Philosophia
https://doi.org/10.1007/s11406-024-00801-w

Rationality, Choice Theory, Self-Torture, and Coherence: On Chrisoula Andreou's *Choosing Well*

Mark Budolfson[1]

Received: 17 October 2024 / Revised: 17 October 2024 / Accepted: 18 November 2024

Abstract
If one rejects choice theory, one still faces the challenge of explaining what one should do in a self-torture case if one is fully rational, and why. In addition, one faces the more fundamental challenge of explaining what rationality is, if it is not choice theory. Responding to these challenges in a systematic way is one of several key contributions of Andreou's insightful book *Choosing Well*, and is my focus here as well. My goal is to highlight questions that arise for Andreou's view, especially from the perspective of those who are sympathetic to choice theory.

Keywords Rationality · Choice theory · Self-torture · Coherence · Chrisoula Andreou · Practical reason

1 Self-Torturer Cases, and the Argument Against Choice Theory as a Theory of Rationality

Standard rational choice theory (hereafter: choice theory) implies that an agent's preferences must be acyclic. Some critics of choice theory use self-torturer cases to argue that cyclic preferences can indeed be rationally permissible, which implies the revisionary conclusion that rationality cannot be identified with choice theory.

Thanks to Chrisoula Andreou, Tristram McPherson, Dean Spears, Sergio Tenenbaum, Patrick Wu, and Cayla Xue for helpful comments.

✉ Mark Budolfson
mark.budolfson@austin.utexas.edu

[1] Department of Philosophy, Department of Geography and the Environment, Population Wellbeing Initiative, University of Texas at Austin, Austin, USA

Published online: 26 December 2024

Self-torturer case

An electric device with 1,001 settings (0, 1, 2, 3,. . ., 1,000) is attached to one. Raising the setting of the device increases the amount of electric current running through one's body, starting at setting 0 with no current and no pain, and increasing to excruciating pain at setting 1,000. For any two adjacent settings on the machine, n and $n+1$, one cannot, with any confidence, distinguish between the settings just by the way one feels; but one gets $10,000 at each advance. One starts at setting 0, and then once a week one can compare all the different settings and then, if one chooses, advance one setting, but one can never permanently return to a lower setting.[1]

For example, Chrisoula Andreou offers a particularly clear statement of this revisionary argument, by endorsing two claims that together imply rational cyclic preferences: first, that it is rational to prefer $n+1$ to n (for every choice point n), and, second, that it is rational to prefer point 0 to 1,000. For this and other reasons, Andreou reaches the revisionary conclusion that rationality is not choice theory, partly on the grounds that cyclic preferences are indeed rationally permissible in self-torturer and other cases.

If one rejects choice theory in this way, one still faces the challenge of explaining what one should do in a self-torture case if one is fully rational, and why. In addition, one faces the more fundamental challenge of explaining what rationality is, if it is not choice theory. Responding to these challenges in a systematic way is one of several key contributions of Andreou's insightful book *Choosing Well*, and is my focus here as well.[2] My goal is to highlight questions that arise for Andreou's view, especially from the perspective of those who are sympathetic to choice theory.

2 Andreou's Revisionary Alternative to Choice Theory

At the core of Andreou's positive view are her proposed additional principles of rational choice that, together with her endorsement of rational cyclic preferences, yields a revisionary alternative to choice theory. On Andreou's view, these additional principles are:

Andreou's additional principles of rational choice

1. There exist rankings of *categories* of options in addition to preference rankings of the options themselves.
2. It is a principle of rationality that one cannot rationally choose options from within a category that one ranks below another category (among the most relevant categories, and insofar as one has a non-cyclic ranking of categories).[3]

[1] Self-torturer cases originate from Quinn, 1993; this formulation is from Andreou, 2023, pp. 19–20 and 49.

[2] Andreou, 2023, pg. 50.

[3] See principle P3* on pg. 61 of Andreou, 2023, and surrounding discussion; see also Chap. 4.

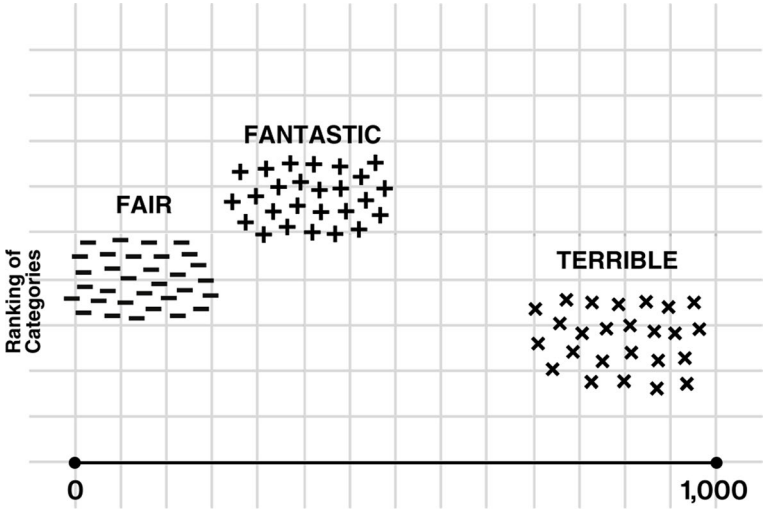

Fig. 1 Ranking of categories in a self-torturer case (which might have vague boundaries)

3. Thus, there is a rational requirement to choose within a category of good options over choosing within a category of bad options in cases like the self-torturer cases, and other cases where rationally incoherent preferences would make one susceptible to a money-pump, etc.[4]

Figure 1 helps illustrate Andreou's principles here via the self-torturer case:

If one follows Andreou's novel additional principles of rationality above, the idea is that one will end up choosing some option x that lies within the FANTASTIC category, even though (according to Andreou's theory of rationally permissible cyclic preferences) there is another option x + 1 that one prefers to x. According to Andreou, this result is exactly as it should be, as it allows us to:

> …make sense of the idea that, in cases involving rationally cyclic preferences, instrumental rationality sometimes calls for showing restraint so as to avoid realizing a self- defeating pattern of choice. For, when preferences are rationally cyclic, we can interpret "showing restraint" as "settling on a dispreferred option" and interpret the demand that one avoid realizing a self- defeating pattern of choice as requiring (perhaps among other things) not that one avoid realizing a dispreferred option, since that is not possible, but that one avoid ending up in an unnecessarily low appraisal category (e.g., ending up with a terrible option when great options are available). (Andreou, 2023, pp. 49).

[4] See Andreou, 2023, Sect. 3.5 for discussion of money pump cases and these principles, and Chap. 4 for discussion of incompleteness and these principles. For general discussion of money-pump arguments, see Gustafsson 2022. On Andreou's view, "It is the combination of cyclic preferences and non-cyclic categories that makes [self-torturer cases] particularly interesting" (Andreou, 2023, pg. 60).

In this way, Andreou's view aims to be revisionary but not radical – revisionary in that it rejects the identification of choice theory with rationality, but not radical because the basic core mechanisms of preference, etcetera described by choice theory still have a central place in Andreou's theory of rationality, but are supplemented by other forces that help shape rationality, all in a way that promises to deliver intuitively reasonable verdicts on cases.

3 The Response on Behalf of Choice Theory

One key question is whether there is a compelling response to self-torturer cases that can be offered on behalf of choice theory. In fact, choice theory has a fairly straightforward response, as has been noted by others (for example, Arntzenius & McCarthy, 1997, and Elson, 2016). Here I'll present this response in a particular way, partly to clarify its relevance to evaluating arguments for alternative views such as Andreou's, and partly to emphasize that it relies only on basic ideas that are familiar from discussions about the regulation of social harms such as pollution.

The response begins by agreeing with the description of the case such that one is progressively better off in cases immediately to the right of the initial point 0, culminating in some FANTASTIC range, and that one is ultimately worse off at point 1,000 than at point 0, and is progressively worse off over the points immediately leading up to point 1,000, and agreeing that the only factors driving those differences in personal good are pain, and money (which can be assumed to change by the same unit in each step, to simplify the case). Choice theory then insists that the simple resolution of all of this is that one has some value function over pain and money, such that initially the marginal value of an additional unit of money exceeds the marginal disvalue of an additional unit of pain, and where ultimately this reverses and the marginal value of an additional unit of money is less than the marginal disvalue of an additional unit of pain, as in Fig. 2:

From this starting point, choice theory extends this reasoning to argue that there must be an optimal peak that maximizes personal good given one's values, modulo uncertainty and vagueness: because one must have a value function over pain and money such that initially the marginal value of an additional unit of money exceeds the marginal disvalue of an additional unit of pain, and where ultimately this reverses and the marginal value of an additional unit of money is less than the marginal disvalue of an additional unit of pain (from above), it follows from this and the description of the case that at some point there is an 'optimum' at which this value peaks, and then subsequently reverses. To illustrate how this might realistically work: consider (i) a linear relationship between units of pain and units of disvalue[5], together with (ii) a familiar diminishing marginal value function for money, and (iii) the basic details of the case described above; then: initially (at point 0) the value of an additional unit of money ($10,000) exceeds the disvalue of an additional unit of pain, but each sub-

[5] We might stipulate this as part of the setup of the case to keep things simple – i.e., we might stipulate that e.g. the technician setting up the self-torture simply calibrates the units of pain to match the subject's values so that the result is a linear function from pain to disvalue.

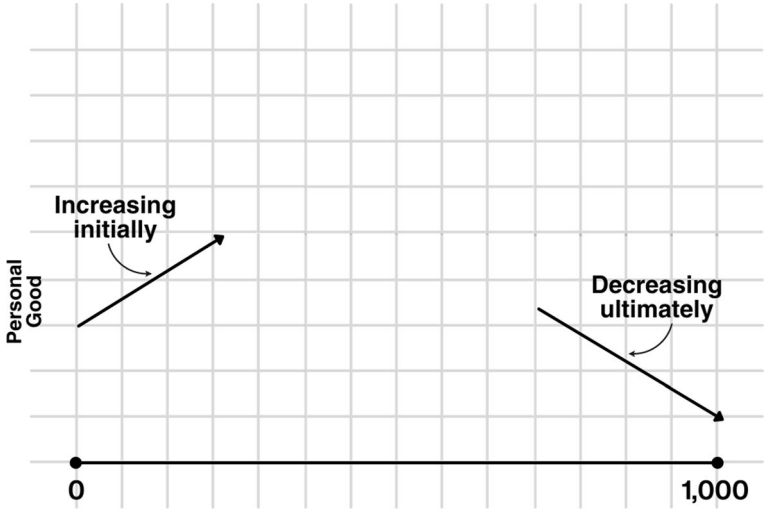

Fig. 2 Personal good is initially increasing, and ultimately decreasing in self-torturer cases

sequent $10,000 transfer has less additional value than the one before, until at some point the value of an additional $10,000 falls low enough that for the first time its value is exceeded by the disvalue of an additional unit of pain.

Choice theory says that the point immediately prior to the point just described is the optimal option that a rational agent must choose (which we can call **optimum**), modulo uncertainty and vagueness, illustrated by Fig. 3:

If this reasoning seems quick at first glance,[6] note that this is essentially the same reasoning used in mainstream economics to explain e.g. the optimal level of pollution in a pollution tax regime – i.e. where 'social good' replaces 'personal good', and we imagine a pollution tax regime in which allowing some initial amount of pollution is net beneficial to society, but as pollution increases there comes a point (the optimum) at which more units of pollution would for the first time become net costly to society, and so we aim to tailor the tax to incentivize the optimum level of pollution.[7] To see that the self-torturer case has the same structure as this textbook environmental regulation case, it may be useful to imagine that the self-torture is delivered not by electric shock, but instead by an initially slightly irritating but ultimately excruciating-at-high-concentrations environmental pollutant that increases at each step from which the tortured agent cannot ever escape. This case is then simply an individualized version of the textbook economic example of harm from pollution. Defenders of choice theory can then argue that just as there is no reason to doubt the textbook analysis of pollution regulation, so too there is no reason to doubt the analysis offered by choice theory in the individual case, both of which imply the conceptual importance of an optimum, modulo uncertainty and vagueness.

[6] See for example Andreou, 2023 pg. 20, especially footnote 7.

[7] See for example Tietenberg, 2018.

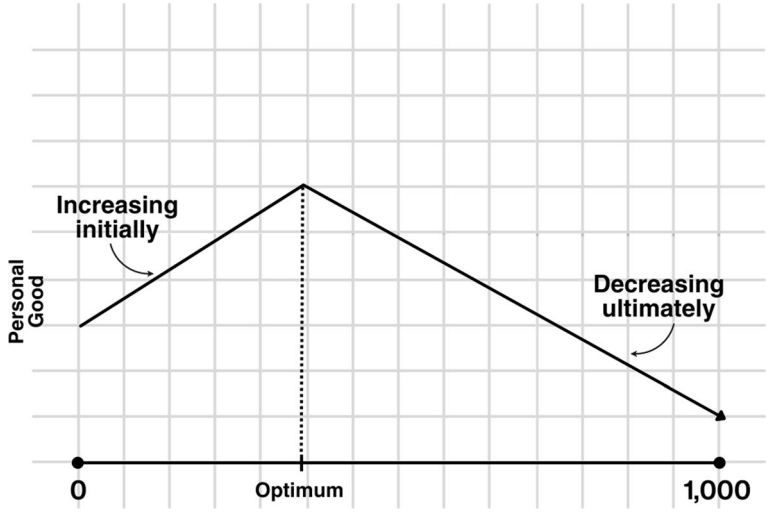

Fig. 3 Optimum is the best amount of self-torture according to choice theory – i.e. it is the optimal balance between more money and more pain given one's values

This leads to a diagnosis of the problem with the argument against rationality as choice theory based on self-torturer cases: the argument depends essentially on two premises: first, that it is rational to prefer $n+1$ to n for every point n, and, second, that it is rational to prefer point 0 to 1,000 (where these two premises together imply a preference cycle). The diagnosis offered on behalf of choice theory is that the first premise is incorrect, since if one is rational there must be some optimal point from which it is not rational to prefer the succeeding option.[8]

If one endorses this response, then choice theory also provides a straightforward explanation of the categorial rankings discussed above that Andreou appeals to in formulating additional principles of rationality, as shown in Fig. 4:

In this way, defenders of choice theory might insist that they have not merely an attractive response, but actually a more attractive explanation of the phenomena in the self-torturer case and our considered judgments about the rationality of choice. From this perspective, if it is also theoretically costly to reject choice theory, it may then seem that self-torturer cases do not give us any reason to prefer an alternative view such as Andreou's, given that choice theory also explains the appeal of Andreou's proposed additional principles.

A question for Andreou is what to say to those who find this response attractive?

[8] See Arntzenius & McCarthy, 1997 for further discussion, and Elson, 2016.

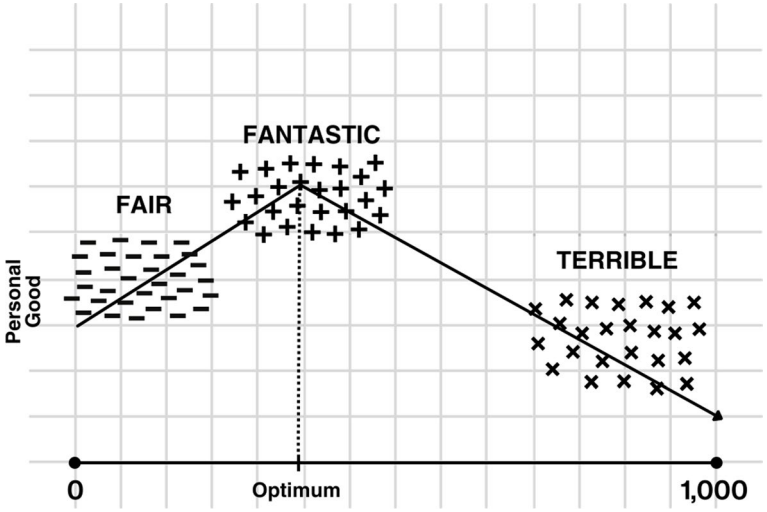

Fig. 4 A coherent ranking of options also explains the ranking of categories

4 A Further Challenge to Choice Theory: The Single-Shot from Optimum Case

We can call the reasoning in the previous section 'the response on behalf of choice theory'. Even if we agree for the sake of argument with all of its claims about the self-torturer case, including that there is an optimum, a further challenge to choice theory as rationality arises based on the following kind of case, which is related to cases discussed by Sergio Tenenbaum and Diana Raffman, and by Andreou:[9]

Single-Shot from Optimum case

Suppose one knows that one is in a self-torturer-like situation, except here the technician in charge of setting up the situation has kindly arranged for one to have only a single choice: one knows one will begin exactly at the optimum in the sense described in the previous section (with all of its associated costs and benefits). The only choice one will have is a single shot choice whether to advance one point further to optimum + 1, in which case one will receive an additional $10,000 and experience greater pain of one increment as in the original self-torturer case. One is certain of all of these facts, including that optimum is the optimum in the sense of best satisfying one's overall preferences and values.

In this case, it might seem rationally permissible to prefer optimum + 1 to optimum, even knowing everything about the choice situation – because even if one supposes

[9] Tenenbaum, 2020, especially pp. 93–4, and for the original, Tenenbaum & Raffman, 2012, pp. 94–95. For Andreou's discussion, see Andreou, 2023, pp. 124–128.

that that optimum + 1 is worse for one than optimum, it is worse only in ways that appear insignificant, whereas optimum + 1 is clearly better in a way that appears significant, namely that one has an extra $10,000. Generalizing, this shows how one might agree with all of the claims of choice theory in the previous section (including that there is an optimum), while still arguing that it is always rational to prefer $n+1$ to n.[10] This reasoning also yields a view consistent with Andreou's arguments that even if the better than relation is transitive and complete, preference cycles are nonetheless sometimes rational.[11]

This helps us focus on a key question: on a view like Andreou's, what ultimately explains the rationality of the preference for $n+1$ to n for every point in the self-torturer sequence (especially in a case like Single-Shot from Optimum where one knows that this will make the outcome worse for oneself)? This is a key question on which the plausibility of the argument against choice theory seems to turn. One might appeal to a principle like the following:

Dominance of Significant Effects

If Y is significantly better than Z along some dimension, and Y is not significantly worse than Z along any other dimensions, then it is rational to prefer Y to Z.

Many will object to this principle on the grounds that it is just the sort of thing that would license an irrational set of choices when there is a sorites-like series of insignificant effects that add up to something terrible (e.g. in self-torturer cases).[12] But if one is already in the spirit of accepting that there can be rational cyclic preferences, including in self-torturer cases, then perhaps one would not be overly concerned about this implication – especially if, like Andreou, one has a positive theory about how a rational agent will via other principles ultimately avoid torturing oneself and land in the FANTASTIC cases.

One question is whether Andreou endorses Dominance of Significant Effects as the more fundamental explanation of the claim that it is rational to prefer $n+1$ to n for every point n in the self-torturer sequence? Or is there a different explanation?

Another complication that Andreou notes is that insofar as she agrees that both (i) optimum is better for one than optimum + 1, and (ii) it is rational for one to prefer optimum + 1 to optimum, this might seem in tension with a principle she endorses called the **Inadvisability Condition**,[13] which is basically the principle that it is irrational to prefer an outcome that one knows and judges to be worse for oneself.

[10] Note that a similar 'one-shot from alleged stopping point' case can be used to challenge any allegedly rational unique stopping point, including e.g. a point n selected via a plan at the initial point in the sequence as a place to stop, etc.

[11] See Andreou, 2023, Chap. 6.

[12] See Elson, 2016.

[13] Andreou's Inadvisability Condition is: X is better than Y (as an option for A) implies that it is rationally inadvisable (for A) to choose Y from any finite set (of alternatives) that includes both X and Y (where the choice of an option counts as rationally inadvisable if rationality advises against choosing that option, all-things considered, given the context of choice). (Andreou, 2023, p. 114).

Andreou provides further discussion and arguments that (i) and (ii) are not ultimately in tension with the Inadvisability Condition, which involve further complexities.[14]

5 Does any of this Really Help One Make Rational Choices in Self-Torturer Cases? Commonsense Advice for Self-Torturers

If you or I are ever thrown into a self-torturer case, I'm not sure that anything that anyone has ever written will be of much use. The problem is that none of the guidance anyone has offered seems actionable given our perceptual and imaginative limitations. For even assuming that you can experience each and every one of the pains associated with each setting before making any choices, that isn't sufficient to know what it is like to *live with* that unprecedented pain, and furthermore it isn't sufficient to know how that disvalue would compare with the value of having an unprecedented amount of additional money (in a context where you are also burdened by unprecedented pain). So, even if one can experience the pain of each setting at the outset, that isn't enough information to know at the outset where the FANTASTIC cases are.

With this in mind, it does not seem useful to be told anything of the form:

Existing Guidance on the Self-Torturer Case

[Make a plan and stick to it to:]

Choose an option [in the neighborhood of] the optimum.[15]

This is not useful guidance because at the initial point one will have little idea where in the self-torture sequence the optimum is, and subsequently one will only imperfectly and gradually accumulate evidence about the location of the optimum as the sequence unfolds. And even when one reaches the optimum, it will remain unclear whether one is at optimum, or optimum-1, or even whether one is in the immediate neighborhood of the optimum. To make this vivid, try to imagine that you are actually put into a self-torturer case, and you are starting from point 0 - that's where you are now, right there at the beginning, and you're allowed to experience how all of the settings feel. Now ask yourself: can you at this point identify a point in the sequence that is close to the optimum, while you are sitting there at point 0 having never made a choice or lived with any of the unprecedented pain and wealth yet? Can you even

[14] See Andreou, 2023, pp. 124–128.

[15] This schema includes Andreou's guidance to choose an option within FANTASTIC (aka the neighborhood of the optimum), as well as the abstract guidance of choice theory, as well other existing recommendations such as making plans on where in the sequence to stop and sticking to them. For the latter, see for example Quinn, 1993, and Bratman, 1998 which adds a carefully formulated 'no regrets' condition (see also Bratman, 2007). Bratman's view may have an advantage in a specific way that this section brings out: in particular, one may learn as the sequence proceeds that e.g. the pain is not as bad as one initially thought, and that the location of optimum is further to the right than one initially thought; presumably one should be able to update one's plans in light of this new information, and the fact that it is an update that satisfies something like the Bratmanian 'no regrets' conditions is a plausible explanation vs. a more rigid version of the view that one always has to stick to plans once one has made them.

identify a point that you are confident is determinately in the neighborhood of the optimum? Is it point 15? How could you possibly know, given that you haven't lived with the choices in the way that will be your way of knowing? One would have to live into the pain and wealth first, and only sequentially gain gradually more and more confidence about whether increased settings continue to be worth it.

In light of this, the idea that you could make a wise plan at the outset to stop at some particular point seems like a nonstarter. And even once you reach the optimum, if you don't know that you're at the optimum, or even whether you're in the neighborhood of the optimum (and thus whether you're in the FANTASTIC cases), then the structural advice from both choice theory and Andreou's theory are not super helpful either. If you can't identify the neighborhood of optimum with confidence until you are past at least most of it, then the existing kinds of advice are unhelpful.

At the same time, this points the way toward some genuinely actionable and commonsensical guidance, which focuses on (i) the fact that one will be better off initially at each step of the sequence but then ultimately be worse off with each step later in the sequence, and (ii) that realistically one won't be able to tell that one is at the optimum when one is there. Given this, the guidance I have in mind is:

Commonsensical Guidance on the Self-Torturer Case

One should continue to increase the setting by one increment until one can tell that one is past the optimal point, and then stop.

Given perceptual and imaginative limitations, this advice may be the best realistic procedure to follow to maximize expected utility, to choose a FANTASTIC option, and/or to make a rational plan and stick to it.[16]

I don't see this guidance as a revolutionary insight, but rather as a bit of commonsense that helps remove what might otherwise seem like a lingering mystery about how one should actually decide what to do in these self-torturer cases, especially once one sees the lack of any actionable guidance from existing theories. The two key points here are that the best strategy in self-torturer cases (i) does not necessarily involve choosing a particular point to stop at the outset, and (ii) might require a lot of learning by going through much of the sequence before one acquires enough evidence to know when to stop. Both of these points seem to me an advance versus previous analyses of self-torturer cases, which tend to assume otherwise. And both of these points should be accepted, I think, regardless of whether one accepts or rejects the response on behalf of choice theory in the previous section.

[16] One's perceptual and imaginative limitations are a form of uncertainty, making the self-torturer case analogous in some ways to the following case of simple uncertainty: One knows that a number has been secretly and randomly drawn today between 1 and 1,000. Each day n, starting tomorrow with day 1, one will have the option of either 'staying in' (in which case if the number drawn today is greater than n, one receives an additional $10,000, otherwise one loses $100), or 'stopping', in which case one keeps all previous winnings and the game ends. In this case, one knows that there is some optimal point, but one doesn't know enough about where it is or even the neighborhood of where it is even as the sequence unfolds, even as one becomes more confident about its likely location. As a result of this ignorance, the rational strategy as the sequence unfolds remains simply to stay in until one can tell one is just past the optimal point, and then stop.

Despite being commonsense guidance, these points nonetheless raise a potential problem for Andreou's view. That's because on Andreou's view, it is rationally permissible to choose options in the lefthand side of FANTASTIC, in the sense of points in Fig. 4 in the left side of the set that constitutes FANTASTIC (since the essence of Andreou's view seems to be that it is rationally permissible to choose any point within FANTASTIC). However, it is difficult to see how one could know (have sufficient confidence) in self-torturer cases that one is within FANTASTIC at points in the lefthand side of FANTASTIC, for the reasons given in this section, even if it is possible to know that one is in FANTASTIC at some point on the righthand side where one can first tell that one is past the optimum (if one stipulates that that point is in FANTASTIC, which it must be if one is to be able to identify any of the FANTASTIC cases on Andreou's own view). So, one could argue that Andreou's view actually misclassifies some choices as rationally permissible that are in fact rationally impermissible: namely, cases where one stops on the far lefthand side within FANTASTIC in a self-torturer case while one is still on the upslope of personal good, and (most importantly) before one is in a position to know (have sufficient confidence) that one is in FANTASTIC.

6 Is there a Fundamental Theory of the Nature of Rationality that Explains all of this?

The preceding summarizes just a few of the many places where Andreou's book brings important issues into better focus, and advances our understanding of the theoretical frontier. As noted, Andreou's view aims to be revisionary but not radical – revisionary in that it rejects the identification of choice theory with rationality, but not radical because the basic core mechanisms of preference et cetera described by choice theory still have a central place in Andreou's theory of rationality, and are supplemented by other forces that help shape rationality on Andreou's view, and promise to deliver intuitive verdicts on cases. A final question is whether there is a plausible high-level theory of rationality that would explain all of this, bring it all together, and make us think that a revisionary view like Andreou's is exactly what we should expect to be true once we understood the fundamental high-level general nature of rationality?

As some relevant background, following Gilbert Harman, we might sharply distinguish principles of rationality, on the one hand, from mathematically precise formal systems that describe coherence conditions for beliefs or choices, such as propositional logic, the probability calculus, bayesian updating, and choice theory. Once we have made this distinction, we can more clearly ask what exactly is the relationship (if any) between principles of rationality and these formal systems?[17] One simple view is to say that rationality is simply to have beliefs or choices that satisfy principles of coherence described by these formal systems – as is claimed by e.g. subjective bayesianism about the rationality of belief, and standard rational choice theory about

[17] Harman, 1973; Harman, 1999. See also Pollock, 1999; Pollock, 2006. Much of this last section outlines issues first highlighted by Harman and Pollock.

the rationality of choice. However, following Harman and Andreou, we might instead insist that contrary to this simple view, there are cases where beliefs or choices are rational even though they violate principles of coherence as given by these formal systems. This is the characteristic claim of a revisionary theory of rationality. But what explains why we should think that rationality would be revisionary in this way?

One attractive fundamental picture that would explain a revisionary approach is that our ends (epistemic and/or practical) would be best served by (among other things) coherent beliefs/preferences if we had no limitations of time, attention, cognitive processing, and the like,[18] because perfect coherence is part of the fundamental explanation of how ideally perfect cognition and decision-making would proceed on a 'frictionless plane' to promote ends. But on this picture, rationality is essentially a concept for real agents who always have limitations – and so in this sense rationality is inherently a concept of non-ideal theory. And when limitations are present, the same ends are best served by beliefs/preferences that are sometimes less than fully coherent. Nonetheless, coherence remains as one of the most fundamental forces that should guide our beliefs and choices over time, akin to a fundamental force of rational gravity.[19]

On this picture, there is a fundamental explanation why, based on the nature of rationality, we should expect a correct theory of rationality to be revisionary (it should reject the identification of instrumental rationality with choice theory, and the identification of the rationality of belief with e.g. standard bayesianism in epistemology), but not radical (because at the same time mathematically precise and coherent theories such as choice theory describe some of the most fundamental forces relevant to rationality). So, on this view it could be argued that the sort of revisionary-but-not-radical view that Andreou and others are attracted to is exactly what we should expect from first principles of the nature of rationality. Thus, coherence as described by choice theory is indeed a fundamental force of rationality that is always relevant, but not in every case a necessary condition for rationality. In this way, choice theory retains a fundamental role.

A final question for Andreou is whether she is attracted to this fundamental picture of rationality, or perhaps an alternative one?[20]

Data availability There is no data associated with this article, as no data is used or reported in the article.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

[18] Perhaps we would add discrimination of borderline cases, depending on view of vagueness, and in what sense it is or is not in play in self-torturer cases.

[19] Some of Harman's arguments suggest that a plausible theory of rationality might maintain that synchronic coherence is a more powerful force than diachronic coherence (even if both are important forces) (Harman, 1999).

[20] An attractive alternative theory is provided by Tenenbaum, 2020, which is a revisionary view that departs more sharply from choice theory than the theory of the nature of rationality that I sketch in this section.

References

Andreou, C. (2023). *Choosing Well*, Oxford University Press.
Arntzenius, F., & McCarthy, D. (1997). Self torture and group beneficence. *Erkenntnis*, *47*, 129–144.
Bratman, M. (1998). Toxin, temptation, and the stability of intention. In J. Coleman, & C. Morris (Eds.), *Rational commitment and social justice* (pp. 141–159). Cambridge University Press.
Bratman, M. (2007). Temptation revisited, in *Structures of agency* (pp. 95–108). Oxford University Press.
Elson, L. (2016). Tenenbaum and Raffman on vague projects, the self-torturer, and the sorites. *Ethics*, *126*, 474–488.
Gustafsson, J. (2022). *Money-pump arguments*. Cambridge University Press.
Harman, G. (1973). *Thought*. Princeton University Press.
Harman, G. (1999). *Reasoning, meaning, and mind*. Oxford University Press.
Pollock, J. (1999). *Contemporary theories of knowledge*. Rowman & Littlefield.
Pollock, J. (2006). *Thinking about acting*. Oxford University Press.
Portmore, D. (2011). *Commonsense consequentialism*. Oxford University Press.
Portmore, D. (n.d.). What's a rational self-torturer to do? *PhilPeople*. Retrieved from https://philpeople.org/profiles/douglas-w-portmore
Quinn, W. (1993). The puzzle of the self-torturer. *Philosophical Studies*, *59*, 79–90.
Tenenbaum, S., & Raffman, D. (2012). Vague projects and the puzzle of the self-torturer. *Ethics*, *123*, 86–112.
Tenenbaum, S. (2020). *Rational powers in action*. Oxford University Press.
Temkin, L. (2011). *Rethinking the good*. Oxford University Press.
Tietenberg, T. (2018). *Environmental and natural resource economics* (11th ed.). Routledge.
White, S. (2015). The problem of self-torture: What's being done? *Philosophy and Phenomenological Research*, *94*, 584–605.