Philosophia (2024) 52:1269–1278
https://doi.org/10.1007/s11406-024-00794-6

Preferences: What We Can and Can't Do with Them

Johanna Thoma[1]

Received: 16 September 2024 / Revised: 16 September 2024 / Accepted: 12 October 2024 /
Published online: 14 November 2024

In her *Choosing Well*, Chrisoula Andreou puts forth an account of instrumental rationality that is revisionary in two respects. First, it changes the goalpost or standard of instrumental rationality to include "categorial" appraisal responses, alongside preferences, which are relational. Second, her account is explicitly diachronic, applying to series of choices as well as isolated ones. Andreou takes both revisions to be necessary for dealing with problematic choice scenarios agents with disorderly preferences might find themselves in. Focusing on problem cases involving cyclical preferences, I will first argue that her first revision is undermotivated once we accept the second. If we are willing to grant that there are diachronic rationality constraints, the preference-based picture can get us further than Andreou acknowledges. I will then turn to present additional grounds for rejecting the preference-based picture. However, these grounds also seem to undermine Andreou's own appeal to categorial appraisal responses.

Keywords Practical rationality · Preferences · Dynamic choice · Instrumental rationality · Diachronic rationality

1 Introduction

There is a classic argument type that goes as follows: Agents who have preferences that are disorderly — for instance, cyclical — foreseeably end up with prospects that are unambiguously bad by their own lights when they act on their preferences in dynamic contexts. Rational agents don't end up with such prospects. And so rational agents don't have disorderly preferences. In *Choosing Well*, Andreou (2023) agrees that rational agents don't foreseeably end up with such prospects. But she defends the rational permissibility of disorderly preferences. What has to give way is the idea

✉ Johanna Thoma
johanna.thoma@uni-bayreuth.de

[1] Universität Bayreuth, Bayreuth, Germany

that rational agents always act in line with their preferences. Rationality may require counter-preferential choice in order to guard against self-defeating prospects.

To explain how it is that rationality can guard agents with disorderly preferences against self-defeating prospects, Andreou puts forth a revisionary account of instrumental rationality. It is revisionary in two respects. The first concerns the goalpost or *standard* of instrumental rationality. Those who think and write about instrumental rationality in a way that is informed by orthodox decision theory often take instrumental rationality to be answerable to an agent's *preferences*: the point of instrumental rationality is to make sure an agent's preferences are served well.[1] Andreou argues that preferences cannot be the only kind of attitude that instrumental rationality is answerable to. Preferences, according to her, are "relational appraisal responses". In addition to such relational appraisal responses, which compare prospects with each other, we also have "categorial" appraisal responses, e.g. whether some prospect is "poor", "good" or "great". The point of instrumental rationality is also to ensure we end up in a high appraisal category, and not in an unnecessarily low one.

The second way in which Andreou's account of instrumental rationality is revisionary is that it is explicitly diachronic. It has to be such in order to tackle the problem cases for agents with disorderly preferences tackled in the book. Take the self-torturer problem due to Quinn (1990) and extensively discussed in *Choosing Well*. The self-torturer has cyclical preferences: He always prefers the higher of two adjacent settings, and he prefers the first setting with no money and no pain to the last setting with heaps of money in torturous pain. But what we can also plausibly say about him is that some of the settings he may end up with are good, perhaps even great, whereas others are terrible — these are his categorial appraisal responses. Still, no single setting increase takes him from, e.g. a determinately "good" to a determinately merely "okay" outcome. If this were the case, he would presumably also have a preference for the lower of the two settings. The boundaries between categorial appraisal responses in the self-torturer problem are vague, and this, one might think, is part of the explanation of why the self-torturer has the preferences he has in the first place. As we just noted, according to Andreou, instrumental rationality is also answerable to categorial appraisal responses. But if we applied that standard only to individual choices at a time, we would get no further in solving the self-torturer problem: The categorial appraisal responses for adjacent settings never clearly favour the lower of two settings as the self-torturer moves his way up to tragedy. What is needed for Andreou's solution is that instrumental rationality directly evaluates *series of choices*: In a series of choices, barring unexpected developments, instrumental rationality requires of an agent that she should not end up in a worse appraisal category than she could have ended up in. Instrumental rationality imposes diachronic constraints on us. This, too, is revisionary, given many decision theorists subscribe to "time slice rationality", the view that all constraints of rationality apply only synchronically.[2]

Andreou takes both revisions to be necessary for dealing with problematic choice scenarios agents with disorderly preferences might find themselves in. I agree with

[1] Something I have called "preference-based instrumental rationality" elsewhere (Thoma 2018).

[2] See Hedden (2015) for a book-length defence.

much of Andreou's analysis. But, focusing on problem cases involving cyclical preferences, I will first argue that her first revision is undermotivated once we accept the second. If we are willing to grant that there are diachronic rationality constraints, the preference-based picture can get us further than Andreou acknowledges. I will then turn to present additional grounds for rejecting the preference-based picture. However, these grounds also seem to undermine Andreou's own appeal to categorial appraisal responses.

2 Doing More with Preferences

Consider an isolated choice with two options the agent has a strict preference between. If we think that instrumental rationality is only answerable to our preferences, it is hard to avoid the conclusion that instrumental rationality requires the agent to choose the preferred option. Because the series of choices in the self-torturer problem consist of such binary choices, the preference-based notion of instrumental rationality gets agents with disorderly preferences into trouble *when applied only synchronically*. The same is true of classic money pump scenarios.[3] But were we to allow *diachronic* application of a preference-based standard, an agent's entire preference structure over the various options she could end up with in a series of choices becomes relevant. This is considerably richer information. If this were to help rule out the unacceptable outcomes in the problem cases for cyclical preferences (that is, ending up in torturous pain or being money-pumped), then Andreou's first revision would need more motivation — provided we accept the second.

Andreou claims repeatedly that no preference-based standard can do the trick.[4] That, I will argue in the following, is too quick. What we would need in order to provide a diachronic but preference-based response to the problematic choice scenarios is a preference-based choice rule for agents with disorderly preferences. Since I focus here on agents with cyclical preferences, we need a preference-based choice rule for agents with cyclical preferences. The most permissive such choice rules will indeed not help avoid being money pumped or self-tortured. Take this rule proposed in Schwartz (1972): an agent should choose a member of a subset of the available options such that (i) no option outside of the subset is strictly preferred to any member of the subset, and (ii) no proper subset of this subset fulfils condition (i). This rule implies that any option in a set over which an agent has a strict preference cycle is permissible as long as there is no further option that is weakly preferred to all options in the cycle. And so it does not rule out any of the options in the self-torturer case or classic money pump scenarios. In a standard version of a classic money pump scenario, you start out with a strict preference cycle over options A, B and C, with B

[3] This is so whether a "naive" or a "sophisticated" choice strategy is followed. See Gustafsson (2022) for an extensive discussion.

[4] E.g. pp. 65–66: "Without categorial appraisal responses, any alternative in a preference loop with a spectrum of options like the one of the self-torturer would be just as rationally permissible as any other." and p. 62: "If the self-torturer had nothing but the relational responses that Quinn describes and these responses were rationally permissible, then there would be no way to show that it is irrational for the self-torturer to end up at 1,000."

preferred to A, C preferred to B and A preferred to C. Option A-ε is the same as A, but where ε is an arbitrarily small amount of money deducted. Given she strictly prefers A to C, for small enough ε, the agent also prefers A-ε to C. Ending up with A-ε is the problematic outcome to be avoided: The agent loses money unnecessarily. Note here that if we also make the very natural assumption that an agent with the described preferences prefers A to A-ε and prefers B to A-ε, we simply have an enlarged strict preference cycle over A, B, C and A-ε. And so Schwartz's rule cannot rule out A-ε.

However, there are more restrictive preference-based choice rules for agents with cyclical preferences. Take the *Uncovered-Choice Rule* first introduced by Miller (1980) in the context of tournament theory: Only options that are not "covered" are rationally permissible to choose, where an option X is covered in case there is some option Y such that Y is strictly preferred to X, and for all other feasible options Z, Y is strictly preferred to Z if X is strictly preferred to Z — that is, X is never strictly preferred to some other option that Y is also not strictly preferred to. The intuitive rationale against covered options is this: For covered options, it will be the case that some other option is in two senses superior: Not only is it preferred; It also comes out favourable when we compare both options to any third option. So the idea is that there is no reason to choose a covered option when we could also pick one of the options that covers it.

In money pump scenarios, the outcome where one is money pumped will usually be covered in this sense, and thus impermissible when compared to the other possible outcomes the agent could reach in the series. For instance, in the standard example described above, A-ε will be covered by A given the assumptions we made about the agent's preferences: A is preferred to A-ε, and A and A-ε rank the same in comparison with the other available options. And so A-ε will not be a permissible choice out of the set of A, B, C and A-ε. In the self-torturer scenario, likewise, if the agent has intuitively sensible preferences, there will also be many options that are covered. An option with a low level of pain and some significant amount of money, for instance, should intuitively cover an option involving torturous pain: The agent prefers it, and it will never not be preferred to an option that the self-torturous option is preferred to. At the same time, there must be options in both kinds of scenario that are not covered, as the uncovered set is necessarily nonempty (see Schwartz, 1990). These should, if the agent's appraisal responses as a whole are coherent, lie in the range of options classed in the highest appraisal category on Andreou's picture, i.e. "good" or "great" options.

And so there is an alternative response to the problematic choice scenarios that goes as follows: Agents with cyclic preferences are rationally required not to foreseeably end up picking a covered option in a series of choices. This will require them to choose counter-preferentially in some binary choices. But once we take a diachronic view of instrumental rationality, this seems no more problematic than the sense in which the Uncovered Choice Rule more generally may require "counter-preferential choice": applied to a one-off choice among many options, the Uncovered Choice Rule may require agents to choose an option to which another option is strictly preferred. In fact, that is precisely what it would require in the classic money pump scenario: In our standard example, all of and only A, B and C would be permissible, and so the agent would have to pick an option to which another is strictly preferred. If

one finds the Uncovered Choice Rule attractive, it thus wouldn't be ad hoc[5] or inconsistent to insist at once that it is sometimes required to act against one's preferences, and that it is also impermissible to end up with an option that serves one's preferences worse than other options — where serving one's preferences worse is understood in terms of ending up with a covered option.

In many ways, this response is like Andreou's: It allows for disorderly preferences while requiring agents to avoid clearly bad outcomes in dynamic choice scenarios by choosing counter-preferentially. But it only grants one of Andreou's revisions: It commits to a diachronic requirement of rationality, while seemingly holding on to the idea that preferences are the standard of instrumental rationality — the Uncovered Choice Rule is formulated in terms of preferences alone. The only obstacle insofar as offering a satisfactory resolution of the self-torturer problem or money pump scenarios is concerned thus appears to be time-slice rationality. Give that up, apply principles of rationality diachronically, and we get the intuitively correct result. It thus seems like we can do more with preferences than Andreou acknowledges. What this shows, I think, is that more would need to be said to establish that preferences are unsatisfactory as the sole attitude that instrumental rationality should be answerable to.

3 The Limits of Preference-Based Instrumental Rationality

The last section showed that we don't necessarily need non-preference-based rationality principles in order to establish a rational requirement to act counter-preferentially to avoid being money-pumped or self-tortured, as long as we are willing to apply preference-based criteria diachronically. So what would nevertheless speak in favour of acknowledging that instrumental rationality is (also) responsive to other attitudes? One strategy would be to look for cases where preference-based criteria and criteria that appeal to other — perhaps categorial — attitudes come apart. In particular, we would need cases where preference-based criteria alone can't give the intuitively correct verdicts, and where non-preference-based ones can support the intuitively correct judgements.

There are some cases where the Uncovered Choice Rule is more restrictive than the demand not to end up in an unnecessarily low appraisal category. For instance, this would be so in money pump scenarios where all options — including the one where one is money-pumped — are in the same appraisal category, say, they are all great. This kind of case will not help establish that categorial responses are also necessary for a satisfactory account of instrumental rationality. In fact, Andreou herself

[5] Andreou brings up the ad hocness charge in a slightly different context, namely when responding to the idea that it is specifically bad to end up in a place that serves one's preferences worse than the option one started out with: "My point, in short, is that, as soon as one grants that, in cases like the case of the self-torturer, it is rationally permissible, and in- deed rationally required, that one stick with an option even though it serves one's preferences worse than another available alternative, then it seems ad hoc to insist that rationality does not permit a series of choices that leads one to an option that serves one's preferences worse than the alternative one began with." (p. 66).

proposes a principle that seems stricter than what can be established by appealing to categorial responses alone to deal with money pump cases:

"P: It is irrational to make a choice or series of choices that leads one to an alternative Y which is such that Y is identical to another alternative X except with respect to one dimension of concern and, in that respect, Y is dispreferred to X." (p.71).

This principle appeals to appraisal responses (indeed relational ones) that are *partial*: An option being preferred or dispreferred in some respect, along some dimension. But it does not appeal to categorial appraisal responses to the options one is choosing between. It is in fact possible that X and Y are in the same appraisal category while P is being violated. I will return to partial appraisal responses below. But for now note that appeal to such partial responses, too, may not be necessary if preference-based criteria like the Uncovered Choice Rule can accommodate the intuitively correct responses in problem cases like money pump scenarios.

So is it possible that an option is not covered and so permitted by the Uncovered Choice Rule, but at the same time is in a lower appraisal category than another available option? If such an option were intuitively impermissible, that would then seem to clearly demonstrate the need for categorial appraisal responses as an additional standard of instrumental rationality. This kind of divergence would require either that an agent directly prefers an option in a lower appraisal category to one in a higher appraisal category, or that there is some third option to which the one in the lower appraisal category is preferred but the one in the higher appraisal category is indifferent to or dispreferred to. The vagueness of categorial appraisal responses in the self-torturer problem, for one, does not seem to furnish us with such cases. The only such cases that occur to me are ones where there seems to be a kind of systematic mismatch between preferences and categorial appraisal responses. And this kind of mismatch at the same time supports the judgement that at least one of these appraisal responses is somehow mistaken. An agent who prefers an option she thinks is merely okay to an option she thinks is great seems to make a mistake: Her different responses are out of whack, not coherent - either she should not have these categorial responses or she should not have these preferences.

I think that preferences of the type we are talking about here — preferences over options, the objects of choice, and which standard decision theory takes to be subject to constraints like acyclicity — are indeed a type of attitude that can be mistaken. And such preferences can be mistaken not merely on objective grounds, but on subjective grounds as well. They can be mistaken representations of what an agent truly cares about. If that is true, we have reason to think instrumental rationality is not ultimately answerable to preferences, quite independently of what is required to get problem cases like money pump scenarios and the self-torturer problem right. Rather, it is answerable to the attitudes that preferences fallibly aim to represent.

I have argued for the thesis that preferences can be mistaken representations of what we truly care about in more detail elsewhere. I'll briefly expand on three main reasons for thinking so here.

1. From a first person perspective, preferences are not the starting point of deliberation, but come rather near the end of deliberation.[6] Whenever a choice is not entirely straightforward, I usually do not start out with a settled preference between the options I am choosing between. Rather, I ask myself what it is that counts in favour or against the various options I am choosing between. On the basis of that, I may form a preference and will eventually choose. What counts in favour of or against an option may of course sometimes be called a "preference" in a more every-day sense: For instance, I may prefer less pain to more pain. But this is not the sense of preference at issue in decision theory and in debates about the permissibility of cyclicality. There we are talking about preferences between the objects of choice. And I am not choosing more or less pain in the self-torturer problem. Rather, I am choosing options that involve more or less pain, but also more or less money. In most real-world contexts, there will in fact be many more than just two dimensions to an agent's objects of choice. What seems to be the starting point of deliberation are attitudes to features along those dimensions, or what I called above *partial* appraisal responses. There will be certain respects in which we evaluate an object of choice positively and others in which we evaluate them negatively. Forming a preference and ultimately choosing requires us to somehow go from a variety of partial attitudes to an all-things-considered one. It seems clear that mistakes can be made here, that it is possible to form a preference that does not accurately represent one's underlying partial attitudes. And when this is the case, intuitively it's the underlying partial attitudes (some of which can of course themselves be quite complex) that instrumental rationality is answerable to (see Thoma 2021b). My second point is a special example of this.
2. There are certain dominance principles that we cannot give a justification for if we take decision-theoretic preferences to be the only type of attitude instrumental rationality is responsive to. Andreou's principle P above is a case in point here. That principle seems highly intuitive, especially in the synchronic case. How could it be anything but instrumentally irrational to choose an option that one takes to be clearly better in some respects and worse in none? But it is conceivable that an agent might form a preference for the dominated option over a dominating one. If instrumental rationality were answerable only to preferences, then instrumental rationality would in such a case require the agent to pick the dominated option and violate principle P. But that seems like the wrong result: the agent should abide by principle P. Her preferences were mistaken. Or to put things in the context of money-pump scenarios: It is irrational to end up being money pumped even if one preferred A-ε to A (provided ε really is unambiguously a loss, and the preference was in that sense a mistake).
3. There is also a (weak) case to be made from (libertarian) paternalist practice. Libertarian paternalism is characterised both by a commitment to interventions in people's choice behaviour that do not constrain their choices, as well

[6] Relatedly, from a third person perspective, preferences over the objects of choice only provide very thin explanations of people's choices. When you ask why somebody picked one of two options where these options differ along a number of dimensions, saying that she preferred one over the other is almost completely uninformative. What we would want to know, for a satisfactory explanation, is *what it is about* the preferred option that tipped the balance. See my (Thoma 2021a) on this point.

as a commitment to intervening only to help agents pursue their own ends. The second element in this is also often referred to as "means paternalism". Behavioural welfare economics studies how to identify an agent's welfare when her (revealed) preferences do not abide by the standard axioms of decision theory. Given the typically subjective understanding of welfare in economics, we can think of this as the exercise of identifying the ends that serve as the standard for means paternalist intervention, expressed in an alternative preference relation. All standard methods for this process of "preference purification" involve taking at least some of an agent's actual (revealed) preferences to be mistaken by an agent's own lights. Preference purification aims to bring to light the preferences an agent ought to have had by her own lights — where these "own lights" must be something other than the agent's actual preferences. And clearly, when purified preferences feed into informing means paternalist interventions, economists take instrumental rationality not to be answerable to the preferences the agent actually has, but to those she should have had on the basis of some further, underlying attitudes (see Thoma 2021c).

There are reasons, then, to think that preferences indeed ought not to be taken as the sole standard of instrumental rationality, even if this was under-motivated by problem cases like money pump scenarios and the self-torturer problem. In fact, these are reasons to think that preferences are not an ultimate standard of instrumental rationality at all — rather they are fallible representations of the true standard. Whenever they fail to accurately capture the true standard, we are not required to serve our preferences. These are also reasons to think that the attitudes instrumental rationality is ultimately answerable to are partial ones: Partial attitudes are where deliberation typically starts, and they ground dominance principles like P. That alone of course does not show preferences are useless in a theory of instrumental rationality. Preferences are summary attitudes that help both agents and analysts systematise what is at stake in a decision problem. Rationality principles formulated in terms of them can still be true and helpful, provided that we make them conditional on the preferences being non-mistaken by the agent's own lights, and as long as they are ultimately justifiable in terms of the true standard of instrumental rationality. The Uncovered Choice Rule applied only to non-mistaken preferences is a candidate for such a principle.

4 Conclusion: Do We Need Categorial Appraisal Responses?

To return to Andreou's revisionary account of instrumental rationality, however, I think the reasons against preference-based instrumental rationality presented here are also reasons to be sceptical that categorial responses of the type Andreou discusses are attitudes that instrumental rationality is ultimately answerable to. This is because the categorial responses she mentions also have as their object the objects of choice, such as, in the self-torturer problem, bundles of pain and money. But as with decision-theoretic preferences, when choosing between two options with multiple choice-relevant features, we don't usually start out appraising one, say, as great and the other as good. Rather, we form such an appraisal response on the basis of

the things that we take to speak in favour of or against each option. And just as with preferences, it is conceivable that we make mistakes in our categorial appraisal responses to the objects of choice. It is even conceivable that we place an option that is dominated in the sense specified by principle P in a higher appraisal category than the dominating option. As with mistaken preferences, in such cases, intuitively instrumental rationality is still bound by the dominance principle, and not beholden to the mistaken categorial responses. Again as with preference, this does not mean that categorial responses to the objects of choice are useless in a theory of instrumental rationality — they may be useful summary attitudes in terms of which we can formulate candidate principles of rationality provided we caveat for their fallibility.

Still, a worry remains: If both preferences and categorial appraisal responses are at best useful but fallible summary representations of what is the true standard of instrumental rationality (attitudes that are partial), then we would only need both if one can't serve the role we want it to on its own. But, as I have argued above, money pump scenarios and the self-torturer problem at least do not support the case that we need both — it seems like principles formulated in terms of preferences can get us just as far. I thus agree with Andreou's rejection of preferences as the sole attitude that instrumental rationality is ultimately answerable to, but I am sceptical that adding categorial responses as a second type of standard for instrumental rationality solves the problem. The true standard of instrumental rationality appears to me to be partial attitudes. Of course partial attitudes may themselves come in a relational (e.g. preferring less pain to more) and a categorial (e.g. loving things insofar as they are boaty) form. But whether we are pluralist or not at *that* level seems to be orthogonal to the discussion of problem cases like the self-torturer and money pump scenarios, and the critical discussion of orthodox decision theory Andreou is engaged in.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing Interests The author declares that she has no conflict of interest.

References

Andreou, C. (2023). *Choosing well: The good, the bad and the trivial*. Oxford University Press.
Gustafsson, J. E. (2022). *Money-pump arguments*. Cambridge University Press.
Hedden, B. (2015). *Reasons without persons: Rationality, identity, and time*. Oxford University Press.

Miller, N. (1980). 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.
Quinn, W. (1990). The puzzle of the self-torturer. *Philosophical Studies*, *59*(1), 79–90.
Schwartz, T. (1972). Rationality and the myth of the maximum, *Nous* 6, pp. 97–117.
Schwartz, T. (1990). Cyclic tournaments and Cooperative Majority Voting: A solution. *Social Choice and Welfare*, *7*(1), 19–29.
Thoma, J. (2018). Temptation and preference-based instrumental rationality. In J. Bermudez (Ed.), *Self-control, decision theory, and rationality*. Cambridge University Press.
Thoma, J. (2021a). Folk psychology and the interpretation of decision theory. *Ergo*, *7*, 904–936.
Thoma, J. (2021b). Judgementalism about normative decision theory. *Synthese*, *198*, 6767–6787.
Thoma, J. (2021c). On the possibility of an anti-paternalist behavioural welfare economics. *Journal of Economic Methodology*, *28*(4), 350–363.