Draft. Forthcoming in *Noûs*. Please contact me before quoting or citing.

Rationality and second-order preferences

Alejandro Pérez Carballo
University of Massachusetts, Amherst

Can I most prefer to have preferences other than the ones I currently have?

Start by distinguishing between first and second-order preferences. My preference for chocolate over vanilla is a first-order preference. My preference for preferring vanilla over chocolate is a second-order preference. Can I have the second-order preference to have first-order preferences other than the ones I have?

The unwilling addict, we are often told, has such a pattern of preferences.[1] He prefers to get his fix over not getting it, but prefers not to have those preferences—he wants to get his fix, but would rather not want to. But we needn't look to addiction in order to find examples of agents that seem to prefer having preferences other than those they actually have. As things stand, I prefer my coffee with some sugar in it. But I wish I didn't want sugar in my coffee. If I could choose my preferences—at least when it comes to my coffee drinking habits—I would pick a set of preferences on which, among different coffee options, coffee without sugar was ranked at the top.

Assume, then, that it is possible to have such a pattern of preferences. Ask now: can such a pattern of preferences be *rational*?

1 Preliminaries

Before moving on, it is worth making explicit a few assumptions I will be making throughout.

1 Cf. Frankfurt 1971, Jeffrey 1974, *inter alia*.

First, I will assume that preferences are defined over propositions. For the sake of readability, I will sometimes talk about preferring ϕ-ing to ψ-ing. But this should be understood as shorthand for the claim about my preferences, where these are defined over propositions.[2]

Second, I will be working with a broadly Bayesian picture of the mind. An agent's state of mind, for the purposes of assessing her rationality, can be represented as a pair consisting of a *credence* function—an assignment of numerical values to a collection of propositions—and a *utility* function—an assignment of numerical values to propositions relative to a given possible world. An agent's preferences, on this picture, supervene in familiar ways on her credence and utility functions (more on this below).

Third, the notion of rationality in play throughout this essay is a relatively thin one in that I'll mostly be concerned with so-called formal or coherence requirements—I will accordingly sometimes call it *minimal* rationality. The particular assumptions I will be making will emerge as we go along, but to anticipate: I will assume that rationality requires that our credences satisfy the axioms of the probability calculus, that our actions be governed by familiar decision-theoretic norms, and that knowledge satisfy a plausible closure condition. This is not meant to be a definition of this notion of rationality. It is an open question whether there are any additional coherence requirements on rational agents.

With these assumptions in place, let me restate our question: is formal coherence compatible with a mismatch between one's first- and second-order preferences?

Thus understood, the answer to our question appears to be *yes*. While an agent whose first- and second-order preferences do not mesh with one another may fail to be ideally rational in some more substantive sense (say, in the sense of fully responding to the reasons she has), such a pattern of preferences does not seem to be ruled out by coherence requirements alone. Indeed, the assumption that a mismatch between one's first- and second-order preferences is compatible with minimal rationality seems

2 Where ϕ and ψ are action tokens, there is no easy way of stating that I prefer the proposition expressed by 'I will ϕ' to the proposition expressed by 'I will ψ'. (To my ears, 'I prefer that I will eat vanilla to that I will eat chocolate' sounds awful.) Thus I will often talk about preferring ϕ-ing to ψ-ing, and preferring most preferring ϕ-ing to most preferring ψ-ing.

to be presupposed in much work on the notions of valuing, autonomy, and freedom of the will.

Much work on second-order preferences was motivated by a desire to capture what a variety of philosophically interesting examples had in common. One such example is the unwilling addict, 'helplessly violated by his own desires', who prefers smoking over not smoking, but who 'hates his addiction and always struggles desperately, although to no avail, against its thrust'.[3] The psychology of such an addict, I suppose, exhibits a very complex structure. But one plausible starting observation, with an eye towards characterizing her psychology, is that she does not endorse those preferences: in some sense, she does not identify with the preferences that, as it were, 'happen' to govern her actions. These notions of 'endorsement' and 'identification' have been used to try to characterize free or autonomous action. And on many such accounts, second-order preferences play a prominent role.

On Harry Frankfurt's well-known view, free action requires that one's preferences be accompanied by a preference to have those very preferences.[4] And Richard Jeffrey, following Frankfurt, suggests we think of Akrates, his akratic agent of choice, as someone who acts in accordance with first-order preferences he would most prefer not to have: part of the appeal of this account in terms of higher-order preferences, according to Jeffrey, is that it is supposed to mesh well with a decision-theoretic picture of rationality:[5]

> [T]he higher-order theory does countenance various other failings—or misfortunes, or conflicts, or tensions, or 'contradictions' in some Hegelian sense. It gives us a canvas on which to paint some very complex attitudinal scenes, from life. That one cannot paint intransitive preference rankings on that canvas makes it all the more interesting that one can paint poor Akrates there, in the various postures we have seen above.

3 Frankfurt 1971, p. 12.

4 Frankfurt 1971. Frankfurt's account isn't quite this: the claim is that free agency requires that one acts based only on desires that are accompanied by a second-order *volition*: a desire that one has that desire and that that desire be 'effective'. Here, I follow Jeffrey 1974 in reformulating Frankfurt's views in terms of preferences.

5 Jeffrey 1974, p. 391.

Admittedly, few now would think that the presence of the relevant second-order preferences is *sufficient* for an agent to count as identifying with her actual preferences. But many are tempted to think it is at least necessary.[6] Eleonore Stump, for example, offers what she takes to be a revision of Frankfurt's account of freedom in terms of second-order desires:[7]

> To express Frankfurt's concept of freedom using this revised understanding of second-order desires and volitions, we should say that an individual has freedom of the will just in case he has second-order desires, his first-order volitions are not discordant with his second-order desires, and he has the first-order volitions he has because of his second-order volitions.

It would be surprising, however, if it turned out that, on the relevant notion of identification, what many take to be a non-trivial, necessary condition on identification is required by coherence alone.

Relatedly, the possibility of mismatch between one's first- and second-order preferences has been used to illustrate the possibility of agents who do not value what they desire. David Lewis, for example, identifies *valuing* X with having a (second-order) desire to desire X:[8]

> [W]e'd better not say that valuing something is just the same as desiring it. That may do for some of us: those who manage, by strength of will or by good luck, to desire exactly as they desire to desire. But not all of us are so fortunate. The thoughtful addict may desire his euphoric daze, but not value it. Even apart from

6 Cf. e.g. Stump 1996, 1988, Taylor 1985, Christman 1987. See also Dworkin 1988, ch. 1 as well as the discussion of 'responsibility-identification' ('the notion of identification relevant to moral responsibility') in Fischer 2012. For additional references, see the discussion of the 'Dworkin/Frankfurt' model of autonomy in Christman 1988, p. 112ff. Even on some 'externalist' conceptions of autonomy (in the terminology of Buss 2014), appeals to second-order preferences or desires are not ruled out as necessary conditions on freedom, or autonomy, or responsibility—cf. e.g. Wolf 1987. For arguments that higher-order attitudes are not necessary for free or 'continent' action, see e.g. Mele 1992.

7 Stump 1988, p. 401. Cf. also Stump 1996, p. 203: "An agent's second-order volitions are authoritative for her because they reflect the all-things-considered judgment of her own mind, and her mind is constitutive of her".

8 Lewis 1989, p. 115. For some proposed modifications on Lewis' account of valuing, see Scheffler 2010. For related appeals to second-order attitudes in accounts of valuing, see Bratman 2000 and Copp 1993. For skepticism on the usefulness of second-order desires, see Harman 1993.

all the costs and risks, he may hate himself for desiring something he values not at all. It is a desire he wants very much to be rid of. He desires his high, but he does not desire to desire it, and in fact he desires not to desire it. [...] We conclude that he does not value what he desires, but rather he values what he desires to desire.

Similarly, Amartya Sen makes a distinction between one's preferences and one's *commitments*, where the latter are meant to correspond to one's values:[9]

Commitment is, of course, closely connected with one's morals. But moral this question is in a very broad sense, covering a variety of influences from religious to political, from the ill-understood to the well-argued.

The notion of commitment is then understood in terms of second-order preferences (or 'rankings of preference rankings'):[10]

[W]e need to consider rankings of preference rankings to express our moral judgments. [...] A particular morality can be viewed, not just in terms of the "most moral" ranking of the set of alternative actions, but as a moral ranking of the rankings of actions

The presumption, I gather, is that coherence requirements cannot by themselves guarantee that one values what one desires—in other words, that while it may be unfortunate for an agent's second- and first-order

9 Sen 1977, p. 329. See also Sen 1974 and Harsanyi 1955.

10 Sen 1977, p. 337. Sen is clearly using the term 'moral ranking' to talk about a ranking of preference systems—essentially a preference order over preferences (cf. also Sen 2004, p. 615ff). But there is an alternative reading of Sen, on which he does not intend to give an analysis of commitment in terms of an independently understood notion of preference as it applies to preference rankings themselves—see e.g., the discussion on p. 339 of the different ways in which the structure of 'meta-rankings' can be put to use. On that reading, the notion of commitment is defined as a preference-like structure that is nonetheless independent of our ordinary notion of preference, in the sense that the agent's meta-ranking says nothing about how action-propositions and preference-propositions compare to one another in terms of the ordinary notion of preference. If that is the best way of understanding Sen's notion of commitment, then Sen is essentially endorsing a view like the one I go on to sketch towards the end of 7, which is not vulnerable to the arguments in this paper. Thanks to an anonymous referee for pressing me on this point of interpretation.

preferences not to align, such a pattern of preferences is not a failure of minimal rationality.

And yet, it seems as if the presupposition of all these accounts has to be false. For there is an argument starting from fairly plausible assumptions to the conclusion that rationality requires that one's first- and second-order preferences align.

But first, an argument for a weaker thesis.

2 An argument

Let an *option* be an action-token that is up to me. (More generally, an *option for an agent* is an action token that is up to the agent.) Let 'ϕ', 'ψ', etc. range over options, and let 'A', 'B', etc. range over arbitrary propositions. Finally, let us say that ϕ-ing is *what I most prefer* iff for any $\psi \neq \phi$, I prefer ϕ-ing to ψ-ing. I will assume that there are no ties among my options—for any two distinct options, I (strictly) prefer one of them over the other. This will simplify the exposition, but nothing hinges on this assumption.

Consider now the following argument:

(1)
- P1. I know that, if I am rational, I will ϕ iff ϕ-ing is what I most prefer.
- P2. If I know that A iff C and I know that B iff D, then: I prefer A to B iff I prefer C to D.

  Therefore, If I know that I am rational and I prefer ϕ-ing to ψ-ing, then I prefer *most preferring ϕ-ing* to *most preferring ψ-ing*.

It bears repeating: the variable 'ϕ' only ranges over *options*: P1 thus amounts to the claim that, whenever ϕ-ing is under my voluntary control (if it is 'up to me'), I know that, if I am rational, I will ϕ iff ϕ-ing is what I most prefer.

For the sake of brevity, let us say that my preferences are *out of whack* iff ϕ-ing is what I most prefer but there is some distinct option ψ such that I prefer most preferring ψ-ing to most preferring ϕ-ing. We can now reformulate the conclusion of argument (1) in a more compact form:

FIRST LEMMA: If I know that I am rational, my preferences cannot be out of whack.

As it stands, the argument for our First Lemma isn't valid. But we can fix that by assuming a weak form of closure for knowledge, viz.[11]

WEAK CLOSURE: If I know that *A* entails *B*, then if I know that *A*, I know that *B*.

To see that, assume that I know that I am rational, and let ϕ and ψ be two distinct options. Given P1 and WEAK CLOSURE, it follows that I know that

I will ϕ iff ϕ-ing is what I most prefer.

Similarly, I know that

I will ψ iff ψ-ing is what I most prefer.

And these two entail, given P2, that I prefer ϕ-ing to ψ-ing only if I prefer *most preferring ϕ-ing* to *most preferring ψ-ing*—in other words, that my preferences cannot be out of whack.[12]

For the remainder of this note, I will assume WEAK CLOSURE. So the question to ask is: which of P1 and P2 can we reject?

11 It is worth noting that this form of closure does not entail the following, stronger principle:

> If *A* entails *B*, then if I know that *A*, I know that *B*.

There are well-known reasons for rejecting this principle. I will not rehearse them here. Suffice it to say that WEAK CLOSURE is much less objectionable.

12 It helps perhaps to reformulate the argument slightly. Let O_1, O_2, etc. be *option propositions*—propositions of the form *I will ϕ*. Say that I most prefer O_i iff for any option proposition O_n, I prefer O_i to O_n. The first premise in our argument is that, if I know that I am rational, then for any i: O_i is true iff I most prefer O_i.

Assume now I satisfy WEAK CLOSURE. Further assume that I know that I am rational. It then follows that I know that (a) O_1 is true iff I most prefer O_1 and (b) O_2 is true iff I most prefer O_2. But if I know that O_1 is true iff I most prefer O_1 and I know that O_2 is true iff I most prefer O_2, we can conclude from P2 that I prefer O_1 to O_2 only if I most prefer most preferring O_1 to most preferring O_2.

3 Preferences and expected utility

Start by noting that we can restate the argument in terms of certainty. By replacing 'I know that' with 'I am certain that' throughout, we get an argument for a stronger conclusion:[13]

(2) P1*. I am certain that, if I am rational, I will ϕ iff ϕ-ing is what I most prefer.
P2*. If I am certain that A iff C and I am certain that B iff D, then: I prefer A to B iff I prefer C to D.
Therefore, If I am certain that I am rational and I prefer ϕ-ing to ψ-ing, then I prefer *most preferring ϕ-ing* to *most preferring ψ-ing*.[14]

Here again, if we rely on the analogue of WEAK CLOSURE for certainty, we get a valid argument for the following lemma:

> SECOND LEMMA: If I am certain that I am rational, my preferences cannot be out of whack.

What to make of argument (2)? P1* is an empirical claim about me. At least on some days, we can suppose, it is true. So the only premise we could try to reject in a principled way is P2*. Unfortunately, we can derive P2* from plausible assumptions.

As I mentioned earlier on, I will be assuming a broadly Bayesian picture of rationality: rational agents are representable in terms of a credence function and a utility function, and their credence functions obey the axioms of the probability calculus. Preferences correspond to comparisons of expected utility in the following straightforward way:

$$A \geq B \text{ iff } EU(A) \geq EU(B),$$

13 I oversimplify. C* is only stronger than C if we assume that knowledge implies certainty. This may well be rejected. But the discussion of the argument below would remain unchanged.

14 I'm using italics in the formulation of this and other conclusions only to make clear the structure of the claim in the consequent of the relevant conditionals. The consequent simply states that my preferences rank the proposition that I most prefer ϕ-ing above the proposition that I most prefer ψ-ing.

where[15,16]

$$EU(X) = \sum P(w \mid X)u(w).$$

Let us assume further that certainty implies maximal credence: if I am certain in A then my credence in A is 1. Now, if I am certain that A iff C, it follows that $P(w \mid A) = P(w \mid C)$ for all w, and thus that:

$$EU(A) = \sum P(w \mid A)u(w) = \sum P(w \mid C)u(w) = EU(C).$$

Similarly, if I am certain that B iff D,

$$EU(B) = EU(D).$$

Hence, if I am certain that A iff C and certain that B iff D, then I prefer A to B iff I prefer C to D.

What about argument (1)? The only plausible barrier for assuming that P1 is true is the possibility that MAXIMIZE below is false (MAXIMIZE is just the embedded clause in P1):

> MAXIMIZE: If I am rational, then I will ϕ iff ϕ-ing is what I most prefer.

It is hard to see why MAXIMIZE could be true but unknowable. But again, from the perspective of a broadly Bayesian picture of rationality, MAXIMIZE seems highly plausible.

On a Bayesian picture, a rational agent maximizes expected utility.[17] And given the intimate connection between preferences and comparisons of expected utility, what maximizes expected utility is what the agent most prefers to do. So, as long as ϕ ranges over action types that are up to her, a rational agent will ϕ iff ϕ-ing is what she most prefers.[18]

15 For the moment, I will be using Jeffrey's definition of expected utility. If you think this is a mistake, say because you are convinced by the well-known counterexamples to evidential decision theory, worry not: for the most part, we can assume that I know all the causal facts, so that Newcomb-like scenarios do not arise. See Joyce 1999, Jeffrey 1983, Lewis 1981. I will return to this issue in §6.

16 For the sake of readability, I adopt the convention of writing '$P(w)$' instead of '$P(\{w\})$'.

17 Again, recall that, for the moment, I am skating over the issue of whether decision theory should be formulated in causal or evidential terms.

18 I will assume throughout that if ϕ-ing is up to me, then I will ϕ if I decide to ϕ. Options, as I understand them, are thus much like the 'basic actions' in Danto 1965. Admittedly,

Once we assume that if I know A, then I assign credence 1 to A, the argument for P2* generalizes straightforwardly to an argument for P2. So if the argument for P2* is sound, and if P1 is true, it follows that I cannot know that I am rational while having my preferences be out of whack.

4 Rationality and self-confidence

If a fully rational agent were required to be certain that she is fully rational, then we would have an argument that it is irrational for your preferences to be out of whack. Or at least, that it is less than fully rational to have your preferences be out of whack. After all, a fully rational agent would be certain that she is fully rational. And a sufficiently reflective agent should be able to see that MAXIMIZE is true. So, from argument (2) we could conclude that her preferences cannot be out of whack. By the same reasoning, if full rationality required knowledge of one's full rationality, we could conclude that full rationality requires my first- and second-order preferences to be in line with one another.

Now, it is controversial whether an ideally rational agent must be certain that she is ideally rational (*a fortiori*, it is equally controversial whether an ideally rational agent must know she is ideally rational). As David Christensen put it, "[t]he fact that [an ideally rational agent] happened to be ideally rational seems like the sort of claim for which some sort of warrant would be needed."[19] An ideally rational agent could well be in possession of misleading evidence to the effect that she is not ideally rational. Ideal rationality offers no protection against misleading evidence.

---

the assumption that there are actions that are up to me in this sense can be contested. But while it may be possible to do decision theory without it—see e.g. Jeffrey 1983, p. 11.9 as well as Pollock 2002—it is safe to say that much work on decision theory rather presupposes the existence of such actions (see e.g. Lewis 1981, p. 7, who defines the agent's options as proposition such that 'he can act at will so as to make any one of [them] hold'). It is an interesting question, one left open by what I say in this paper, whether the arguments below generalize to formulations of decision theory that do without assuming the existence of basic actions. Thanks here to an anonymous referee.

19 Christensen 2007, p. 327f.

Still, our results so far allow us to conclude that an ideally rational agent cannot know she is fully rational while having her preferences be out of whack. And the best explanation of this seems to be that an ideally rational agent cannot have her preferences be out of whack. For while an ideally rational agent may not know she is ideally rational, it is hard to see how merely having her preferences be out of whack could get in the way of her knowing (let alone of her being certain) that she is fully rational.

What is more, given P2 and its analogue P2*, we can conclude that a rational agent cannot know (or be certain) that she acts so as to maximize expected utility while having her preferences be out of whack. She need not have a view on whether she is fully rational. All that we need is that she takes (or knows) herself to be an expected utility maximizer.

(3) P3. I know that I will ϕ iff ϕ-ing is what I most prefer.
P2. If I know that *A* iff *C* and I know that *B* iff *D*, then: I prefer *A* to *B* iff I prefer *C* to *D*.
Therefore, if I know that I am an expected utility maximizer and I prefer ϕ-ing to ψ-ing, then I prefer *most preferring ϕ-ing* to *most preferring ψ-ing*.

And by the same token, we have:

(4) P3*. I am certain that I will ϕ iff ϕ-ing is what I most prefer.
P2*. If I am certain that *A* iff *C* and I am certain that *B* iff *D*, then: I prefer *A* to *B* iff I prefer *C* to *D*.
Therefore, if I am certain that I am an expected utility maximizer and I prefer ϕ-ing to ψ-ing, then I prefer *most preferring ϕ-ing* to *most preferring ψ-ing*.

We thus have apparently sound arguments for two additional lemmas:

> THIRD LEMMA: If I know that I am an expected utility maximizer, my preferences cannot be out of whack.
>
> FOURTH LEMMA: If I am certain that I am an expected utility maximizer, my preferences cannot be out of whack.

The best explanation of these two results seems to be that when it comes to expected utility maximizers, their preferences cannot be out of whack. For consider the alternative explanation, one on which it is perfectly rational for an agent's preferences to be out of whack but only if she does not take herself to be rational. On this view, having a particular pattern of coherent preferences can get in the way of knowing that one is minimally rational. This explanation requires accepting a thesis that is as surprising as the claim that rationality requires that one's preferences not be out of whack—it is hard to see how merely having one's preferences be out of whack could get in the way of knowing that one is rational. But it also requires positing a type of rationality requirement that is different in kind from familiar coherence requirements. After all, most rationality requirements on preferences (if not all) take the form of structural principles—e.g. the requirement that one's preferences be transitive.[20] In contrast, a requirement not to have one's preferences be out of whack *unless* one fails to be certain in one's own rationality is not a structural principle, for it makes explicit mention of a particular proposition. An explanation that requires introducing a new kind of coherence requirement lacks the unity of, and is less simple than, an explanation that posits an additional structural requirement of rationality.

In short, on the assumption that minimal rationality requires (i) that one be an expected utility maximizer and (ii) that one satisfy WEAK CLOSURE,[21] we seem to have two arguments for the following claim:

> INCOMPATIBILITY: Minimal rationality is incompatible with one's preferences being out of whack.

The arguments appeal to a fairly straightforward form of Inference to the Best Explanation (IBE), with P4 and P4* being straightforward generalizations of the Third and Fourth Lemmas:

20 A system of preferences is *transitive* just in case for any A, B, and C, $A \geq B$ and $B \geq C$ implies $A \geq C$.

21 Depending on what one takes the objects of credence to be, one might think that (ii) is redundant. But that would be a mistake, for a Bayesian conception of epistemic rationality, as I understand it, says nothing about the conditions an agent must satisfy in order to count as knowing that p. The analog principle for *certainty*, however, is a straightforward consequence of the tenets of Bayesian rationality.

(5) P4. A minimally rational agent cannot know that she is an expected utility maximizer while having her preferences be out of whack.
P5. The best explanation of P4 is that minimal rationality requires one's first- and second-order preferences to align.
Therefore, minimal rationality is incompatible with one's preferences being out of whack.

And similarly:

(6) P4*. A minimally rational agent cannot be certain that she is an expected utility maximizer while having her preferences be out of whack.
P5*. The best explanation of P4* is that minimal rationality requires one's preferences to align.
Therefore, minimal rationality is incompatible with one's preferences being out of whack.[22]

5 Open-mindedness

What to make of argument (6)? If argument (4) is sound, and if we think that IBE is a sound rule of inference, we must either accept INCOMPATIBILITY or reject P5*. Does the case for P5* stand up to scrutiny?

I assumed that, unless minimal rationality is incompatible with a mismatch of first- and second-order preferences, we could not explain why certainty in one's own minimal rationality was incompatible with a mismatch in one's first- and second-order preferences. But what if

22 If rationality required certainty that one is an expected utility maximizer, we would have a much more direct argument for INCOMPATIBILITY. I think considerations similar to those discussed above—against the claim that rationality requires certainty in one's own rationality (Christensen 2007)—tell against the idea that an ideally rational agent must be certain that she is an expected utility maximizer. However, see Ahmed 2014, p. 95f for an argument to the contrary—or rather, for an argument that, at least when facing a particular type of decision problem, an expected utility maximizer must take herself to be acting in accordance with expected utility theory. Unfortunately, Ahmed's argument relies on assumptions that I'm inclined to reject, but even if I were not, I could not rely on them in the present context without much further argument. Thanks to an anonymous referee for bringing Ahmed's argument to my attention.

rationality is incompatible with certainty in one's minimal rationality, whether or not one's preferences are out of whack?

Say that a credence function is *strictly coherent* or *regular* iff it is probabilistically coherent and assigns a value greater than 0 to any proposition that is possibly true.[23] Thus, while a probabilistically coherent credence function will assign 0 to any impossible proposition, a strictly coherent one will *only* assign a value of 0 to impossible propositions. Given that the negation of any contingent proposition is possible, it follows from the axioms of the probability calculus that a regular credence function will only assign value 1 to necessary propositions. In particular, if my credence function is regular, I will not assign value 1 to the proposition that I am an expected utility maximizer.

Now, it has been suggested that regularity is a rationality constraint on our credence functions.[24] The thought is that a rational agent should be 'open minded': I can be wrong about any contingent proposition, so I should never put myself in a position where I cannot change my mind about a given contingent proposition. And if having credence 0 in p by itself guarantees I will never assign anything but 0 to p, I should never assign credence 0 to p, unless p is necessarily false.[25] In particular, if regularity is a rationality constraint, a fully rational agent should not assign credence 0 to her not being an expected utility maximizer. Consequently,

23 Cf. Lewis 1980, p. 267 and Williamson 2007, p. 173. There are two additional ways of characterizing regularity. The first depends on taking the objects of credence to be *sentences* rather than propositions. On this way of thinking about it, regularity is a matter of assigning a value greater than 0 to any non-condradictory sentence. Another way involves taking the atoms of the algebra of propositions to be doxastic possibilities. Regularity would then amount to the claim that one assigns a value greater than zero to any proposition that is doxastically possible. Nothing in the discussion to follow hinges on how we characterize regularity.

24 E.g. Lewis 1980, Stalnaker 1970, Shimony 1955, Skyrms 1980, McGee 1994.

25 Of course having credence 0 in p does not, by itself, guarantee any such thing. If I acquire a piece of evidence in the form of a proposition I previously assigned credence 0 to, my updating on that evidence may well result in my changing my credence in p. It is true that if I assign non-zero credence to q, and assign 0 to p, after updating with q I will continue to assign 0 to p. And it is true that, on a ratio analysis of conditional probability, I can only conditionalize on events I assign non-zero credence to. But there are many reasons for rejecting the ratio analysis—see Hájek 2003 for discussion—and for allowing for update on propositions one assigns zero credence to. For a thorough, critical discussion of arguments in support of regularity as a rationality constraint, see Easwaran 2014.

if regularity is a rationality constraint, a fully rational agent should not be certain that she is an expected utility maximizer.

I do not intend to settle the question whether rationality requires that one's credence function be regular. Let us assume, for now, that it does. Would this provide an alternative explanation of P4*?

If we were to explain the truth of P4* by appealing to regularity as a rationality constraint, it would have to be that a violation of regularity alone guarantees a misalignment of one's first- and second-order preferences. But not any violation of regularity guarantees that one's preferences are out of whack.[26] So the truth of P4* cannot be explained by appeal to the fact that certainty in one's being an expected utility maximizer involves a violation of regularity.

Matters are less straightforward when it comes to argument (5). If regularity is a constraint on rationality, then knowledge had better not require certainty. But if knowledge does not require certainty, the argument for the claim that a minimally rational agent cannot know that she is an expected utility maximizer while having her preferences be out of whack—that is, the argument for P4—fails. And so too does argument (5).

Still, while this would block one route to INCOMPATIBILITY, we have a different argument for the same conclusion that is not vulnerable to the regularity objection.

6 Causal Decision Theory: a way out?

Can we respond to the argument for INCOMPATIBILITY by insisting that decision theory be formulated in causal terms?

26 *Proof*: Consider the partition given by the *value-level propositions* (the partition corresponds to the equivalence classes of the following equivalence relation: w and w' are in the same *equivalence class* iff the agent is indifferent between being in w and being in w'—in other words, iff $u(w) = u(w')$). Assume there is an equivalence class with more than one element, and let w^* be in that class. Shifting the probability from the proposition true in exactly those worlds in the equivalence class to the proposition true in exactly those worlds in the equivalence class different from w^* involves shifting to a state in which regularity is violated. But such a shift will not affect the agent's preferences, since it will not affect the computation of expected utility.

Recall that, according to Causal Decision Theory (CDT), rationality does not require that I act so as to maximize expected utility, where the expected utility of A is defined by:

$$EU(X) = \sum P(w \mid X)u(w).$$

Proponents of CDT believe that, in cases where you are uncertain as to what the causal structure of the decision problem you are facing is, it may be rational to act in ways that do not maximize expected utility. For example, consider this well-known case:

> NEWCOMB: There are two boxes in front of you, A and B. You must decide whether to take the contents of box B alone or to take the contents of both A and B. You know box A contanis \$10. You know box B contains \$100 iff someone you know to be highly reliable predicted you would choose to take the contents of box B only, and is empty otherwise.

If rationality requires maximizing expected utility, then rationality requires that you chose one box as long as your credence that the predictor is perfectly reliable is sufficiently high.[27] But, according to proponents of CDT, rationality requires that you chose two boxes. After all, what you do will have no causal impact on the contents of the boxes, so it would be silly to refuse to take box A—either box B will be empty, in which case you will be \$10 richer than if you only took box B, or it will contain \$100, in which case you will be \$100 richer than if you only take box B.[28]

27 The expected monetary gain of two-boxing is given by

$$P(\texttt{empty} \mid \texttt{two.boxes}) \times \$10 + P(\neg\texttt{empty} \mid \texttt{two.boxes}) \times \$110.$$

The expected monetary gain of one-boxing is given by

$$P(\texttt{empty} \mid \texttt{one.box}) \times \$0 + P(\neg\texttt{empty} \mid \texttt{one.box}) \times \$100.$$

The latter will be greater than the former whenever

$$P(\neg\texttt{empty} \mid \texttt{one.box}) = P(\texttt{empty} \mid \texttt{two.boxes}) > 0.55.$$

28 Lewis 1981, Skyrms 1980, Stalnaker 1981, Gibbard & Harper 1981, *inter alia*.

To see what CDT recommends, we need to introduce a bit of terminology. A *dependency hypothesis* is a maximally specific proposition 'about how the things he cares about do and do not depend causally on his present actions'.[29] The collection of dependency hypotheses are mutually incompatible (since maximally specific) and jointly exhaustive, so they form a partition $\mathcal{K}$. Given this partition, we can define the *causal* expected utility of X as follows:

$$CEU(X) = \sum_{K \in \mathcal{K}} P(K)EU(K \cdot X).$$

According to CDT, rationality requires maximizing causal expected utility. In NEWCOMB, you are uncertain as to how what you care about depends on your actions: you do not know whether, if you opt for two boxes rather than one, you will end up with $110 or with $10. This is borne out by the fact that the causal expected utility of choosing two boxes is higher than that of only choosing box B.[30]

The bearing of all this on our original problem may now be clear. For concreteness, let us focus on argument (1):

(1)
- P1. I know that, if I am rational, I will ϕ iff ϕ-ing is what I most prefer.
- P2. If I know that A iff C and I know that B iff D, then: I prefer A to B iff I prefer C to D.

  Therefore, If I know that I am rational and I prefer ϕ-ing to ψ-ing, then I prefer *most preferring ϕ-ing* to *most preferring ψ-ing*.

In arguing for P2, I explicitly relied on the following claim linking preferences and expected utility:

> LINK: I prefer A to B iff the expected utility of A is higher than that of B.

29 Lewis 1981, p. 11. You can think of dependency hypotheses as conjunctions containing exactly one counterfactual of the form $a \mathbin{\square\!\!\rightarrow} V$, for each value-level proposition V (in the sense of fn. 26), and each option a available to the agent. Cf. Joyce 1999, p. 170.

30 Here again NEWCOMB would provide a reasonable illustration. Even though the expected utility of one-boxing is higher than that of two-boxing, rationality requires that I choose both boxes.

A proponent of CDT may not object to using 'preference' so that LINK is true. But then she would argue that MAXIMIZE is not true:

> MAXIMIZE: If I am rational, then I ϕ iff ϕ-ing is what I most prefer.

For according to CDT, there are circumstances in which rationality requires that I ϕ even though the expected utility of ϕ-ing is lower than that of some of its alternatives. On the assumption that LINK is true, this would mean that there are cases in which rationality requires that I do something other than what I most prefer (among the relevant options).

Alternatively, a proponent of CDT may grant MAXIMIZE. But then she would reject LINK: a rational agent would, in NEWCOMB, most prefer choosing two boxes, so it cannot be that preferences go by way of comparison of expected utilities.

Let us stipulate that LINK is true. The proponent of CDT can thus object to argument (1) by rejecting P1, since she thinks that MAXIMIZE is false. Still, this leaves her with argument (2), which does not presuppose the truth of MAXIMIZE:

(2) P1*. I am certain that, if I am rational, I will ϕ iff ϕ-ing is what I most prefer.
P2*. If I am certain that *A* iff *C* and I am certain that *B* iff *D*, then: I prefer *A* to *B* iff I prefer *C* to *D*.
Therefore, If I am certain that I am rational and I prefer ϕ-ing to ψ-ing, then I prefer *most preferring ϕ-ing* to *most preferring ψ-ing*.

The proponent of CDT cannot reasonably insist that I cannot be certain that CDT is false.[31] So we still have a puzzle: if I am certain that rationality requires I ϕ iff the (evidential) expected utility of ϕ-ing is higher than that of its alternatives, then my preferences cannot be out of whack.

31 An anonymous referee rightly points out that a proponent of CDT can claim that rationality requires that one not be certain that CDT is false *if* she thinks regularity is a requirement of rationality. As I argued in 5, however, appealing to regularity cannot be what explains why I cannot rationally be certain that I am rational while having my preferences be out of whack.

A proponent of CDT is also not in a position to object to argument (4):

(4)
P3*. I am certain that I will ϕ iff ϕ-ing is what I most prefer.
P2*. If I am certain that *A* iff *C* and I am certain that *B* iff *D*, then: I prefer *A* to *B* iff I prefer *C* to *D*.
Therefore, if I am certain that I am an expected utility maximizer and I prefer ϕ-ing to ψ-ing, then I prefer *most preferring ϕ-ing* to *most preferring ψ-ing*.

The notion of rationality plays no role in this argument, so the question of whether or not CDT is true has no bearing on the soundness of argument (4).

To be sure, it is only if we accept something like MAXIMIZE that we will be in a position to argue that *rationality* requires that one's preferences not be out of whack. But regardless of what we take to be the best theory of rational choice, we are left with the puzzle of explaining the truth of our Second and Fourth lemmas.

Furthermore, suppose I know which dependency hypothesis is true. Then, even if I am a firm believer in CDT, I will be certain that if I am rational, I will ϕ iff ϕ-ing is what maximizes expected utility—after all, on the supposition that I know which dependency hypothesis is true, expected utility and causal expected utility coincide. So we can modify argument (5) to get a different argument for INCOMPATIBILITY:

(7)
P6. A minimally rational agent (in the causal sense) cannot know that she is a causal expected utility maximizer while having her preferences be out of whack unless she is uncertain as to which dependency hypothesis is true.
P7. The best explanation of P6 is that minimal rationality (in the causal sense) requires one's first- and second-order preferences to align.
Therefore, minimal rationality (in the causal sense) is incompatible with one's preferences being out of whack.

Perhaps the case for P7 is not as strong as the case for its analog, P5. Nonetheless, pending an alternative explanation of P6, we seem forced

to accept that minimal rationality is incompatible with one's preferences being out of whack.

7 Some alternatives worth exploring

There is something to the distinction between those preferences we identify with and the rest. There is something to the idea that a minimally rational agent may have preferences she does not identify with—that her values do not coincide with what she happens to prefer. It seems incredible that minimal rationality could guarantee that an agent will always identify with the preferences that she actually has.[32] A struggling alcoholic may well have as strong a preference for whisky over water as his preference for soup over salad. But we would be missing something if we could not account for the fact that the agent sees no problem with the latter and is deeply unhappy about the former.

If I am right, second-order preferences are not well suited to capture the phenomena. This may not be come as a surprise to those who think an adequate picture of rational agents must go beyond 'pristine belief/desire psychology'.[33] Perhaps the notion of *intention* will turn out to be indispensable for giving an account of valuing or identification.[34] But it would be nice if we could find alternative ways to model the phenomenon with the simple ingredients of a Bayesian picture of the mind.

One possibility worth exploring would be to abandon LINK. Perhaps the lesson from all this is that preferences best correspond to comparisons of *causal* expected utilities. Doing so might allow us to make room for a minimally rational agent whose preferences are out of whack. To see why, go back to NEWCOMB. Suppose you are certain that the predictor is perfectly reliable, so that you are certain that you will opt for two boxes iff box *B* is empty. Given LINK, it follows that you prefer two-boxes over one-boxing iff you prefer that box *B* be empty. But note that the causal expected utility of choosing two-boxes need not equal the causal expected utility of box *B* being empty.

32 Cf. Frankfurt 1971, p. 11: "[A] rational creature, who reflects upon the suitability to his desires of one course of action over another, may nonetheless be a wanton."

33 Mele 1992, p. 281.

34 Cf. Bratman 1987, 2000.

Unfortunately, while there is much to be said for using causal expected utilities to evaluate courses of *action*, it is far from clear what bearing causal expected utilities have on the evaluation of preferences—at least if we assume that which preferences I have is not under my control.[35]

Another option would be to try to capture the phenomena not in terms of the expected utility (whether causal or not) that an agent assigns to her having a given pattern of first-order preferences, but in terms instead of features of her *conditional* preferences. We could, for example, say that I identify with a given pattern of first-order preferences iff, conditional on certain contingent facts about myself being different from what I know them to be, I still have those first-order preferences.[36]

Here's one way of spelling this out. Let us say that my preference for A over B is *robust* iff, conditional on my preferring B over A, I prefer A over B.[37] Or, to put it in slightly more vivid terms, say that my preference for A over B is robust in this sense if I would advise a counterpart of mine who prefers B over A to nonetheless opt for A over B.[38]

For example, my preference for vanilla over chocolate is not very robust. Conditional on my preferring chocolate over vanilla, I no longer prefer vanilla over chocolate. Thus, while

$$\text{I eat vanilla} \geq \text{I eat chocolate},$$

nonetheless

$$\text{I eat chocolate} \geq^{C} \text{I eat vanilla},$$

35 Cf. Joyce 1999, p. 253: "evidential decision theorists have been right all along about the nature of rational desire, but they have mistakenly thought that all desires provide reasons for action. The fact that A would be better news than B does not give an agent a reason to choose A over B unless what is meant is that A's news value *on the subjunctive supposition that it is performed* is greater than B's news value *on the subjunctive supposition that it is performed*. The moral, then, is that Jeffrey's theory is not really a logic of *decision* but a logic of *rational desire*." See also Joyce 2000.

36 Cf. Jeffrey's solution to the so-called paradox of ideal evidence—Jeffrey 1983, p. 196f.

37 In the end, we may think it best not to think of the proposition(s) we use for determining how robust a set of preferences is as being *about* first-order preferences. Perhaps we should think of them as propositions describing certain phenomenal features of one's experience (say, the particular way that vanilla tastes, and so on). We may want to ascribe robust first-order preferences to agents that do not have the conceptual capacities to think about their own preferences. Doing so in full generality, however, may require some hard work.

38 Cf. Parfit 1984, §59 on desires that are conditional on their own persistence.

where $\geq^C$ denotes my preference ranking conditional on my preferring chocolate over vanilla. This preference ranking is here defined in terms of my conditional expected utilities:[39]

$$A \geq^C B \text{ iff } EU(A \mid C) \geq EU(B \mid C),$$

where

$$EU(A \mid B) = \sum_w P(w \mid A\&B)u(w).$$

The unwilling addict, the one who doesn't endorse having the first-order preferences that she has, is one who, conditional on her preferring not smoking over smoking, would not prefer smoking over not smoking. The committed vegetarian that is disgusted by the taste of meat, in contrast, is one who prefers not eating meat over eating it, *even conditional* on her coming to prefer the flavor of the non-vegetarian options on the menu.

A rational agent can be certain that she is an expected utility maximizer and yet have first-order preferences that are not very robust. We know, from the argument for P2*, that for an agent who is certain that she is an expected utility maximizer the expected utility assigned to the proposition that she ϕs will equal the expected utility she assigns to the proposition that she most prefers to ϕ. So, the expected utility she assigns to smoking (say) will equal the expected utility she assigns to most preferring smoking. But conditional on her preferring not-smoking over smoking, say, the expected utility she assigns to smoking may be much lower than the one she assigns to not smoking.

A toy example might help. Suppose we have an agent whose credence is defined over four different worlds. In w_1, she smokes and most prefers smoking; in w_2, she smokes and most prefers not smoking; in w_3, she doesn't smoke and most prefers not smoking; and in w_4 she doesn't smoke and most prefers smoking. Her utility function assigns 10 utiles each to w_1 and w_4 and 5 utiles each to w_2 and w_3. Let S be the proposition that she smokes and M the proposition that she most prefers smoking.

Assuming she is certain that she is an expected utility maximizer, our agent's credence is concentrated on w_1 and w_3, so that the expected

39 It may be best to use preferences conditional on counterfactual suppositions, to accommodate agents that are certain of their actual preferences, but this is not a question I will address here.

utility of smoking is given by $P(w_1 \mid S) \times 10$ and the expected utility of not-smoking is given by $P(w_3 \mid \neg S) \times 5$, so that the expected utility of smoking is twice as that of not smoking. Conditional on her most preferring not smoking, however, her preferences are reversed, for[40]

$$P(w_4 \mid S\&\neg M) \times 10 > P(w_2 \mid \neg S\&\neg M) \times 5.$$

To be sure, this does not yet tell us how to distinguish a third kind of smoker—she who prefers smoking over not-smoking, but who is neither committed to her preferences nor willing to condemn them. In principle, we could identify this third kind with someone who is indifferent, conditional on her preferring not-smoking over smoking, between smoking and not-smoking. This may not be the best strategy all things considered. But its existence does shows that appealing to robustness gives us enough structure to allow for minimally rational agents who do not identify with the preferences that they have.

At any rate, even if it improves on the appeal to second-order preferences as a way of characterizing the phenomenon of endorsing preferences, the appeal to robustness leaves something to be desired. The unwilling addict, we are tempted to say, experiences a certain *conflict*. It is tempting to say, of such an addict, that she wants to smoke, but that she also wants to *not* smoke.

To allow for conflicts among preferences, however, we need more structure than the simple Bayesian story provides us with. For any agent whose preferences correspond to comparisons of expected utilities will have no conflict among preferences: if the preference ordering mirrors the comparisons of expected utilities for some pair of a credence function and a utility function, the preference ordering will be antisymmetric: a preference for smoking over not smoking will rule out a preference for not smoking over smoking.

This leads us to a final, more radical suggestion. This strategy requires acknowledging that the motivational structure of agents like the unwilling

40 I am allowing for conditioning on events of zero probability to be well-defined. In this particular case, we have that S & $\neg M$ entails that w_4 is actual, so that

$$P(w_4 \mid S\&\neg M) = 1.$$

addict cannot be captured by a single system of (conditional and unconditional) preferences. We could start by positing *two* systems of preferences, one which favors smoking over not-smoking, and another one which favors not-smoking over smoking—this would allow for some kind of *fragmentation* at the level of preferences.[41] By itself, this would not be enough. Talk of a conflict among two systems of preferences suggests a certain kind of symmetry that seems not to be there in the case of the unwilling addict. And for talk of two systems of preferences to have any purchase, we need to say more about what distinguishes those two systems, and about what reason we have for positing two such systems in the first place.

Allan Gibbard, in his discussion of weakness of will in *Wise Choices*,[42] makes a distinction between what he calls two *systems of control*—two distinct motivational systems present in human agents.[43] To a first and very rough approximation, we can think of one such system as being one we share with other animals, and accordingly call it an *animal motivational system*. The other system, one peculiar to reflective creatures, we can call the *normative motivational system*:[44]

> This is a picture of two motivational systems in conflict. One system is of a kind we think peculiar to human beings; it works through a person's accepting norms. We might call this kind of motivation *normative* motivation, and the putative psychological faculty involved the *normative control system*. The other putative system we might call the *animal control system*, since it, we think, is part

41 For discussion of fragmentation as a way of modeling agents with conflicting beliefs, see: Lewis 1982, Stalnaker 1991, Egan 2008. (Cf. also Greco 2014 for a proposal to model cases of 'epistemic akrasia' in terms of fragmentation at the level of belief that is structurally very similar to the present suggestion.) Strictly speaking, the kind of fragmentation we would need in order to model the unwilling addict is not just a fragmentation among preferences: an agent with fragmented belief states will in all likelihood have different corresponding preference orderings, but that conflict among preferences would not reflect any conflict at the level of the agent's motivational state. Better to model our agent with two distinct utility functions, each one of which plays certain roles that the other one does not.

42 Cf. Gibbard 1990, p. 56–76.

43 Of course, the suggestion that we think of agents as endowed with two distinct motivational systems goes back at least to Plato (e.g. *Phaedrus*, 237e-238). Cf. also Watson 1975.

44 Gibbard 1990, p. 56.

of the motivational system that we share with the beasts. Let us treat this picture as a vague psychological hypothesis about what is going on in typical cases of "weakness of will".

We can think of each such system as determining a particular pattern of behavioral dispositions. Their difference is that they are manifested in different contexts, and are responsive to different considerations. The normative control system, as Gibbard describes it, is the one that governs avowal in, and (crucially) that is sensitive to, normative discussion:

> [W]e should think of the motivation I have been calling 'normative' as motivation of a particular, linguistically infused kind—a kind of motivation that evolved because of the advantages of coordination and planning through language.[45]

The animal control system is one that is not (as) sensitive to normative discussion—one that is manifested by cravings and appetites.[46]

On Gibbard's norm-expressivistic theory of normative judgment, when I say that I ought not smoke, I am giving voice to a feature of my normative control system—very roughly, that in contexts governed by my normative control system, I act as if I prefer smoking over not-smoking. But we needn't take on all of Gibbard's metanormative commitments in order to make use of the idea of a motivational system that is only triggered in certain contexts. The unwilling addict, we could say, is one whose animal control system includes a disposition to opt for smoking over non-smoking, even though his normative control system includes a disposition to opt for not-smoking over smoking.

Needless to say, more needs to be said in order to flesh out the thought that the unwilling addict exhibits a conflict between two systems of preference, and to tell a complete story about what distinguishes these two

45 Gibbard 1990, p. 57.

46 Gibbard later goes on to add some subtlety to this distinction—in particular, to think that the conflict present in cases of weakness of will is one between the norms an agent accepts and those she is 'in the grip' of, where being in the grip of a norm is a matter of behaving in ways that are sanctioned by the norm even if one is not disposed to explicitly avow to, nor to appeal to in practical reasoning, the relevant norm. I will stick to the less subtle distinction for the sake of simplicity, though I expect a fully worked out theory of preference endorsement will need to take those subtleties into account.

systems. But I'm inclined to think that something much like this story has got to be the way to go—endorsing preferences just cannot be a matter of preference among preferences.[47]

References

Ahmed, Arif. 2014. *Evidence, Decision and Causality*. Cambridge: Cambridge University Pres.

Bratman, Michael E. 1987. *Intentions, Plans, and Practical Reason*. Cambridge, Mass.: Harvard University Press.

Bratman, Michael E. 2000. Valuing and the Will. *Philosophical Perspectives* 14. 249–265.

Buss, Sarah. 2014. Personal Autonomy. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2014.

Christensen, David. 2007. Epistemic Self-Respect. *Proceedings of the Aristotelian Society* 107. 319–337.

Christman, John. 1987. Autonomy: A Defense of the Split-Level Self. *Southern Journal of Philosophy* 25(3). 281–293.

Christman, John. 1988. Constructing the Inner Citadel: Recent Work on the Concept of Autonomy. *Ethics* 99(1). 109–124.

Copp, David. 1993. Reasons and Needs. In Ray Frey & Chris Morris (eds.), *Value, Welfare, and Morality*, 112–137. Cambridge: Cambridge University Pres.

Danto, Arthur C. 1965. Basic Actions. *American Philosophical Quarterly* 2(2). 141–148.

Dworkin, Gerald. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Pres.

Easwaran, Kenny. 2014. Regularity and Hyperreal Credences. *Philosophical Review* 123(1). 1–41.

Egan, Andy. 2008. Seeing and believing: perception, belief formation and the divided mind. *Philosophical Studies* 140(1). 47–63.

Fischer, John Martin. 2012. Responsibility and Autonomy: The Problem of Mission Creep. *Philosophical Issues* 22(1). 165–184.

Frankfurt, Harry G. 1971. Freedom of the Will and the Concept of a Person. *Journal of Philosophy* 68(1). 5–20.

47 Thanks to David J. Barnett, David Braddon-Mitchell, Phil Bricker, Rachael Briggs, Adam Elga, Kenny Easwaran, Alex Gregory, Dustin Locke, Chris Meacham, Eliot Michaelson, David Plunkett, James Shaw, Sam Shpall, and Katia Vavova for helpful comments on earlier versions of this paper. Thanks also to Earl Conee and two anonymous referees for this journal. Special thanks to Tom Dougherty multiple comments and advice.

Frey, Ray & Chris Morris (eds.). 1993. *Value, Welfare, and Morality*. Cambridge: Cambridge University Pres.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, Mass.: Harvard University Press.

Gibbard, Allan & William L. Harper. 1981. Counterfactuals and Two Kinds of Expected Utility. In William L. Harper, Robert C. Stalnaker & G. Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance and Time*, 153–190. Dodrecht: Reidel.

Greco, Daniel. 2014. A puzzle about epistemic akrasia. *Philosophical Studies* 167(2). 201–219.

Hájek, Alan. 2003. What Conditional Probability could not be. *Synthese* 137(3). 273–323.

Harman, Gilbert. 1993. Desired Desires. In Ray Frey & Chris Morris (eds.), *Value, Welfare, and Morality*, 138–157. Cambridge: Cambridge University Pres. Reprinted in Harman 2000, pp. 117–136.

Harman, Gilbert. 2000. *Explaining Value and Other Essays in Moral Philosophy*. Oxford: Oxford University Press.

Harper, William L., Robert C. Stalnaker & G. Pearce (eds.). 1981. *Ifs: Conditionals, Belief, Decision, Chance and Time*. Dodrecht: Reidel.

Harsanyi, John C. 1955. Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy* 63(4). 309–321.

Jeffrey, Richard C. 1974. Preference Among Preferences. *Journal of Philosophy* 71(13). 377–391.

Jeffrey, Richard C. 1983. *The Logic of Decision*. University Of Chicago Press.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. New York: Cambridge Univ Press.

Joyce, James M. 2000. Why We Still Need the Logic of Decision. *Philosophy of Science* 67. S1–S13.

Lewis, David. 1980. A Subjectivist's Guide to Objective Chance. In Richard C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, vol. II, 263–293. Berkeley, CA: University of California Press. Reprinted, with Postscripts, in Lewis 1987, pp. 83–132.

Lewis, David. 1981. Causal Decision Theory. *Australasian Journal of Philosophy* 59(1). 5–30.

Lewis, David. 1982. Logic for Equivocators. *Noûs* 16(3). 431–441.

Lewis, David. 1987. *Philosophical Papers*. Vol. II. New York: Oxford University Press.

Lewis, David. 1989. Dispositional Theories of Value. *Proceedings of the Aristotelian Society* 63. 113–137.

McGee, Vann. 1994. Learning the Impossible. In Ellery Eells & Brian Skyrms (eds.), *Probability and Conditionals: Belief Revision and Rational Decision*, 179–199. Cambridge: Cambridge University Pres.

Mele, Alfred R. 1992. Akrasia, Self-Control, and Second-Order Desires. *Noûs* 26(3). 281–302.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Pollock, John L. 2002. Rational Choice and Action Omnipotence. *The Philosophical Review* 111(1). 1–23.

Scheffler, Samuel. 2010. Valuing. In *Equality and Tradition: Questions of Value in Moral and Political Theory*, 15–40. Oxford: Oxford University Press.

Sen, Amartya K. 1974. Choice, Orderings and Morality. In Stephan Körner (ed.), *Practical Reason*, 54–67. Oxford: Basil Blackwell.

Sen, Amartya K. 1977. Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy and Public Affairs* 6(4). 317–344.

Sen, Amartya K. 2004. *Rationality and Freedom*. Cambridge, Mass.: Harvard University Press.

Shimony, Abner. 1955. Coherence and the Axioms of Confirmation. *The Journal of Symbolic Logic* 20(1). 1–28.

Skyrms, Brian. 1980. *Causal Necessity*. New Haven: Yale University Press.

Stalnaker, Robert C. 1970. Probability and Conditionals. *Philosophy of Science* 37(1). 64–80.

Stalnaker, Robert C. 1981. Letter to David Lewis. In William L. Harper, Robert C. Stalnaker & G. Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance and Time*, 151–152. Dodrecht: Reidel.

Stalnaker, Robert C. 1991. The problem of logical omniscience, I. *Synthese* 89(3). 425–440. Reprinted in Stalnaker 1999, pp. 241-254.

Stalnaker, Robert C. 1999. *Context and Content*. Oxford: Oxford University Press.

Stump, Eleonore. 1988. Sanctification, Hardening of the Heart, and Frankfurt's Concept of Free Will. *Journal of Philosophy* 85(8). 395–420.

Stump, Eleonore. 1996. Identification and Freedom. *Philosophical Topics* 24(2). 183–214.

Taylor, Charles. 1985. What is Human Agency? In *Human Agency and Language: Philosophical Papers 1*, 15–44. Cambridge: Cambridge University Pres.

Watson, Gary. 1975. Free Agency. *Journal of Philosophy* 72(8). 205–220.

Williamson, Timothy. 2007. How Probable Is an Infinite Sequence of Heads? *Analysis* 67(3). 173–180.

Wolf, Susan. 1987. Sanity and the Metaphysics of Responsibility. In Ferdinand David Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, 46–62. Cambridge: Cambridge University Pres.