Temptation and Preference-Based Instrumental Rationality

July 31, 2017

1 Introduction

We all seem to be prone to temporary shifts in our preferences. I might plan, for instance, to only stream one episode of a TV show during my coffee break. But then, once I have watched the first, I come to prefer to watch another. In the dynamic choice literature, such cases have come to be known as 'temptation problems', since I can be said to be tempted to watch a second episode of TV.[1] Temptation problems confront us with a puzzle about instrumental rationality. On the one hand, an agent seems to do better by her own lights if she does not give into the temptation, and does so without engaging in costly commitment strategies. This seems to indicate that it is instrumentally irrational for her to give into temptation. On the other hand, resisting temptation also requires her to act contrary to the preferences she has at the time of temptation. But that seems to be instrumentally irrational as well.

My starting point here is that to make any progress in the resolution

[1] See, for instance, Gauthier (1996) and McClennen (1998). Bratman (1998) and Holton (2009) consider the same kinds of problems, but speak primarily of 'evaluative judgements' rather than preferences.

of this puzzle of instrumental rationality, we need to be more explicit about what we take to be the standard of instrumental rationality against which an agent's actions are evaluated. This chapter argues that it is a pervasive, but usually implicit assumption in rational choice theory, that the agent's preferences over the objects of choice form the standard of instrumental rationality against which the agent's actions are evaluated. I call this assumption 'preference-based instrumental rationality'. With this notion of instrumental rationality in hand, I consider the two most prominent types of argument for why resisting temptation could be instrumentally rational, even though it requires us to act counter-preferentially. I argue that both arguments fail under preference-based instrumental rationality.

The first type of argument is a two-tier argument, whereby not the agent's individual actions, but her deliberative strategies over time are assessed instrumentally. Individual actions, in turn, are judged instrumentally rational if they are endorsed by the best deliberative strategy. A strategy that has the agent resist temptation is then argued to be instrumentally best. The core problem for two-tier accounts is that preference-based instrumental rationality implies that in temptation cases, the standard of instrumental rationality itself shifts. I argue that this means that we can no longer say that a strategy of resisting temptation is instrumentally best. According to the second type of argument, resisting temptation is the result of mutually beneficial cooperation between the agent's 'time slices'. Agents then have the same kinds of reasons to engage in this intrapersonal cooperation as they have to engage in interpersonal cooperation. I argue that, given preference-based instrumental rationality, no plausible account of mutually beneficial cooperation between an agent's time slices can be given.

One might think that giving up preference-based instrumental rationality will help these arguments. However, I argue that this is not so. Doing so either doesn't do away with the problems, or it makes the arguments redundant, save for a special case. Giving up preference-based instrumental rationality creates the possibility that the agent's preferences misrepresent the true standard of instrumental rationality. But if the true standard of

instrumental rationality is still shifting in a particular temptation case, the same problems as before arise. If, on the other hand, the true standard of instrumental rationality is stable in a temptation case, then we already have a straightforward justification for why resisting temptation is instrumentally rational: It is best according to the true, stable standard of instrumental rationality.

The choice we thus face is the following: Either we stick with preference-based instrumental rationality, in which case we are left to conclude that resisting temptation is instrumentally irrational — unless we find a better argument to the contrary. Or we abandon preference-based instrumental rationality, in which case temptation cases may turn out to be much less puzzling. This chapter will conclude by suggesting that the latter option makes both better sense of the phenomenon of temptation, and has independent appeal.

2 A Temptation Case

Suppose that I like to stream an episode of a TV show when I take my afternoon coffee break. As my break starts, at t_1, I prefer to watch only one episode, and then get back to work. But after I have watched that first episode, at time t_2, I prefer to watch another one over stopping. Once I have watched the second episode, I then return to my earlier preferences and would prefer just having watched the one episode.

Let O_0 be the outcome of watching no TV: I will get all my work done, but my coffee break will be boring. Let O_1 be the outcome of watching one episode, namely that I have an interesting coffee break and also get all my work done afterwards. O_2 is the outcome of watching two episodes: While I get to watch two episodes of an interesting show, I will not get my work done. Let $\succ$ represent strict preferences between outcomes. My preferences

at the different points in time are the following:

$$t_1 : O_1 \succ O_0 \succ O_2$$

$$t_2 : O_2 \succ O_1 \succ O_0$$

$$t_3 : O_1 \succ O_0 \succ O_2$$

The dynamic decision problem I face can now be illustrated by the decision tree in Figure 1. The square nodes here represent choices I need to make. In each case, I can decide whether to go 'up' or 'down'.

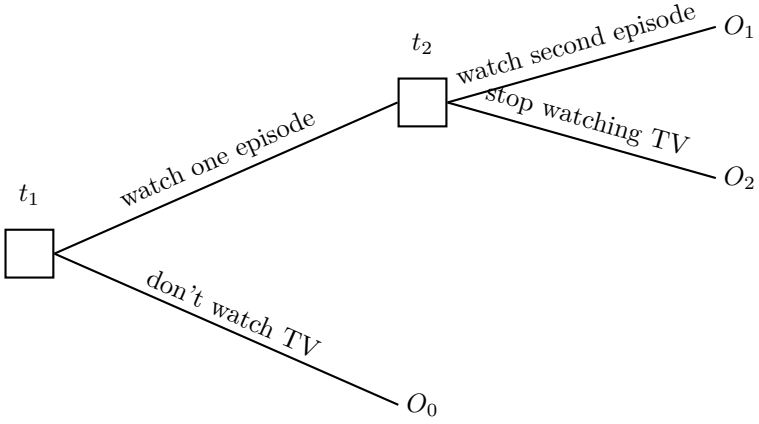

Figure 1: Temptation Problem

Now suppose that if I make a decision at time t_2, I simply choose according to my preference between the available outcomes O_1 and O_2, and watch a second episode. Further suppose that I predict that I will do so at t_1, and treat this as certain. I then take myself to effectively face the choice between O_0 and O_2 at t_1. If again, I simply go with my preference over these outcomes, I will choose to not watch any TV. I do so even though, at every point in time, I prefer watching one episode to watching no episodes.

As we said, this kind of case is often referred to as a 'temptation problem' in the dynamic choice literature. The choice behaviour I have just described, in turn, is referred to as 'sophisticated'.[2] At each point in time,

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

the agent predicts what she will rationally do in the future. And, under conditions of certainty, she chooses in accordance with what act brings about the outcome she most prefers at that point in time, out of the ones that are still available to her then. If the agent treats her prediction of future rational behaviour as certain, sophisticated choice simply follows from a rule that demands maximizing with regard to one's preferences at every point in time. Such a rule seems to be an archetypical requirement of instrumental rationality.

If I follow this sophisticated choice strategy in the temptation case, however, I will end up not watching TV. It is rationally impossible for me to watch one episode and then not give into the temptation to watch another. Yet many philosophers have thought that it must be rationally possible for agents to resist temptations of this sort. And then there must be something wrong with sophistication as a requirement of instrumental rationality, or else resisting temptation is in conflict with instrumental rationality.

Indeed, one main source of the intuition that resisting temptation can be rational is itself instrumental in nature. And that is that the agent in these examples seems to end up worse off by her own lights. It seems like her life would go better if she had the capacity, at t_2, to not act in accordance with her temporary preferences. In fact, we can interpret the choice to not even watch the first episode as a kind of costly pre-commitment mechanism. I forego the first episode to bind myself not to watch another. The cost of pre-commitment, however, only seems to buy me something I could have had for free, had I only been able to resist the temptation. We may consequently want to provide an argument that claims that it is rational to act against one's preferences at the time of temptation because doing so leaves one ultimately better off.

Two types of such arguments can be distinguished. On the first, it is sometimes rational to resist temptation because doing so is called for by the best deliberative strategy, by the agent's own lights. According to this type of argument, usually referred to as a 'two-tier' argument, the rationality of the individual action should be assessed by whether it is endorsed

by the best deliberative strategy. On the second type of argument, resisting temptation is rational because it is the product of mutually beneficial cooperation between the agent's time slices. Let us call these 'time-slice cooperation arguments'.

In the following, I want to show that the instrumentalist arguments in favour of resisting temptation either fail, or are redundant, save for a special case. To see why, we have to focus our attention on the question of what we take to be the standard of instrumental rationality when making these arguments. That is, what do we take to be the conative attitude by which we evaluate the agent's actions or deliberative strategies?

3 Preference-Based Instrumental Rationality

Instrumental rationality is traditionally understood as requiring agents to take the best means to ends they desire. But note that ends and desires do not appear in standard rational choice theory. Nor did they feature in the description of our temptation problem. Instead, standard rational choice theory, as well as much of the wider literature on practical rationality features binary preferences. Given this ubiquity of preferences, how should we then think about the requirements of instrumental rationality?

On a broad understanding of instrumental rationality, actions or principles of choice are evaluated in light of the agent's own conative attitudes, or pro-attitudes.[3] If we adopt such a broad understanding, there is then an open question as to which of the agent's conative attitudes should be the basis of evaluation of the agent's actions. Rational choice theorists typically assume that this basis of evaluation should be the agent's preferences over

[3]Williams (1979) arguably articulates such a broad understanding of instrumental rationality when he argues that an agent only has a reason to do x if doing x somehow advances an element in her "subjective motivational set" S. This subjective motivational set, according to Williams, could contain various different pro-attitudes, plans or commitments.

the objects of choice, which, in the case of choice under certainty, are outcomes. According to what, in the following, I want to call 'preference-based instrumental rationality', instrumental rationality is about acting well in the light of one's preferences over outcomes. That is, preferences form the standard of instrumental rationality. Outcomes, in turn, then play the role of ends in preference-based instrumental rationality. This notion of instrumental rationality requires that we take preference to be a binary kind of conative attitude, which matches the intuitive sense of preference we have been using so far.[4]

The move to a preference-based notion of instrumental rationality is very common, but often implicit, and seldom argued for.[5] Crucially for us, preference-based instrumental rationality appears to justify a requirement to maximize with regard to one's preferences, and thus sophistication, instrumentally. If instrumental rationality requires us to act well in the light of our preferences over outcomes, then, provided there is a most highly ranked outcome, instrumental rationality seems to require us to take the action that leads to it. If I choose in this way, I will not frustrate any of my binary prefer-

[4]Preference is sometimes also interpreted behaviourally, as a kind of disposition to choose, in particular by economists. Doing so would require us to look for the standard of instrumental rationality elsewhere, as Section 8 does.

[5]Many authors use desire and preference interchangeably. Others equate ends with outcomes. In this passage, for instance, Morris and Ripstein (2001) claim that rational choice theory requires agents to have rankings of ends: "The traditional theory of rational choice begins with a series of simple and compelling ideas. One acts rationally insofar as one acts effectively to achieve one's ends given one's beliefs. In order to do so, those ends and beliefs must satisfy certain simple and plausible conditions: For instance, the rational agent's ends must be ordered in a ranking that is both complete and transitive." (p.1) Yet others claim that ends and desires are different from preferences over outcomes, but still abide by preference-based instrumental rationality. Gauthier (1987) claims that ends may be inferred from preferences, but that preferences are primary, and that rationality is about maximizing a measure of preference (pp.22-26). Nozick (1993), too, claims that preferences are basic, and that ends and desires can be derived from them through some process of filtering or processing (p.144). Hampton (1994) provides a critique of standard rational choice theory that relies on interpreting rational choice theory in terms of preference-based instrumental rationality.

ences. Preference-based instrumental rationality thus seems to lend support to one side of the puzzle we started out with: If resisting temptation requires acting against our preferences, that seems instrumentally irrational. In the following, I will consider whether instrumentalist arguments in favour of resisting temptation, and thus of acting counter-preferentially, nevertheless go through under preference-based instrumental rationality.

4 Two-Tier Arguments

Two-tier arguments proceed from the observation that agents sometimes serve their ends best if they do not, at every point in time, take their reasons directly from their ends. Or, in terms of preference-based instrumental rationality, agents sometimes serve their preferences best if they do not act in accordance with their preferences at every point in time. This is the basic insight David Gauthier (1994) provides in his "Assure and Threaten". Given this basic insight, Gauthier argues that instrumental rationality in fact demands that we assess not individual choices, but entire deliberative procedures by how well they serve our preferences. We then regard actions as rational if and only if they are in accordance with the best deliberative procedure — even if that procedure calls for a choice that serves the agent's preferences at the time of action less well than another.

There are various worries about the two-tier nature of this account. For instance, we do seem to have a strong intuition that whether an action is instrumentally rational depends on how well it serves the agent's preferences at the time of action. Bratman (1998) calls this the *standard view*. Denying it would suggest that we can be moved by the 'dead hand of the past', that is, by past preferences or by plans previously made. But what preferences I once held but no longer hold does not seem instrumentally relevant at the time of action. Neither do plans previously made that do not serve my current preferences. Under preference-based instrumental rationality, these considerations seem to support sophistication.

However, here I want to raise another, more fundamental problem for two-tier arguments for the rationality of resisting temptation. And that is that in temptation cases, we cannot in fact establish that a deliberative strategy that endorses resisting temptation really serves the agent's preferences best. And so, under the assumption of preference-based instrumental rationality, the argument does not get off the ground. While this is a problem for two-tier accounts in general, let me first look at Gauthier's own in more detail.

Gauthier (1994) appeals to the counterfactual consideration that the agent at each point in time thinks that she is better off going through with a resolution than she would have been had she made no resolution at all. The example that originally motivated Gauthier's argument is an intertemporal Prisoner's Dilemma between two agents first described by Hume (2007/1739), III.2.5 520-521. In this example, two farmers A and B would benefit from helping each other harvest their crop rather than doing it each on their own. However, for each, it would be even better if he received help with harvesting his field, without having to reciprocate. Now we imagine that the dynamic structure of the case, illustrated in Figure 2, is such that A's field is ready to harvest earlier.

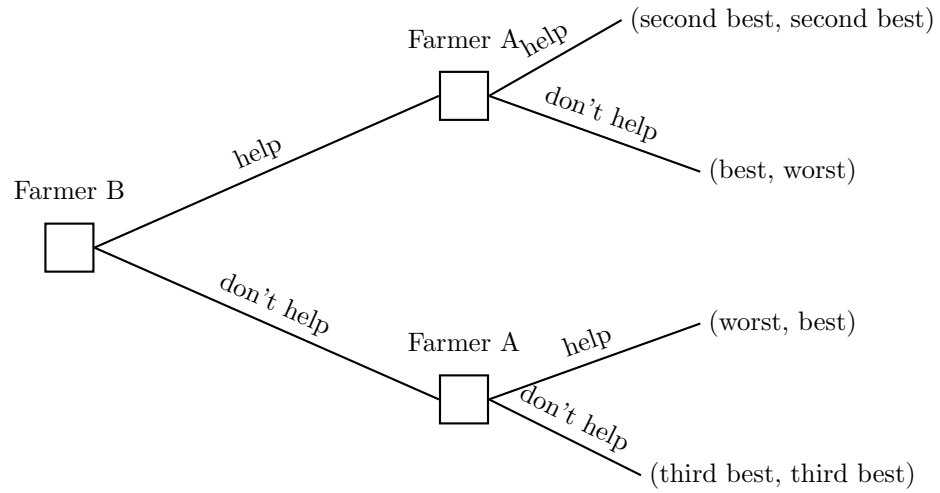

Figure 2: Intertemporal Prisoner's Dilemma

In this dynamic choice problem, if A maximizes with regard to his preferences once it is his turn, then whether he was himself helped or not, he will decide not to help B. Either way, he will be better off not helping. But knowing that, B will not help A in the first place. He would know that the favour would not be returned, and so is better off not helping. The farmers thus each end up with a worse outcome than they could have had: They each end up harvesting their fields alone, when they could have helped each other and both been better off.

The farmers could achieve the better outcome, Gauthier argues, if A could make a sincere assurance to reciprocate to B, provided B helps him first. If B believes this, he will in fact help, in order to secure A's help in return. But, according to Gauthier, A can only make a sincere assurance that B will believe if he will take himself to have reason to follow through on the assurance when it comes to doing so. The problem here is that he will not take himself to have such reason if he takes his reasons directly from his preferences over outcomes at the moment of choice. In this kind of case, then, A does better with regard to his preferences, if he uses a deliberative procedure that requires of him to go through with his assurance, even if it means at times not choosing the act that he prefers.

In a nutshell, the specific deliberative strategy Gauthier defends in this kind of case is the following. When it comes to acting on an assurance, the agent should ask herself two questions. First, 'how well would I have done if I had never made any assurance'? And second, 'how well will I do if I act on the assurance'? If the agent judges she would do worse acting on the assurance than she would have done having never made one, then she is free to just maximize with regard to her preferences at each point in time. But otherwise, she should act on the assurance.

Let us assume that in the absence of an assurance, both agents are sophisticated and know this about each other. They thus know that in the absence of an assurance, they will each have to harvest their fields alone. Then Gauthier's deliberative strategy leads to the desired result in the intertemporal Prisoner's Dilemma, provided A has made the assurance. When

it comes to helping B, A judges that he will do better acting on his assurance than he would have done never having made an assurance at all.

Interestingly, if we simply substitute 'resolution' for 'assurance', this deliberative strategy can also be used to justify going through with resolutions to resist temptations in the kind of case presented above. Suppose the agent made a resolution to watch only one episode and then stop. At t_2, when it comes to following through, she considers what would have happened had she not made this resolution. Again, we assume that in the absence of a resolution, the agent is sophisticated. She would then not even have watched the first episode. That means, according to the agent's preferences at t_2, she would have done worse not having made a resolution than she would do acting on the resolution. Even then, she prefers only watching one episode to watching none. And then, according to Gauthier's deliberative strategy, she should follow through with the resolution.[6]

5 The Failure of Two-Tier Arguments

There is, however, a crucial difference between the temptation cases and the intertemporal Prisoner's Dilemma, and I want to argue that this shows that Gauthier's argument is unsuccessful in the case of temptation, even if it were

[6]While Gauthier (1996) argues in favour of extending the two-tier account to justify resolution in temptation cases, Gauthier (1997) in fact expresses some scepticism about this. The motivation Gauthier (1997) states for not extending the account to temptation is that the agent does not relate to herself over time as she does to other people. Part of the reason in favour of cooperation, Gauthier here claims, is that the agent views other people as 'ends in themselves'. But according to Gauthier, she does not view her previous selves in that way. Gauthier here seems to abandon our presupposition, and the presupposition he makes in "Assure and Threaten", that resolution is to be justified in terms of instrumental rationality alone. And so while I agree in the following that temptation cases are crucially different from the intertemporal Prisoner's Dilemma, the difference Gauthier himself points out is one that he cannot appeal to if he wants to offer a true instrumentalist two-tier account.

successful in the case of the intertemporal Prisoner's Dilemma. In the intertemporal Prisoner's Dilemma, each farmer's preferences over the possible outcomes of the game remain constant. These constant preferences can be used as instrumental standards by which to evaluate the deliberative strategy Gauthier proposes. The temptation cases are different in this respect. It is in fact a defining feature of temptation cases as we characterized them that the agent does not have constant preferences over outcomes. Under preference-based instrumental rationality, these same changing preferences form the standard of instrumental rationality.

For a two-tier account to apply, we need to identify a deliberative strategy that is best by the agent's own lights. The deliberative procedure Gauthier proposes results in the best outcome according to the agent's preferences at t_1. But it does not lead to the best outcome according to the agent's preferences at t_2. At t_1, the agent thinks that the best course of action is one whereby she watches only one episode and then stops. But at t_2, according to her preferences, the best course of action for the whole choice problem is the one where she watches the first episode and then goes on to watch another. According to the agent's preferences at t_2, a deliberative procedure that endorses this course of action would be best.

Gauthier's proposed deliberative strategy can endorse making a resolution to not watch a second episode and going through with it, as we have seen. But it would not equally endorse making a resolution to watch two episodes and going through with that resolution — which is the best course of action according to the agent at t_2. At t_2, the agent would have no problem going through with such a resolution, of course. But at t_1, the agent takes it to be better to have made no resolution at all than to act in accordance with it and watch the first episode. This is because, at t_1, she prefers watching no TV over watching two episodes.

We can, however, imagine possible alternative deliberative strategies that would allow the agent to make a resolution to watch both episodes and go through with that resolution. The agent at t_2 would prefer such a deliberative strategy. Gauthier's proposed deliberative strategy is thus not

the best deliberative strategy according to the agent's preferences at each point in time. It is the best deliberative strategy according to the agent's preferences at t_1. But it is not the best deliberative strategy according to the agent's preferences at t_2. Therefore, an argument that requires an agent's deliberative strategy to be best by her own lights in order for it to be rational to follow it does not go through. At the time when the agent is tempted, she does not think that a deliberative strategy that requires her to resist temptation is best. And so according to such an argument, she would not be rationally required to follow it.

Gauthier (1997) proposes a different deliberative strategy specifically for the context of temptation. There he notes that often, in cases of temptation, while the agent's proximate preferences for, e.g. watching a second episode, change, the agent retains 'vanishing point' preferences that still favour watching only one episode. These vanishing point preferences are preferences about how to choose in similar situations in the future. So even while the agent is tempted, she may prefer not to give into a similar temptation at future points in time. Let us grant that this is so in our TV consumption case. Even as I am tempted to watch another episode, I prefer that I only watch one episode at my coffee break the next day. Gauthier thinks that this makes it the case that the best deliberative strategy is one whereby the agent ignores her proximate preferences, but acts in accordance with the vanishing point preferences she holds at other times.

It is clear that, given her vanishing point preferences, the tempted agent judges that she will do much better by adopting a deliberative strategy that will make her resist temptation at all points in time than she would do if she adopted a deliberative strategy whereby she always gives into temptation. However, that does not make it the case that the agent takes the deliberative strategy of always going with her vanishing point preferences to be *best*. In particular, a deliberative strategy whereby she can make just this one exception would be preferred by the tempted agent. Gauthier's argument only goes through on the assumption that the agent is committed to adopting deliberative strategies that treat similar decision problems alike. However,

such a commitment is not required by instrumental rationality. Without any desire for such consistency, the agent could always formulate deliberative procedures that allow for exceptions that are indexed to a specific time or place.

At this point, we might want to make a two-tier argument at a higher level, to the effect that agents who don't allow themselves to make exceptions generally do better in life. But again, as long as the agent's shifted preferences are the standard of instrumental rationality, the best deliberative strategy at this higher level will be one that allows just this one exception to not making exceptions. The underlying problem for both of Gauthier's accounts is that, given preference-based instrumental rationality, as preferences change, the standard by which to evaluate deliberative strategies changes.[7] This is in fact fatal for any two-tier account. According to two-tier accounts, an action is rational if and only if it is endorsed by the best deliberative procedure. This approach shields the tempted agent's actions from being evaluated in terms of her shifted preferences directly. However, given preference-based instrumental rationality, the shifted preferences reappear at the higher level of deliberative strategies. And at the time of temptation, the agent is not only tempted, but would endorse a deliberative procedure whereby she would give into temptation.

6 Time-Slice Cooperation Arguments

Edward McClennen (1998) offers a treatment of temptation cases that is more explicit about the changing nature of the agent's preferences, which

[7] I am assuming here that preference-based instrumental rationality is about doing well by the preferences the agent actually holds at the time of action. It is not, e.g., about doing well by all the preferences the agent has ever held, or will ever hold, or about doing well by the preferences the agent has held and will hold within some smaller window of time. By doing so, I am rejecting a temporally extended view of the agent's interests. I do so for the same reasons as I reject interpreting McClennen's appeal to Pareto optimality between time slices as a two-tier account below.

makes it impossible for us to judge the benefits of a deliberative procedure against a single set of preferences. He still thinks an appropriate, unchanging instrumental standard for this context can be formulated, however. His instrumentalist argument is based on intertemporal, intrapersonal optimality instead, a standard he had already advocated in McClennen (1990) in a slightly different context.[8]

At first sight, his account may look like another two-tier account. The deliberative strategy McClennen defends as rationally called for under many circumstances is resolution. Let a plan be a set of choices, one for each decision node the agent could find herself at in a given decision tree. Under certainty, each plan has one outcome associated with it. A resolute agent considers which plan or plans she prefers most at the outset, adopts one, and then simply carries it out. McClennen thinks that there are instrumental advantages to resolution whenever it makes possible a series of choices that is judged at least as good or better by the agent at each point in time in the decision problem, than the alternative where she is sophisticated. That is, resolution can be justified by appealing to what we may think of as Pareto improvements between an agent's 'time slices': Resolution leaves some time slices better off and no time slice worse off.

Resolution in the temptation cases above indeed yields such an intrapersonal Pareto improvement. If the agent makes a resolution to only watch one episode and does not give into temptation, she ends up with O_1. If, instead, she is sophisticated and acts according to her preferences at each point in time, she ends up with O_0, as we have seen above. But at each point in time in the dynamic choice problem, she prefers O_1 to O_0. And so the resolute strategy is superior according to McClennen's criterion. And in fact, no further Pareto improvements are possible here, since there is no other outcome that is judged better by the agent at each point in time.

There is a substantive reason and a reason of argumentative strategy for

[8]Intrapersonal optimality had also already been discussed in the economic literature as a choice criterion for agents with changing preferences. See Peleg and Yaari (1973).

not interpreting McClennen's argument as a two-tier argument. The substantive reason is that intrapersonal optimality is implausible as a standard of instrumental rationality. This is because an agent need not care about her preferences at different points in time. But treating intertemporal optimality as a standard of instrumental rationality would make it non-optional for such an agent to cater to her past and future preferences. A requirement to cater to one's past or future preferences even if one does not care about them does not sound like a requirement of *instrumental* rationality (even if it may be a non-instrumental requirement of rationality). Instrumental rationality, I take it, is about doing well by the ends we actually hold at the time of decision.[9] Like Gauthier, McClennen himself claims to be in the business of establishing requirements of instrumental rationality. Under preference-based instrumental rationality as we understand it, if the agent did care about achieving intertemporal optimality in a way that is relevant for instrumental rationality, she would have ranked the Pareto optimal outcome most highly in her preferences. Given that we postulated that the agent does not rank resisting temptation most highly at the time of temptation, she thus does not sufficiently care about achieving intrapersonal, intertemporal Pareto optimality.

The strategic reason for not interpreting McClennen's argument as a two-tier argument is that, if we do so, everything I say below about abandoning preference-based instrumental rationality will apply to his argument thus understood. If instrumental rationality is also about catering to one's past and future preferences, then, if an agent's temporary preferences are in tension with such an intertemporal standard, as they are in temptation cases, they misrepresent the true standard of instrumental rationality — in

[9]It is sometimes assumed in the decision theoretic literature that choosing rationally consists in choosing well for your future self. Jeffrey (1965/1983) appeals to this idea when arguing for his version of evidential decision theory. Briggs (2010) uses it to analyze various decision theoretic paradoxes. And LA Paul (2015) presupposes this when she argues that rational choice is impossible when we can't know what our future attitudes will be. Of course most of us care to some extent how we will view our decisions in the future. But this is rarely all that matters for us, and it may matter to us in different ways.

which case we do not need a two-tier argument to explain why it could be instrumentally rational to resist temptation.

McClennen's appeal to optimality is thus not best understood as part of a two-tier argument. Instead, the best way to interpret his appeal to optimality is in analogy with the role of Pareto optimality in interpersonal choice problems like the Prisoner's Dilemma. In those games, there is no agent whose end it is to achieve Pareto improvements. It is simply the case that achieving a Pareto improvement serves both agent's preferences. This provides the basis for authors like Gauthier to argue for the rationality of decision rules that make cooperation possible. Each agent has a reason to do her part in making cooperation possible, because each agent stands to gain from it. McClennen suggests that analogously, in the temptation cases, the agent's 'time slices' can engage in mutually beneficial cooperation.[10] Adopting a choice rule that makes such cooperation possible is advantageous for each time slice.

7 The Failure of Intrapersonal Optimality Arguments

Regardless of the merits of the argument in the interpersonal case, this analogy ultimately fails. McClennen leaves it vague what time slices are and how they relate to the agent. But however we think of them, the analogy to interpersonal cooperation is suspect.

At one end of the spectrum, we could think of the time slices as separate agents that exist in succession (but presumably retaining memory of resolutions made by earlier time slices). Carrying out a resolute choice strategy now requires different agents to do their part: One needs to form a resolution, and the other ones need to carry it out. The problem on this

[10] Ainslie (1992) similarly suggests that willpower in the face of the preference reversals caused by hyperbolic discounting is the result of a kind of intrapersonal cooperation.

interpretation, apart from the implausible picture of agency it paints, is that the time slice at t_2 whose turn it is to resist the temptation is asked to act on a resolution that she did not make herself. She never made any assurance to the time slice at t_1 that she would resist the temptation, and had no say in the formation of the resolution. She could not have done so, since she was not around at the earlier points in time.[11]

And so, if this case resembles a case of interpersonal cooperation, it resembles one where a cooperative scheme is forced on an agent. In the farmer case, suppose that farmers A and B have not communicated at all. Farmer A harvests half his field alone and then takes a break. When he comes back, B has harvested the rest of the field for him. Even if A knows that B would only have done this had he expected A to return the favour, it does not seem instrumentally irrational of A not to return the favour. It might be nice to do so, or even called for by some social norms. But unless A cares about these social norms or about being nice, instrumental rationality seems to in fact require A to not help his neighbour in return.

At the other end of the spectrum, we could think of time slices as different stages of the same agent. In the temptation cases, this same agent merely changes her preferences over time. But in this kind of case, we usually simply assume that when an agent changes her preferences, she changes her mind, and the new preferences simply override the old preferences. In that case, there is no reason for the agent to still act on preferences she does not hold anymore. If the agent is cooperating with herself in temptation problems, as we are supposing, she is in fact cooperating with an agent who has changed her mind about the terms of cooperation. In interpersonal cooperation, at least, there appears to be no reason to make good on an assurance to your cooperator if doing so would not benefit her anymore, due

[11] Bratman (1995) objects to appeals to intrapersonal optimality on the basis that the earlier time slice is not around anymore once the later time slice makes a choice. The concept of cooperating with the dead, as it were, seems odd. I take that objection not to be entirely decisive. Gauthier's proposed deliberative strategy, at least, does recommend going through with an assurance to the dead.

to a shift in her preferences.

For instance, in the farmer case, suppose farmer A secured farmer B's help with an assurance. But just before it comes to reciprocating, B changes his preferences such that he now prefers harvesting alone after all. Perhaps he took a sudden dislike to A. It seems implausible that in this kind of case, there is anything to be said for A "helping" B. In fact it would be bizarre for A to impose his help against B's will. Likewise, it seems, in the case where the tempted agent is cooperating with herself, there is nothing to be said for catering to the agent's earlier preferences once they have changed.

The best way to think about time slices may lie somewhere in the middle. But two requirements would need to be met in order for the argument to resemble interpersonal cooperation. First, it would need to be the case that time slices are unified enough such that a later time slice recognizes a resolution made by an earlier time slice as her own. But they can't be so unified that the preferences of later time slices override the preferences of earlier time slices. I don't see how these two requirements could plausibly be met together.

8 Giving Up Preference-Based Instrumental Rationality

I have argued that the two most prominent kinds of instrumentalist arguments for the rationality of resisting temptations fail. Two-tier accounts fail because in temptation cases the standard by which to evaluate deliberative procedures shifts. And no plausible account of mutually beneficial cooperation between time slices of an agent can be given. However, my argument relied on the assumption of preference-based instrumental rationality. But this assumption may well be false. I now want to suggest that giving up preference-based instrumental rationality does not help those who want to make the instrumentalist arguments for resisting temptation we discussed.

If there is to be any hope of rational choice theory formulated in terms of preferences to serve as a theory of instrumental rationality, then preferences should at least normally or ideally stand in a close relationship to the true standard of instrumental rationality. This would be so, for instance, if we understood preferences as conative attitudes that act as a summary representation of the agent's underlying desires and concerns that form the true standard of instrumental rationality. Or it would be so if preferences were understood as dispositions to choose that are ideally responsive to the agent's underlying desires and concerns as a whole. Giving up preference-based instrumental rationality opens up the possibility that preferences express or represent this true standard of instrumental rationality incorrectly or incompletely. We may in fact suspect that this is the more appropriate analysis of temptation cases: Under the influence of some tempting situation, the agent's preferences shift to diverge from her underlying, true desires. As I want to argue here, however, conceding this does not help those who want to make the instrumentalist arguments we discussed.

Suppose that at any point in time, only one unique preference ranking can accurately capture the true standard of instrumental rationality. We can then distinguish two exhaustive possibilities of how the tempted agent's shifted preferences relate to the true standard of instrumental rationality. First, whether the agent's actual preferences correctly represent her underlying desires or not, the preferences that would do so are not stable. That is, the underlying true standard of instrumental rationality in fact shifts significantly over time. In that case, all the problems we discussed in the foregoing still arise, and the instrumentalist arguments still fail to establish the rationality of resisting temptation. The second possibility is that the agent's underlying desires would in fact only be correctly expressed by a stable preference ranking. In that case, the underlying true standard of instrumental rationality is in fact stable.

Thinking of temptation cases along the lines of this second possibility may in fact be what explained the intuitive instrumental irrationality of

giving into temptation all along.[12] The fact that the preference reversal in temptation cases is only temporary could be seen as evidence that tempted agents never stop having the goal of being temperate, but are only momentarily confused about what they really want.[13] However, in this case, it seems like we don't need the instrumentalist arguments we have been considering in the foregoing anymore. What is instrumentally rational is to do well by one's underlying desires and concerns. If the true standard of instrumental rationality, all the way through, uniquely supports only watching one episode, then, even as the agent is tempted to watch another episode, instrumental rationality requires her not to do so. This is so for straightforward reasons. As long as the agent refrains from watching the second episode, however she manages to do so, she is instrumentally rational. Moreover, the instrumentalist puzzle we started out with easily resolves: Agents are now at best instrumentally required to maximize with respect to their preferences if the preferences correctly capture their underlying desires — which we are supposing they don't in temptation problems.

If these are indeed the only two possibilities of how the tempted agent's shifted preferences can relate to the true standard of instrumental rationality in temptation problems, then there would be no use for the instrumentalist arguments we considered above. Either they fail, or they are redundant. I would, however, like to point to an interesting third possibility where a two-tier argument may again be of use. This possibility may arise if we allow for non-uniqueness in the sense that several different preference rankings express

[12]This is suggested, for instance, by Sarah Paul (2015) who claims that the stable, more long-term preferences an agent has before and after being tempted have a better claim to 'speak for the agent' (even at the time when she is tempted). Gauthier's (1997) argument that the agent should act on her 'vanishing point' preferences may also in part have been motivated by this intuition.

[13]However, note, too, that there may be cases where the agent's momentary preferences have a better claim to accurately representing what she truly cares about. This could be so, for instance, for a woman requesting an epidural when in labour despite an earlier, well-informed resolution not to do so. See Andreou (2014) for this example. One advantage of the view described here is that it could explain why, in these kinds of cases, instrumental rationality may demand giving into 'temptation'.

the true standard of instrumental rationality equally well. Suppose, then, that (1) there is at least one stable preference ranking that would correctly capture the agent's underlying desires at every point in time, but (2) at any point in time, it is also true that several different preference rankings would accurately capture the agent's underlying desires. For instance, suppose that throughout, the agent's underlying desires underdetermine whether she should have the preferences she does in fact have at t_1, or those she has at t_2 — both are permissible. As a matter of fact, the tempted agent has shifting preferences. But she could have stable preferences that would capture the true standard of instrumental rationality correctly at every point in time. This general possibility may arise both when the standard of instrumental rationality is stable, and when it shifts only slightly over time, so that permissible preference rankings overlap.

In these circumstances, a two-tier argument may actually give the agent reason to stick with one of the permissible preference rankings throughout, or to act as if she did. Adopting a deliberative strategy that demands this kind of stability in the face of non-uniqueness will keep her from ending up with an outcome that is definitely worse according to her underlying desires, such as the outcome of not watching any TV. While this is an interesting possibility, this seems to me to be a special case, and only some real life temptation cases will be accurately described by this analysis, if non-uniqueness is even a coherent possibility. And in any case, the kind of two-tier argument sketched here differs substantially from the ones typically presented in the literature on temptation cases.

9 Conclusions

If we stick to preference-based instrumental rationality, the instrumentalist arguments for resisting temptation fail. Only if we abandon it will we be able to give an instrumental argument for resisting temptations. There are in fact good independent reasons for abandoning preference-based instrumental

rationality, at least if we have in mind preferences as they feature in standard rational choice theory. For one, in ordinary speech we often take conative attitudes over features of outcomes to explain our preferences over outcomes: I may prefer O_1 to O_2 because I desire to get my work done, and I take this to outweigh my desire to watch TV. The kind of instrumental failure that may be involved in temptation according to this picture also seems familiar. I often find myself forming all-things-considered attitudes over my options that on reflection did not do full justice to everything I care about in those options. And then I take myself to be instrumentally criticizable. Lastly, I argue elsewhere[14] that standard instrumentalist arguments in favour of the core requirements of rational choice theory do not work on the preference-based picture.

Where does abandoning the preference-based picture leave us with respect to the rationality of resisting temptation? As we said above, if we can show that the agent's underlying concerns in fact are stable, and that the agent's preferences merely momentarily misrepresent this fact, then resisting temptation is instrumentally rational for straightforward reasons. But this response depends on the true standard of instrumental rationality in fact being stable, which may or may not be true, depending on the case. We can thus no longer give an argument that instrumental rationality *requires* that agents resist temptation as we characterized it. After all, instrumental rationality cannot demand that the agent have any particular ends. The best we can do, if we want to appeal to instrumental rationality alone, is to argue that agents ordinarily have desires that support resisting temptations in a wide variety of cases. Alternatively, we could make it a defining feature of temptation cases, properly understood, that the agent's underlying true desires in fact stably support not giving into temptation. If we do so, it must be clear that there is no more puzzle about how resisting temptation can be rational. The challenge then lies only in how agents can be motivated to do what is rational.

[14] See Thoma (2017).

References

George Ainslie. *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*. Cambridge University Press, 1992.

George Ainslie. *Breakdown of Will*. Cambridge University Press, 2001.

Chrisoula Andreou. Temptation, resolutions, and regret. *Inquiry*, 57(3): 275–292, 2014.

Michael Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.

Michael Bratman. Planning and temptation. In Larry May, Marilyn Friedman, and Andy Clark, editors, *Mind and Morals: Essays on Ethics and Cognitive Science*, volume 293-310. Bradford/MIT, 1995.

Michael Bratman. Toxin, temptation, and the stability of intention. In Jules Coleman, Christopher Morris, and Gregory Kavka, editors, *Rational Commitment and Social Justice: Essays for Gregory Kavka*, pages 59–83. Cambridge University Press, 1998.

Rachael Briggs. Decision-theoretic paradoxes as voting paradoxes. *Philosophical Review*, 119(1):1–10, 2010.

David Gauthier. *Morals by Agreement*. Oxford University Press, 1987.

David Gauthier. Assure and threaten. *Ethics*, 104(4):690–721, 1994.

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.

David Gauthier. Resolute choice and rational deliberation: A critique and a defense. *Nous*, 31(1):1–25, 1997.

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

Richard Holton. *Willing, Wanting, Waiting*. Oxford University Press, 2009.

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

Richard Jeffrey. *The Logic of Decision*. University of Chicago Press, 2nd edition, 1965/1983.

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

Edward McClennen. Rationality and rules. In Peter Danielson, editor, *Modeling Rationality, Morality, and Evolution*, pages 13–40. Oxford University Press, 1998.

Christopher Morris and Arthur Ripstein. Practical reason and preference. In Christopher Morris and Arthur Ripstein, editors, *Practical Rationality and Preference*. Cambridge University Press, 2001.

Robert Nozick. *The Nature of Rationality*. Princeton University Press, 1993.

L.A. Paul. *Transformative Experience*. Oxford University Press, 2015a.

Sarah Paul. Doxastic self-control. *American Philosophical Quarterly*, 52: 145–158, 2015b.

B. Peleg and M. Yaari. On the existence of a consistent course of actions when tastes are changing. *Review of Economic Studies*, 40(3):391–401, 1973.

Johanna Thoma. *Advice for the Steady: Decision Theory and the Requirements of Instrumental Rationality*. PhD thesis, University of Toronto, 2017.

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