumental justification that she should. In fact, as long as we grant that some acyclical preferences would represent the agent's underlying feature-attitudes well, feature-based instrumental rationality would

[17]Of course, one might wish to argue that some non-instrumental rational failure is involved here, but then we would be abandoning the Humean interpretation of decision theory.

permit the agent to act in accordance with those hypothetical preferences in money pump scenarios, thereby avoiding being money-pumped, while keeping the cyclical preferences.

Therefore, we cannot assume *uniqueness* if we want to make a successful Money Pump Argument. We should accept non-uniqueness at least in the sense that both the original cyclical preferences, as well as at least one acyclical preference relation are permissible representations of the underlying feature-attitudes. The latter is needed if the Money Pump Argument is to be successful, and the former takes the intuitive plausibility of the agent's original preferences seriously. The problem now is that any kind of non-uniqueness implies that there is no requirement, in terms of feature-based instrumental rationality, that the agent's actions should be guided by the preferences she actually adopts, for the reasons given above.

What does this mean for the Money Pump Argument? The Money Pump Argument only gets off the ground if the agent is guided by her preferences in action, for instance by following a sophisticated choice strategy. If the agent is guided by her preferences in action, she needs to adopt acyclical preferences to guarantee she can't be put in a situation where she is money-pumped. We have just found that, under feature-based instrumental rationality, we can no longer give an independent justification for *preference-guidance*. But one might think that the Money Pump Argument provides a *joint* instrumental justification for *acyclicity* and *preference-guidance*. After all, given stable preferences over time, *acyclicity* and *preference-guidance* together guarantee that the agent is not money-pumped in the ways we described.

I will grant for now that the agent in fact sticks with whatever preference relation she adopts throughout the series of choices she is offered. Still, the Money Pump Argument does not provide a joint justification for *acyclicity* and *preference-guidance*. Assuming acyclical preferences, acting in accordance with one's preferences is a good way (though not the only way) of avoiding being money-pumped. And assuming preference-guidance, having acyclical preferences is the only way of avoiding being money-pumped in the kinds of scenarios we described. But we have no independent instrumental justification for either *preference-guidance* or *acyclicity*. This makes it rationally permissible that, instead of fulfilling both, an agent avoids being money-pumped by violating both, that is, by keeping her cyclical preferences, and by refraining from always being guided by them in action.

This alternative response to money pump scenarios may make good sense in cases where cyclical preferences are especially tenacious, such as in the Self-Torturer Problem. In the light of the pairwise indiscernibility in pain of adjacent settings, you might find your strict preference for the higher of any two adjacent settings hard to shake. But you might still acknowledge that serving your respective desires for money and for being pain-free well requires you to stop at a sensible setting – against your preference at that point in time.

In fact, this coheres well with at least one prominent recent solution to the puzzle of the self-torturer.[18]

One may now respond that part of the point of having preferences is action-guidance, and that this speaks decisively in favour of responding to money pump scenarios by adopting acyclical preferences. However, either this argument relies on reinterpreting preference, or it is ineffective. There are several possible ways of understanding this appeal to the supposed action-guiding purpose of preferences. Firstly, we could redefine preference in such a way that we cannot choose to act counter-preferentially. For instance, we could say that preferences are dispositions to choose, as is popular particularly in economics in the guise of 'revealed preference theory'. But this would be a departure from the notion of preferences we have been dealing with so far, which takes preferences to be all-things-considered conative attitudes that express the agent's underlying feature-attitudes. The next section will argue that such a reinterpretation of preference does not in fact help the Money Pump Argument.

Secondly, we may mean that agents should form preferences over outcomes with the purpose of guiding choice in mind. Since only acyclical preferences can always be choice-guiding without leading to the agent being money-pumped, this may give the agent reason to form acyclical preferences. The problem with this argument is that we lack an instrumental justification for why agents should form preferences, understood as all-things-considered attitudes, with the purpose of always being choice-guiding in mind. That requirement looks a lot like *preference-guidance* itself, and we just saw that we cannot justify that requirement instrumentally.

Thirdly, perhaps the claim is just that agents in fact do form preferences with the purpose of guiding choice in mind. However, such a descriptive claim cannot ground the normative claim that agents *should* respond to money pumps by forming acyclical preferences and maximizing with regard to them. It would only establish that agents usually do that. But even that seems dubious. Note that cyclical preferences can be action-guiding outside of money pump contexts without problematic consequences. For instance, there might be nothing wrong with the self-torturer following her preferences in contexts where she is only offered a small subset of the original series of choices. And so cyclical preferences can fulfil the purpose of action-guidance for the most part, just not always. To the extent that they can't, the prevalence of cyclical preferences may just be evidence that agents do not form preferences with the purpose of always being action-guiding.

Thus, if we stick to a notion of preference that takes preference to be an all-things-considered conative attitude to outcomes on the basis of the agent's attitudes to features of those outcomes, then there is no decisive reason to avoid being money-pumped by adopting acyclical preferences rather than by selectively acting against one's preferences. The

[18] See Tenenbaum and Raffman (2012).

Money Pump Argument thus fails to provide a general justification for *acyclicity* according to feature-based instrumental rationality as well. While we could, with principle Q, explain what is instrumentally irrational about being money-pumped, and thus establish P2, feature-based instrumental rationality fails to establish that only agents with acyclical preferences can rationally avoid being money-pumped (P1). The next section argues that reinterpreting preference behaviourally does not salvage the argument.

7 Preference as Disposition to Choose

Rather than as conative attitudes to outcomes, the preferences that feature in decision theory are sometimes understood behaviourally. Many economists, in particular, take choice to be revealed preference, and, in turn, preference to be hypothetical choice. Savage (1972) explicitly interprets preferences as hypothetical choices for the purposes of his decision theory (p.17). In the philosophical literature, too, preferences are sometimes understood as dispositions to choose in the thin sense of hypothetical choice.[19]

Understanding preference as disposition to choose may help us, since under such an understanding of preference, the question of whether one ought to be guided by one's preferences in action appears to become otiose, and justifying *preference-guidance* appears to become unnecessary. To have a preference for one thing over another would simply mean that one would choose one over the other if offered the choice. Perhaps, then, the Money Pump Argument in favour of *acyclicity* may go through after all, although we are now arguing for the acyclicity of a different kind of relation — that of hypothetical binary choice.

To assess this, we need to be more precise on which hypothetical choices preferences are meant to pick out. We said that on the present understanding of preference, preferences are dispositions to choose one outcome over another if both are on offer. If we don't want to rule out cyclical preferences from the start, then we can't mean that preferences are dispositions to choose one rather than the other no matter what further options are also on the table. As we have seen, when choosing amongst a set of outcomes over which you have cyclical preferences, you must frustrate at least one of your binary preferences. But if preferences are dispositions to choose one option over another if both are available, then it can't have been true that you had that preference after all.[20] And so this notion of preference is

[19] I have argued for such an interpretation elsewhere, see [redacted].

[20] McClennen (1990) formulates conditions on choice, such that the agent's choice behaviour can be described by a pair-wise preference relation that forms a weak ordering, which is a stronger condition than *acyclicity*, and with regard to which the agent maximizes (pp.22-25). The condition that needs to be met is that choice is *context-free*. Roughly, context-freedom requires that an agent's choices from a set of options are not affected by adding further options, unless those further options are themselves chosen. If we define preference in a way that already presupposes that such a condition is met, then we are already guaranteed

unhelpful. Being guaranteed *acyclicity* through our definition of preference would be helpful if we also thought that we in fact all have preferences in the sense of preference we are appealing to. But sceptics about *acyclicity* would question that.

The better approach is to think of a preference of a over b simply as a disposition to choose a over b, when only those two options are available. But note that now *preference-guidance* when choosing amongst a larger set of options still needs justification. It is still an open question why an agent should choose in accordance with her binary choice dispositions when many options are on the table. Luckily, this problem does not affect the Money Pump Argument, since in the Money Pump Argument, the agent is only confronted with binary choices. With a notion of preference as binary choice disposition in hand, it is no longer open to us to say that an agent can avoid being money-pumped by keeping her cyclical preferences, and at some point acting against her preference.

Unfortunately for the Money Pump Argument, however, there is still an alternative way of avoiding being money-pumped open to the agent. If normative decision theory works with a notion of preference that identifies it with binary disposition to choose, then it is only a feature-based theory of instrumental rationality insofar as these dispositions to choose are appropriately responsive to the agent's feature-attitudes. For instance, if I desire money in my current circumstances, then I should be disposed to choose Apartment A over Apartment A - ϵ. But this means that, for the same reasons considered above, the preference relations that are appropriately responsive to an agent's feature-attitudes may be non-unique in the cases we are concerned with. Indeed, for the Money Pump Argument to be successful, we need to assume that they are non-unique at least in the sense that the original cyclical preferences, as well as at least one acyclical preference relation are appropriately responsive to the agent's underlying feature-attitudes.

Earlier, non-uniqueness caused trouble for the Money Pump Argument, because without it, we cannot justify *preference-guidance*. Now, non-uniqueness is problematic because it seems to make it rationally permissible for an agent to switch between different permissible preference relations over time, even if it is conceptually impossible for an agent to act against the preferences she does have at the time of action. This again makes it possible to rationally avoid being money-pumped without adopting acyclical preferences.

What we need for an agent to not be money-pumped is for her to stop trading at some point before she has lost money. Under the notion of preference we are now considering, this means that she must at some point prefer to stick with an outcome rather than trade it for the next one offered. It is enough if she has that preference at the time when somebody attempts to money pump her. But this is compatible with her having cyclical preferences

acyclicity. That seems to be exactly what we are doing when we think of preferences as dispositions to choose that are in a similar sense unaffected by what further options are on the table.

at every point in time. Consider the Self-Torturer Problem, for instance. For the agent to avoid being money-pumped, it just needs to be true that at some point, she will be disposed to stick with the lower of two settings she is offered. Although that temporarily breaks the preference cycle we originally characterized her with, her preference relation as a whole need not be acyclical, even at that point in time. Suppose, for instance, that at the point in time when she is offered the choice between S_{300} and S_{301}, she is disposed to choose S_{300}. She could at the same time retain all her other preferences, including, for instance, the following cycle:

$$S_1 \prec S_2 \prec ...S_{300} \prec S_{302} \prec ...S_{1000} \prec S_1$$

Let us assume that S_{300} is also the setting that would be picked out as the last permissible stopping point by the acyclical preference relation the agent would adopt, were she to pick one. If both the original cyclical preference relation as well as an acyclical preference relation with the last permissible stopping point of S_{300} are appropriately responsive to the agent's underlying feature-attitudes, it is eminently plausible that the preferences just described also express the agent's feature-attitudes correctly. After all, they are only minimally different from the cyclical preferences that seemed most natural to the agent. They just include the minimal change necessary to keep the agent from being money-pumped. If anything, they are thus more accurately based on the agent's underlying attitudes than fully acyclical preferences would be.

Of course, given the preference cycle still contained in the agent's preferences, the agent is still in danger of being money-pumped when she is offered a different series of trades — namely one that leaves out setting S_{301}. But actually being money-pumped would require that the agent's preferences remain stably what they are over that series of choices. Again, the agent could avoid being money-pumped by adopting a crucial preference to not trade any further at some point in time in that new series of trades. If we give up on the idea that the agent needs to have stable preferences over time, and across different choice situations, then agents can avoid being money-pumped without having fully acyclical preferences.

The second interpretation of the resolute dynamic choice strategy discussed earlier can be understood as just this kind of response to the Money Pump Argument. Resoluteness there required adjusting one's preferences within a dynamic choice problem only, and only insofar as is necessary to end up with the outcome one would choose in a single choice (that is, not money-pumped). The lesson is that agents can keep their original cyclical preferences for the most part, if they just make the necessary adjustments to their preferences needed to make sure they aren't money-pumped. These adjustments are specific to each dynamic choice problem they face.

We are thus left again with two ways of avoiding being money-pumped. One can adopt

stable, acyclical preferences, or one can adopt cyclical preferences that are unstable in just the right ways. Can we say anything in favour of avoiding being money-pumped in one way rather than the other? What may count in favour of adopting stable acyclical preferences is that we are often not in a position to know what further choices we will be offered, and we sometimes forget what choices we were offered in the past. If that is so, we might fail to adopt preferences that are cyclical but unstable in just the right way to avoid being money-pumped. A strategy that requires me to adopt a disposition to choose specifically to avoid being money-pumped requires me to keep track of potential money pumps to try and avoid them. But often, this will be difficult. Adopting stable, acyclical preferences circumvents this difficulty. With such preferences, agents cannot just stumble into being money-pumped.

However, there is a related disadvantage to the strategy of adopting stable, acyclical preferences. And that is that it requires an agent to have a specific disposition to choose even in situations where there is no danger of her being money-pumped. And that may be unnecessarily restrictive. This comes out especially in the Self-Torturer Problem. Suppose that the agent avoids being money-pumped in the series of choices we described above by adopting acyclical preferences that have her stop at S_{300}. If those are her stable preferences, then the agent is presumably also required to choose S_{300} over S_{301} at later points in time, in situations where she is only offered that one choice. But that seems unnecessarily restrictive. Much speaks in favour of choosing S_{301} in that situation.

In order to offer a Money Pump Argument in favour of *acyclicity* as an unconditional requirement, we need agents to decisively side with the strategy of adopting stable, acyclical preferences in the face of potential money pumps. And there is no such decisive reason, independently of the agent's other desires. At best, there are competing considerations that count in favour of adopting one or the other strategy of avoiding being money-pumped. Feature-based instrumental rationality thus does not require agents to adopt acyclical preferences independently of the content of the agent's desires. The project of providing a general instrumental justification for *acyclicity* has thus reached a dead end.

8 Conclusions and a Conditional Instrumentalist Defence of Acyclicity

If we want to understand decision theory as a theory of instrumental rationality, we need to provide instrumental justifications for its central requirements. One such central requirement is *acyclicity*. The standard instrumentalist defence of this requirement is the Money Pump Argument. The argument aims to show that agents who violate *acyclicity* can be placed in situations where they end up money-pumped. Being money-pumped, in turn, is deemed

to be instrumentally irrational. I have argued that this argument fails to provide a general instrumental justification of *acyclicity*. In fact, the common acceptance of the argument seems to rely on a fatal equivocation about the standard of instrumental rationality.

The Money Pump Argument, in its standard form, presupposes that agents choose in accordance with their stable preferences over outcomes. A requirement of *preference-guidance* in fact seems to be well justified if we take preferences over outcomes to be the fundamental conative attitude which forms the standard of instrumental rationality. But in order to show that being money-pumped is indeed instrumentally irrational, we need to reject preferences over outcomes as the standard of instrumental rationality. Instead, I argued, we should adopt a feature-based notion of instrumental rationality. According to this notion, instrumental rationality requires an agent to do well by her attitudes to features of outcomes. But on this notion, *preference-guidance* no longer holds as a general requirement in money pump scenarios, and neither do agents generally have to stick to a stable preference relation over time. And then agents can rationally avoid being money-pumped without adopting acyclical preferences.

Confidence in the Money Pump Argument thus seems to be based on equivocation about the standard of instrumental rationality. On one way of understanding the standard of instrumental rationality, *preference-guidance* is plausible, so that agents with cyclical preferences who abide by it indeed end up money-pumped. But we cannot show that this is instrumentally irrational. On the other way of understanding the standard of instrumental rationality, we can explain why being money-pumped is instrumentally irrational. But agents with cyclical preferences can rationally avoid being money-pumped. Ultimately, the Money Pump Argument fails to provide an unconditional instrumental defence of *acyclicity* on either way of thinking about the standard of instrumental rationality.

What we haven't ruled out, however, is that the Money Pump Argument could successfully provide a conditional instrumental defence of *acyclicity*. I here want to point to a promising way of doing so under the assumption of feature-based instrumental rationality, which we argued provides the more plausible picture of instrumental rationality. If Q is a requirement of feature-based instrumental rationality, then being money-pumped is arguably instrumentally irrational. The problem was that there are ways of avoiding being money-pumped that do not involve adopting acyclical preferences, and that are not generally ruled out by instrumental rationality. However, some agents may have desires that favour the response of adopting acyclical preferences, which would rationally rule out the other possible responses. For those agents, the Money Pump Argument would justify a requirement of *acyclicity*. The Money Pump Argument can thus provide us with an instrumental defence of *acyclicity* that is conditional on the desires in question.

When we are operating with the behavioural conception of preference described in the

last section, one plausible candidate for such a desire is the desire to have stable preferences, that is, stable binary choice dispositions, over time and across different choice contexts. Such a desire helps because it effectively resolves the problem of non-uniqueness: If an agent has a desire for stability of preference, then once she has settled on a preference relation, this preference relation becomes the only relation that expresses her feature-attitudes correctly from then on. In the money pump scenario, the strategy of avoiding being money-pumped by adopting cyclical preferences that are unstable in just the right way then performs worse according to the agent's feature-attitudes than the strategy of adopting stable acyclical preferences. The agent will be frustrating her desire for stable preferences when another available course of action, which would have also been permissible given the agent's other feature-attitudes, would have not done so. In fact, avoiding being money-pumped without adopting stable acyclical preferences would thus also violate principle Q.

We can hence make a successful Money Pump Argument in favour of *acyclicity* that is conditional on a desire for stable binary choice dispositions. And in fact, there may be something attractive about being settled in one's choice dispositions in this way. Still, and importantly, instrumental rationality cannot require us to have the desire to have stable preferences over time. This is because instrumental rationality is supposed to be silent about what desires an agent may have. As long as we remain Humeans about decision theory, the best we can say about *acyclicity* is that agents who have the right kinds of desires, such as a desire for stable preferences, have to abide by it. For the rest of us, instrumental rationality is more permissive. The alternative, of course, is to accept that there are non-instrumental coherence requirements on our preferences, and argue that *acyclicity* is one such requirement.

References

Arif Ahmed. Exploiting cyclic preference. *Mind*, fzv218, 2016. doi: 10.1093/mind/fzv218. URL +http://dx.doi.org/10.1093/mind/fzv218.

Jonathan Aldred. Intransitivity and vague preferences. *The Journal of Ethics*, 11(4):377–403, 2007.

Chrisoula Andreou. Cashing out the money-pump argument. *Philosophical Studies*, 173(6): 1451–1455, 2016.

Frank Arntzenius and David McCarthy. Self torture and group beneficence. *Erkenntnis*, 47 (1):129–44, 1997.

Donald Davidson, J. C. C. McKinsey, and Patrick Suppes. Outlines of a formal theory of value, I. *Philosophy of Science*, 22:140–160, 1955.

Franz Dietrich and Christian List. A reason-based theory of rational choice. *Nous*, 47(1): 104–134, 2013.

James Dreier. Rational preference: Decision theory as a theory of practical rationality. *Theory and Decision*, 40(3):249–276, 1996.

David Gauthier. Commitment and choice. In F. Farina, S. Vannucci, and F. Hahn, editors, *Ethics, Rationality, and Economic Behaviour*, pages 217–243. Oxford University Press, 1996.

Claudia González-Vallejo. Making trade-offs: A probabilistic and context sensitive model of choice behavior. *Psychological Review*, 109(1):137–155, 2002.

Johan E. Gustafsson. The irrelevance of the diachronic money-pump argument for acyclicity. *The Journal of Philosophy*, 110(8):460–464, 2013.

Jean Hampton. The failure of expected-utility theory as a theory of reason. *Economics and Philosophy*, 10(2):195–242, 1994.

Jean Hampton. Does Hume have an instrumental conception of reason? *Hume Studies*, 21 (1):57–74, 1995.

Daniel Hausman. *Preference, Value, Choice, and Welfare*. Cambridge University Press, 2012.

David Hume. *A Treatise of Human Nature*. Clarendon Press, 2007/1739.

James Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

Pekka Korhonen, Herbert Moskowitz, and Jyrki Wallenius. Choice behavior in interactive multiple-criteria decision making. *Annals of Operations Research*, 23(1):161–179, 1990.

Isaac Levi. Money pumps and diachronic books. *Proceedings of the Philosophy of Science Association*, 3:S235–S247, 2002.

David Lewis. Desire as belief. *Mind*, 97:323–332, 1988.

Edward McClennen. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge University Press, 1990.

Nicholas Miller. A new solution set for tournaments and majority voting: Further approaches to the theory of voting. *American Journal of Political Science*, 24(1):68–96, 1980.

Philip Pettit. Decision theory and folk psychology. In Michael Bacharach and Susan Hurley, editors, *Foundations of Decision Theory: Issues and Advances*, pages 147–175. Blackwell, 1991.

Warren Quinn. The puzzle of the self-torturer. *Philosophical Studies*, 59(1), 1990.

Wlodek Rabinowicz. Money pump with foresight. In M. Almeida, editor, *Imperceptible Harms and Benefits*, pages 123–154. Kluwer, 2000.

Wlodek Rabinowicz. Safeguards of a disunified mind. *Inquiry*, 57(3):356–383, 2014.

Frank P. Ramsey. Truth and probability. In R.B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, pages 52–94. Routledge, 1928/1950.

Leonard Savage. *The Foundations of Statistics*. Dover Publications, second revised edition, 1972.

Frederic Schick. Dutch bookies and money pumps. *Journal of Philosophy*, 83(2):112–119, 1986.

Thomas Schwartz. Rationality and the myth of the maximum. *Nous*, 6(97-117), 1972.

Howard Sobel. Cyclical preferences and world Bayesianism. *Philosophy of Science*, 64(1): 42–73, 1997.

Sergio Tenenbaum and Diana Raffman. Vague projects and the puzzle of the self-torturer. *Ethics*, 123(1):86–112, 2012.

Alex Voorhoeve and Ken Binmore. Transitivity, the Sorites paradox, and similarity-based decision-making. *Erkenntnis*, 64(1):101–114, 2006.

Bernard Williams. Internal and external reasons. In Ross Harrison, editor, *Rational Action*, pages 101–13. Cambridge University Press, 1979.