Published as: Duncan MacIntosh, "Preference-Revision and the Paradoxes of Instrumental Rationality", Canadian Journal of Philosophy, Vol. 22, No. 4 (December 1992a), pp. 503-530.

This manuscript version differs in small ways, e.g., in footnoting. Please refer to published version.

Preference-Revision and the Paradoxes of Instrumental Rationality[1]

Duncan MacIntosh
Dalhousie University

1. Introduction

On the received theory of practical rationality, rational choices must maximize one's expected utility.[2] But sometimes--in Paradoxical Choice Situations (PCSs)--one's basis for choices (e.g., preferences, beliefs, dispositions, intentions) will affect the choices of other agents. They will choose favourably to one satisfying one's preferences only if one's choice basis is not maximization on one's current preferences. Paradoxically, it seems maximizing not to maximize. David Gauthier thinks it rational here to adopt a disposition constraining oneself from maximizing, and that non-maximizing choices are then rational; an action is rational if dictated by a disposition it maximized to adopt.[3]

I give another solution: do not refrain from maximizing, but adopt whichever preferences it would maximize to have as your choice basis given your current preferences; maximize on your revised preferences in subsequent choices.[4] We thus get the choices Gauthier thought would

---

1. For helpful discussion I thank Neera Badhwar, David Braybrooke, Bob Bright, Douglas Butler, Peter Danielson, Bob Martin, Tory McGeer, Howard Sobel, Kadri Vihvelin, Sheldon Wein and especially Julie Colterjohn, Richmond Campbell, and Terry Tomkow. Thanks also to my commentator, David Zimmerman, at the 1988 CPA, to which part of an earlier version was given. Finally, I am very grateful to referees Gregory Kavka and Michael Webster, to two anonymous referees and to the editors for advice on form and for deep questions. A Killam Post-Doctoral Fellowship from Dalhousie University funded early research.

2. For details, see Richmond Campbell and Lanning Sowden, eds., Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem (Vancouver: The University of British Columbia Press 1985).

3. David Gauthier, Morals By Agreement (Oxford: Clarendon Press 1986), Chs. I, V, VI, and `Deterrence, Maximization, and Rationality,' Ethics, 94 (1984) 474-495.

4. I suggested this in `Two Gauthiers?,' Dialogue, 28 (1988) 43-61 and later papers. It appears briefly in different forms in Gregory Kavka, `Some Paradoxes of Deterrence,' The Journal of Philosophy, 75 (1978) 285-302, reprinted in John Perry and Michael Bratman, eds., Introduction to Philosophy: Classical and Contemporary Readings (New York: Oxford University Press 1986)

follow a choice of dispositions, but need not call non-maximizing choices rational. We save the received theory by applying it not just to choices of actions--means to ends--but also of ends (given current ends).

This, too, however, offends orthodoxy. For we normally think it rational to prefer something, x, only on three pretexts: x has an objectively preferable property, or a property one prefers things to have, or x's obtaining would advance a preference one has. To these I add that preferring x would advance one's preferences; one prefers x not because of properties of x, but of the preference for x. This is a study of the rationality of revising one's preferences on that pretext.

2. A Paradox For Instrumental Rationality

To begin, we need a PCS. One would be a Prisoners Dilemma (PD) where the agents will know each other's choice basis and may amend their own before the game.[5] But the Deterrence Dilemma (DD) is more vivid[6]: unless you, a nuclear super-power, intend to retaliate if attacked by another (unless you "Intend") and unless this would guide you in an attack, you will be attacked. You prefer that harms to everyone be minimal. Since Intending would deter harms, it maximizes. But if you are attacked, it is not maximizing to retaliate (to "Act"); that would only cause more harm. (Everyone will die, so there is no future deterrent advantage to Acting.) Actions are rational if maximizing, and retaliating is not, so it cannot be rationally intended. But intending maximizes and so seems rational, even though the act intended is non-maximizing. What to do?

Gauthier thinks it rational to Intend for that maximizes, rational to Act for that expresses a

---

516-526 (references are to the latter), David Lewis, `Devil's Bargains and the Real World,' in Douglas Maclean, ed., The Security Gamble: Deterrence Dilemmas in the Nuclear Age (Totowa, N.J.: Rowan and Allenheld 1984) 141-154, and Amartya Sen, `Choice, Orderings and Morality,' in Stephan Korner, ed., Practical Reasoning (Oxford: Basil Blackwell 1974) 54-67. It is mentioned skeptically in Jordan Howard Sobel, `Maximizing, Optimizing, and Prospering,' Dialogue, 27 (1988) 233-262, and in passing in some of his earlier papers. Edward McClennen, in his `Prisoner's Dilemma and Resolute Choice,' in Campbell and Sowden, 94-104, and his `Constrained Maximization and Resolute Choice,' Social Philosophy & Policy, 5 (1988) 95-118, may give views on preferences for the Prisoner's Dilemma like those I push here for PCSs in general; if so, I am developing his idea. But he may not see agents undergoing revisions in their preferences in PCSs; they would just have different ones then, or just resolve to act differently. I hope to examine elsewhere what the former might involve. I argue that resolutions cannot rationalize non-maximizing choices in my `McClennen's Early Co-operative Solution to the Prisoner's Dilemma,' The Southern Journal of Philosophy, 29 (1991) 341-358.

5. E.g., see Gauthier, Morals By Agreement, Chs. I, IV, V, and my `Preference's Progress: Rational Self-Alteration and the Rationality of Morality,' Dialogue, 30 (1991) 3-32.

6. For the DD, see Gauthier, `Deterrence,' Lewis, `Devil's Bargains,' and Kavka, `Some Paradoxes.'

rational intention, and actions are rational if rationally intended. The rationality of Intending determines that of Acting.[7]

Some commentators, however, think one cannot rationally intend and perform a non-maximizing action.[8] E.g., Mark Vorobej agrees that if it is rational to Intend, it is rational to Act. But he thinks it irrational to Act for that is not what the intention it would maximize to have post-attack would demand; it would say to refrain from retaliating. But if it is irrational to Act, it is irrational to Intend. The rationality of Acting conditions that of Intending. Thus rational agents can't Intend. Nor, then, can they rationally Act. Deterrence by rational threat of rational retaliation is impossible for agents who disprefer harms.[9]

But Gauthier would think it no argument against Acting's rationality that it is not maximizing, nor against Intending's that the act intended is not, nor even that Acting expresses an intention not now maximizing to have. An intention is rational if adopting it maximized; an action, if it expresses such an intention. It maximized to Intend pre-attack, so retaliation would express an intention it then maximized to adopt. Thus Intending and Acting are rational.[10] Still, to call them such for his reasons we must accept his account of rationality. But I think we can rationalize them on the standard account.

3. Rational Revisions of Preference-Functions

I now rehabilitate a possibility rejected (for reasons we will soon address in general terms) by David Lewis[11], Gregory Kavka[12], and Gauthier[13]: a rational agent would come to prefer to retaliate. For it maximizes to deter. He can only deter if he would rationally retaliate if attacked so his enemy fears him. He would only find retaliation rational if he preferred it. Thus, pre-attack, it is maximizing and so rational for him to so revise his preferences as to find it

---

7. Gauthier, `Deterrence.'

8. Kavka, `Some Paradoxes,' 519-521, and Mark Vorobej, `Gauthier on Deterrence,' Dialogue, 25 (1986) 471-476.

9. Vorobej, `Gauthier on Deterrence.'

10. I think Vorobej's criticism can be made conclusive. See my `Retaliation Rationalized: Gauthier's Solution to the Deterrence Dilemma,' Pacific Philosophical Quarterly, 72 (1991) 9-32, my `Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem,' The Southern Journal of Philosophy, 26 (1988) 399-425, and my `Preference's Progress,' where I try to show that rational choices must maximize.

11. Lewis, `Devil's Bargains,' 153-154.

12. Kavka, `Some Paradoxes,' 520-525.

13. Gauthier, `Deterrence.'

maximizing and so rational to retaliate, to become one who prefers that harms be minimal except after an attack, where he prefers retaliating to minimizing harms. The new preference then makes rational the previously irrational intention to retaliate. Retaliation from the new preference, and so from the intention it rationalizes, will maximize on the new preference and so be standardly rational.[14] If he prefers for its own sake that harms be minimal, he should come now instead to prefer to retaliate if attacked later, for its own sake; he has conclusive instrumental grounds for seeing retaliation as a better intrinsic good. This deters, for retaliation then maximizes; it is the choice he would rationally make post-attack, the one his enemy fears. Intending and Acting are rational. But classical rationality is preserved: rational choices maximize on the preferences had when choosing. One maximizes on one's initial preferences in choosing to supplant them with retaliatory ones, then maximizes on them in choosing to retaliate.

The guiding thought: to choose rationally is to maximize on one's preferences. But sometimes it maximizes to commit to deciding one's future choices in some way other than that of maximizing on the preferences one now has. Still, rational choices must maximize. So to choose ways of making future choices that will make your future choices rational, you must choose ways choices from which would maximize. This requires choosing a new preference-ordering for deciding choices. Any other determinant would yield irrational choices because it would have you choose against your preference-ordering, and so, irrationally. For if you must choose a new way of making future choices, that is because it maximizes now to be inclined to choose later against your original ordering. But since rational choice must express the ordering you have when choosing, to choose rationally later, you must have changed the ordering.

Is this stoicism?[15] That is either the view that one shouldn't care about what happens (for one can't affect it), or shouldn't care about whether what one cares to happen happens (same reason), or should care only that what will inevitably happen, happen (for one can't make it otherwise). But I say one should care about things the caring about which will advance one's prior cares. Stoicism is defeatist: yield to the ineluctable. I counsel strategy: change the world by changing your values. Stoicism assumes agents want to care about what they can get (or its rationales are non-sequiturs). I assume that to care about something is to be prepared, ceteris paribus, to do what is needed to get it.

So: we can get Gauthier's conclusions on which actions are rational from the received theory since it justifies preference-functions on which retaliation maximizes.[16] Agents face a second-

---

14. I give a similar rationale for revising one's preferences before facing some PDs in my `Co-operative Solutions to the Prisoner's Dilemma,' Philosophical Studies, 64 (1991) 21-33, and my `Preference's Progress.' Also, see McClennen's papers.

15. The question is from Kavka's referee's report.

16. In `Two Gauthiers?,' I argue that a similar reading of Gauthier's Constrained Maximizer disposition best serves his conclusions on the rationality of morality as exemplified in rational choices in some PDs. See also McClennen, `Constrained Maximization.' Gauthier insists in `Morality, Rational Choice, and Semantic Representation: A Reply to My Critics,' Social Philosophy & Policy, 5 (1988) 173-221, that his proposal is different and independently viable.

order choice not among dispositions constraining preference expression, but among preferences. These do the same job, but what it is to have a disposition able to rationalize retaliation is now straightforward: it is to have preferences which dispose one to retaliate because they make it maximizing.

4. Objections to the Rational Preference to Retaliate

Doesn't the old preference rationalize dropping the retaliatory one post-attack, when action on it can only increase harms? No, for you now prefer post-attack retaliation to post-attack harm-minimizing; you no longer prefer harm-minimizing simpliciter.

But you do not initially have a preference to retaliate, only that harms be minimal. This makes you wish to adopt the former only as a means to minimizing harms. So since it is only an instrumental preference, and since it is pointless after the aim it was to serve failed, won't dropping it be rational? No. For while you adopt it for instrumental reasons, it is not a preference for retaliating as a means, but for its own sake. So to act rationally from it post-attack is to retaliate, not to drop it and minimize harms. You acquire it for instrumental reasons, but it is intrinsic.

One has an intrinsic or basic preference for x just if one prefers x as an end--"intrinsic" because one prefers its target for itself, not its power to cause something else, and "basic" because such preferences are the basis for deciding which other preferences to have, e.g., for things which help satisfy one's basic preferences. (These are not necessarily irrevocable or hard-wired.) If one prefers x only for its power to cause something else, y, preferred as an end, one instrumentally prefers x. The difference between the two preference types is in what one finds preferable in the target, x. In the former, one finds x preferable, in the latter, x's power to help cause what one prefers in itself (so that one prefers x only if and while it so helps).

This distinguishes preference types. But we must also distinguish kinds of reasons for them. The tradition sees only these reasons for preferring a condition, x: first, x has objective value. But this is outside the theory of rationality given preferences (where the utility one gets if x obtains measures its value, not something in x). The second: x has a property one already prefers in things for itself; the third: x would cause ends already preferred. To these, I add: preferring x would cause such ends; in a DD, intrinsically preferring retaliation would serve one's basic aversion to harms.

But doesn't one enslave oneself to a preference which, post-attack, would be irrational? For it now works against the anti-harm preference it was to serve. So isn't one now just the helpless automaton of a blind and isolate rogue preference? Doesn't action on it fail to maximize on the balance of one's "real" preferences? How is retaliation now any more rational and voluntary than if one had put the matter beyond one's control by delegating the decision to retaliate to a machine, or to someone else?

Well first, one is never a slave. One willingly acquires new preferences for their deterrent effect, and then willingly retaliates because, given them, one then wants to if attacked. Nor is one "blind"; one's eyes are open in adopting the new preferences, and in retaliating from them. Second, one does not adopt an unintegrated preference, one yielding an incoherent preference function in which one prefers retaliating to not harming and not harming to retaliating; one does not acquire a preference discordant with one's continuing preferences, then irrationally act on it

against them. Rather, one changes the ranking (acquires a preference function in which, post-attack, retaliation ranks higher than harm prevention) as recommended by the present ranking, then retaliates rationally as that then maximizes, reflects what one all-in prefers post-attack. One's preferences are integrated, coherent. Finally, whether an action maximizes depends on the preferences had when acting, not on ones no longer had. One's new ones are (now) one's "real" ones.

But suppose you think the enemy can be deterred by you preferring to retaliate. So you do. He attacks. But then you learn his attack was locked in: your preferring could not have deterred. There was never any point in it, so surely retaliating is also pointless. But you now prefer to and to be rational it seems you must. The lesson: do not prefer to retaliate simpliciter, but to minimize harms except when attacked by those who would have been deterred by your readiness to retaliate, in which case, to retaliate. (In this game, the enemy cannot help being either locked-in or threat-sensitive.) You then prefer minimum harms in normal situations, but in attacks, prefer not to retaliate against those who would not have been deterred, but to retaliate against those who would. So you need only retaliate where a readiness to do so would have deterred. (Deterring is just lowering the chance of attack. Thus an attack does not prove that you did not deter; the chance of your readiness affecting the enemy's choice is fixed independently of his decision in this case.) Adopting this ranking, then, better maximizes on your preferences. For it is sensitive to future information and commits you only to the potential for harming needed to minimize the odds of all harms.[17]

So: it is rational to revise one's preferences so that one intrinsically prefers minimal harms only if that does not involve letting off deterrable attackers, otherwise preferring to retaliate against the threat-sensitive; thereafter it is rational, because maximizing, to retaliate. This answers Vorobej: it is rational to act on those preferences. Thus as far as that goes, it is not irrational to acquire them.

But is this not bad faith, willful irrationality? For one first hates harms, and later prefers to harm if attacked. How can this happen without one either acting against one's preferences or self-deceivedly thinking it no harm to retaliate? Well, the new ranking reflects one's original one when one adopts it, being the maximizing and so rational one to acquire given the original. Post-attack, it rationalizes actions which do not reflect the original, but no matter, for one no longer has it and rational choice proceeds from current preferences. (Likewise, one has no reason to regret having become a retaliator, for one retrospects the decision from retaliatory values.) One

---

17. Is Gauthier's account better here? For he may think it rational to retaliate only if the disposition to do so was maximizing, i.e., deterring. If one learns it never was, one needn't retaliate. But I think Gauthier holds that a disposition maximized not if it in fact made a desired outcome likely, but was believed to on the evidence when adopted. So unless he carefully formulates the disposition, his agent too must retaliate against those later discovered to be undeterrable. For the disposition to retaliate against them would still be one it "subjectively" maximized to adopt earlier. Thus he must have the disposition be to retaliate only against those still believed to have been deterrable (as I must have the rational retaliatory preference), or he must say that things "maximize" relative to the facts, not just beliefs. (Thanks to an anonymous referee for the issue, and to Terry Tomkow for help replying.)

constantly (and rationally) believes that retaliation harms, but one's attitude towards harming in a DD changes.

But it would be irrational to alter one's beliefs here, so why not one's preferences? Because rational belief follows evidence. If you believed just for practical reasons (e.g., to get money) your belief would be irrational by epistemic standards. But there are no standards which an instrumentally rationalized change in basic preferences violates. Or are there...?

5.On the Very Idea of a Rational Revision of Basic Preferences; Their Nature and Rational Kinematics[18]

Normally one rationally prefers something when one sees it as either needed for, or of a kind with, what one already basically prefers. In the former, one does not come to like it for itself, only its use as a means; in the latter, one does, but not because one has changed one's basic preferences, only because one sees it as like things already preferred. So can one really rationally revise basic preferences just because they are served by their own revision? The seeming oddness of this may owe to our precritical theories of the nature of basic preferences, their origins, and the prospects for their rational revision. We now consider these and their implications for our proposal.

Theory (a): one gets basic preferences non-rationally: by heredity, association and reward conditioning. We are born with desires for food, air, etc. We come to intrinsically value other things by associating them with satisfying our "natural" preferences. So we come to like the dinner bell for itself by associating it with food. One is neither rational nor irrational in one's first preferences--one is just born with them--nor in acquiring new ones--one just does by habitual association of their objects with ones already basically preferred. (Think of Hume or B.F. Skinner.)

Discussion: but even if the process of acquiring a basic preference is non-rational--if one cannot rationally directly adopt one--it may yet be rational to arrange conditioning that will result in one.[19]

Theory (b): one only acquires "new" basic preferences when one sees that some conditions have properties one already basically prefers as a kind, and so one comes to intrinsically prefer ones on which one is newly enlightened. This is just what it is to come to like something for its own sake.[20] But one never rationally acquires basic preferences for new kinds.

Discussion: if (b) says one cannot acquire a basic preference for a new kind, it is likely

---

18. Thanks to Richmond Campbell, Douglas Butler and Howard Sobel for discussion on the issues in this section.

19. See Kavka, `Some Paradoxes,' `The Toxin Puzzle,' Analysis, 43 (1983) 33-36, and `Responses to the Paradox of Deterrence,' in Maclean, 155-159. He sees some justification for doing it, but may not see it as conclusive, especially for rational moral agents. See below and my `Kavka Revisited: Some Paradoxes of Deterrence Dissolved,' (unpublished manuscript).

20. Thanks to Sobel for this suggestion.

false--people can be conditioned into almost anything; if that one cannot rationally arrange to acquire one, this seems groundless given rationales like the above for acquiring one in a DD. Now, "coming to like something for its own sake" is ambiguous between coming to like its intrinsic properties--for whatever reason--and coming to like it because of those properties; it conflates the resulting attitude with a reason for adopting it. It can be a reason to prefer something that one already prefers things of its general kind, to prefer it because of its intrinsic properties. But sometimes one should prefer something because there is advantage to preferring it. There are more reasons for preferring things than are dreamed of in (b).

But if it is a reason to prefer something that one already prefers things of its kind, surely it is a reason not to prefer something that one disprefers things of its kind. So surely in dispreferring harms we must disprefer actions increasing them, and so disprefer retaliating?[21] One has some reason to disprefer it, for it would cause intrinsically dispreferred harms. Preferring to retaliate will cause retaliation on attack. That is instrumentally dispreferred. So, were this the only thing relevant, preferring to retaliate should be instrumentally dispreferred. But is this conclusive?

You must prefer having whichever preferences would minimize harms (the deterring basic preference to retaliate), but you disprefer doing what will pointlessly increase harms (retaliating). Which should you prefer, deterring attack, or not retaliating? Well, which maximizes your expected utility (EU)? To decide the EU of retaliating, multiply the possible utilities and disutilities of doing so by their odds, then add. Say 50 harms will certainly be caused by retaliating, refraining, sure to cause none; an EU for refraining of 50. To decide the EU of preferring intrinsically to retaliate, multiply the possible utilities and disutilities of doing so by their odds, then add. The best that could happen is 100 harms saved (no attack, no retaliation), the chance, high, say .99, the worst, 100 harms in the outcome (attack, retaliation), the chance, low, .01. An EU of 97.9. 97.9 beats 50, so you should come to intrinsically prefer retaliating. (Assume one cannot retaliate unless one prefers to, nor prefer to retaliate and not do it if attacked. Since preferring to retaliate and refraining from retaliating are mutually exclusive, choosing one is omitting the other. So their EUs are comparable, the higher of preference revision decisive over the lower of refraining from retaliating.) You should prefer having the deterring preference, for you can most probably prevent more harms with it than by not deterring and refraining from retaliating. Thus, your preference for minimum harms yields a stronger instrumental preference for (is a better reason for having) a basic preference for retaliating.

That you would normally instrumentally prefer refraining is not decisive in whether you should intrinsically prefer retaliating. This depends on whether preferring it, or not so preferring in order to be able to refrain, would maximize on the basic preference grounding the normal instrumental preference. For satisfying an instrumental preference yields no utility itself; only if the condition satisfying it causes one satisfying an intrinsic preference. So you have no independent rational duty to your instrumental preferences,only your basic ones--here best served by their own revision. This justifies abandoning one's strongest or highest-ranked preference as a means to satisfying it. (Actually, we should not speak of dropping or adopting a preference; rather one reranks preferences, changes one's set of preference-functions, of rankings of conditions in a preference-ordering. If I use the former locutions, I mean the latter.)

---

21. This worry is in Kavka, `Some Paradoxes.'

Precritically,it may seem an argument against (conditionally) preferring to inflict retaliation's harms that one now prefers minimum harms. But this very preference makes acquiring a retaliatory one rational in PCSs. Oddly, that one prefers to minimize harms should make it easier, not harder, to become one who sometimes prefers to harm.

Note: you do not come to prefer retaliation because retaliation has a property you originally intrinsically preferred (for it doesn't), but because preferring it does--that minimizes harms. This is the novelty here, is how the rational acquisition of basic preferences differs from that of instrumental ones, and from coming to prefer something for its having a trait already intrinsically preferred.

Theory (c): we get a preference for x by seeing x's objective value, as if preferableness were an objective property, like mass (as in moral realism). Retaliation is not objectively preferable. So it would be bad faith, willful irrationality, to let instrumental factors make us prefer it, just as it would be to let them affect our beliefs. Preferring is more like believing or perceiving than like wanting; thus the rational kinematics of preference is epistemic, not pragmatic. So acquiring a preference for retaliation would be like epistemic irrationality since one knows it to be objectively dispreferable.

Discussion: value realism is notoriously problematic. But even were it true, the preference for retaliation might objectively have the moral property of it being obligatory to acquire it (for the harming deterred), retaliation, that of it being obligatory not to perform it (since it would cause gratuitous harms), and our paradox returns. Also, the solution: presumably, one must act on the strongest obligation, which, surely is to do what would likely result in the least objectively bad conditions; one must deter and so prefer to retaliate. So if value is objective, and if its accurate perception can give one a DD, it also affords an escape. For to have a DD, one must find that preferring to retaliate (and so to harm after an attack) and refraining from retaliating (and so from harming) have objective value, but that the former's is higher in the circumstances. Otherwise, Intending would not maximize (objective value) and one would not have the dilemma.

But one sees that harming has negative value, and seeing that preferring conditionally to harm has positive value does not change this. So isn't to adopt the preference to overlook the negative value of the condition preferred, like a willful, epistemically irrational misreading of value facts about harms?

Perhaps the tension can be eliminated this way: in the paradoxical situation of the DD, harming by retaliating cannot have lower value than refraining. For since preferring to retaliate is objectively preferable, and since one can only rightly prefer the truly preferable, retaliating must be too. To pursue the analogy with belief, if one has evidence that not p, but stronger evidence that p, one must take p as probably true. One does not overlook the counter-evidence; but the total evidence favours p. Now, if one has evidence that retaliating is objectively dispreferable (evidence consisting, perhaps, in one's current aversion to it, for here, preferring is perceiving, one's preferences therefore tending to track objective value), but stronger evidence that it is preferable (in one's current stronger preference to deter, and so, to prefer to retaliate), one is justified in preferring to retaliate; for one is justified in finding it probably objectively preferable all things considered.[22] And we could defend it by saying that it is not just harming, but

22. My thanks to Douglas Butler for this approach.

implementing a morally obligatory value; harming for an exonerating reason (for whoever had to so value). This is like Gauthier's move on rationality: retaliation is prima face irrational given an aversion to harms, but is made rational by its expressing a disposition rational to adopt to serve the aversion. But while Gauthier thought one's preferences could remain constant and a non-maximizing action be rational, I think one must revise them so that one's actions will maximize (but on the new preferences). Likewise, here, maybe the morally correct preferences vary with situations, the objective moral values of acts, with their historical contexts.

If this still seems implausible, perhaps we must question the conception of value on which EU measures it. For it may seem refuted by this argument: if it is right, an action's value is its EU. Since the EU of deterring (which entails retaliating if attacked) is (pre-attack) higher than that of not deterring (which entails refraining), retaliating must have higher value than refraining. Absurd; objectively, retaliating must be at least as bad as deterring is good. So that theory of value must be false.

But this is unfair to the subjective theory of value; for we managed to pose (and, perhaps, resolve) the DD even for the objective theory. So how things rank in value, and whether a present attitude's value can determine that of a future action, are independent of value's nature. Preferring to retaliate and retaliating might have higher objective value than preferring to refrain and refraining. In any case, if we accept a conception of value fitting instrumental rationality (where nothing has value independently of preferences), we must accept these consequences for value order in DDs. And if we cannot analyze instrumental rationality without intruding objective value, this troubles all rational choice theory, not just our proposal. Moreover, taking it as a problem on grounds of the above reductio may make DDs (situations where one is not just conflicted, but has a way out) impossible. If it correctly portrays the objective theory's consequences, DDs can only arise if the subjective theory is true (where retaliating can be justified by the value of preferring it). (More in Section 9.)

I learned another argument for (c) from Kavka. He says that on my view,

> the only justification for having certain preferences...may have nothing to do with the qualities of the states of affairs these are preferences for, but only with the instrumental value of having these preferences for satisfying earlier ones. But these...in turn, might be of only instrumental value, and so on. Unless basic preferences are somehow anchored in the features of the targeted states of affairs, why does and should it matter to us whether...they are satisfied?[23]

His worry: it would/should only matter to one that one's preferences be satisfied if one feels/is justified in them. I say a preference is only justified by its advancing another one. But then the same for it, etc. So there is no final justification. So it does not/should not matter to us that our preferences be satisfied. But it does/should, so my theory is false. If value is objectively in things, however, there will be final justifications: I am objectively justified in preferring x if x is objectively good or preferable, subjectively justified if I think it is. My reason for preferring x is that x is good, or so I think, and in thinking this I have reason for it to matter to me that x obtain.

---

23. I quote from his referee's report.

But then it is not the instrumental value of a preference that justifies having it or caring that it be satisfied, but facts (or beliefs) about its target.

But Kavka makes some dubious assumptions: first, it can't matter to one that the preference for x be satisfied without reason to prefer x. This may hold for instrumental preferences, but not basic ones. For to have a basic one for x is for it to matter to one that x obtain. If one prefers x, for x to obtain is for that preference to be satisfied; thus in having a basic preference for x, it matters to one that it be satisfied. And one might just have been born with it; that one did not get it by rational reflection or that there is no justification for it, does not mean one can't or shouldn't care that it be satisfied. If one basically prefers x, that is reasonable if there is no alternative preference whose possession would maximize on the first, i.e., if preferring x is not less conducive to x's obtaining than preferring something else, y. One does not need a reason to prefer x for it to matter to one that x obtain, except when rationally adjusting one's preferences. And when it would advance them to have new ones, that is a reason for the new ones.

A basic preference is rationally obligatory if adopting it advances a prior one. The issue only arises if one has preferences, so acquiring one's first ones is non-rational; they cannot be justified by prior ones. But a preference need not be justified to be reasonably held; we only explain what it is for one to be justified if it is. The first can, however, be rationally assessed by whether having different ones would best advance them (and by whether they are coherently ordered).

One can have a basic preference whose satisfaction matters to one even with no reason; why not also instrumental ones? One only cares about their targets for their power to cause what one values as an end. So one cannot have an instrumental preference without a basic one, and it can't matter that a condition instrumentally preferred obtain unless that would cause something intrinsically preferred. With no ends, one logically cannot have a preference for a means as such. Perhaps one could have an "instrumental preference" for x, another for y because y conduces to x, etc., but if one finds nothing intrinsically valuable, the objects of these preferences are not means to an end; only to a means. And if one's every preference is for a condition's power to cause another condition, nothing preferred for itself, one has infinitely many preferences. Otherwise one's first one, the one rationalizing the rest but itself groundless, is basic; one does not have only (so-called) instrumental preferences. I don't know whether there could be agents with only such preferences, i.e., with infinitely many; nor whether we could ever rationalize their actions: we might be unable to give finite explanations, and "infinite" ones in terms of "infinite reasons" might not really count as such. But this is academic, for I define a preference as instrumental just if its target is only preferred for its power to advance a basic preference.

Kavka's second assumption: one only has reason to prefer x given a final answer to the question, why prefer x? But true or not (see above on regresses), one has an answer on my account: preferring x advances one's basic preferences. It is not needed for them to justify other ones that one have a reason for them, only no reason against them. Once one has them, one cares that their targets obtain; and this can justify instrumental preferences and such alternative basic ones as would advance the originals.

The third assumption: if the only reason for preferring x is that this advances another preference, the same holds for it, etc.; with only instrumental reasons for preferring things, there is nothing one prefers for itself to be a final reason. But we saw that one doesn't need a reason for all of one's preferences for one to ground another, nor for there to be things one cares about for themselves. Thus there can be final reasons, so satisfaction of our preferences can matter to us.

And it suffices for a preference's adoption to be rational that it advances one's prior basic preferences. Thus one can have only instrumental rationales for one's basic preferences (if one has reasons for any of them) without threat to their being basic. This is only implausible if it implies that one prefers nothing for itself. But it really implies that one has preferences which are ones for things for themselves. (Do not confuse instrumental reasons for a preference with instrumental preferences. One can have instrumental reasons for basic preferences.) So one can have rational basic preferences without value realism. There need be no objectively preferable conditions, only ones whose intrinsic properties one prefers.

Besides, value realism cannot explain why things matter to us. x can be objectively preferable and it not matter to me whether x obtains. I can even believe x to be good and it not matter to me--I may not respect the good. For it to matter to me that objectively preferable conditions obtain, I must subjectively prefer them. Another queerness in objective goodness: its recognition is not inherently motivating. This does not trouble the subjective theory: to subjectively prefer x is, ceteris paribus, to be moved to make x obtain, to care about its obtaining. And philosophers only want an objective theory to make subjective values criticizable. But our theory advances that hope in affording some basis for judging basic values.

Kavka might agree that it can only matter that one's instrumental preferences be satisfied if they serve a basic one; also, that to have a basic one is to care about its target's obtaining. But he may yet wonder, if the only reason for having any preference is that it advances another, do we have any reason for a basic preference? If yes, it is either that it advances another one, or something about the condition preferred. If the former, we have a regress; if the later, a preference is not made rational by advancing a prior one.

But there is only a problem if every reasonable preference must have a reason; and it needn't. A held preference is rationally permissible if having another one would not better advance it. Most are reasonable, for most are self-advancing; having them maximizes the odds that their targets will obtain because rational agents try to create what they prefer. But a reasonable preference need not be uniquely rational. Nature can arbitrarily give one initial preferences: I may be born with a preference for eating chocolate, you for anything but. Both are presumptively reasonable since they motivate and so maximally advance their own satisfaction. They are only irrational if having other preferences would better advance them.

Finally, one's basic preferences can be anchored in features of targeted states of affairs, as when, having a basic preference for conditions with certain properties, one prefers a condition on seeing it has them. My innovation is a theory of the rationality of basic preferences for new kinds of condition. Even there, when one acquires a new preference because that advances a prior one, it is because of the properties of the first one's targets that one acquires the later: one adopts it to cause the targets of the first. (Here, though, it is not the condition targeted in a preference that rationalizes it, but that a target of an earlier preference is made likely by acquiring one with a different target.)

But suppose I basically prefer whatever has property F, and see that a has F; I have reason to basically prefer a. But suppose the preference for Fxs, or a stronger basic preference, would be advanced by me dispreferring Fxs; I have reason not to prefer a. Isn't there a conflict of rationales? No. For whether one's basic preferences are sound is prior to whether a condition fitting them should be preferred. But this is only revealed in PCSs. Usually the subordinate consideration operates because normally one needn't drop a preference to advance it. But that

consideration only applies to preferences rational to keep; it is over-ridden if revising them maximizes.[24] An agent's preferences might leave him indifferent between revising and not; or he might have incoherent values advising both revising and retaining the same preference. But in the first case he could rationally choose by a coin toss; and in the second, his ill-ordered values make him irrational on any account and so no counter-example to ours.

Theory (d): we are born with fixed actual or potential basic preferences, but which ones are "to the fore" depends on the availability of goods, our training, our beliefs about our circumstances, etc.

Discussion: for (d) to fully explain what determines a preference's manifestation on a given occasion, we would need versions of (a)-(c), the relevant facts being either associative (a), deductive from active preferences (b), involving perceptions of objective value properties (c), or pragmatic (our theory). Moreover, whether one has all the preferences one can ever have at birth, or just the potential for them, recalls Locke on innate ideas. And we might complain in a Lockean spirit that there is no difference between a preference not activated except under certain conditions, and one not had except then. So the verdict on (d) is in those on (a)-(c), giving no further reason against pragmatic preference revision. Doesn't Decision Theory require (d) for agents to have a firm basis for rational choices? If one's preferences are not givens, but are variable, how can we speak of the rationality of a choice given preferences? But this only requires determinate preferences at each time of choice, not the same ones at all times. And we preserve this: in choosing to revise one's preferences, one chooses from one's current ones; in later choosing actions, one chooses from one's new ones. So our theory does not deprive rational choice of a basis.[25]

6. Preference Revision: Action or Not (and Does it Matter)?

A change is direct if seeing its aptness suffices to induce it, voluntary if it occurs just because one wants it to, costly if it lowers one's utility (though it may raise it overall by its consequences' utility). Is rational basic preference revision voluntary, like action; or is it like the non-voluntary, direct acquiring of new beliefs; or must one arrange to be conditioned into it by non-rational (and maybe costly) processes?

In the normal rational kinematics of preference, preferences change non-voluntarily directly one sees the conditions justifying it. Likewise, an epistemically rational agent non-voluntarily acquires a belief directly on seeing conclusive evidence for it. Normal belief-change on new evidence is costless[26], as is normal preference-change. The reason normal belief and preference

---

24. Thanks to Terry Tomkow for discussion here.

25. Thanks to Peter Schotch and an anonymous referee for help on this. And see my `Retaliation Rationalized,' pp. 28-29.

26. There may be a utility cost if one disprefers the beliefs one acquires. But our agent only prefers that harms be minimal. So if he comes to believe that harms have occurred, it is they which trouble him, not his belief in them.

changes have all these properties is that they are not actions. They are direct responses to events, not meddlings in them, and are caused by evidence, not volition. Intention acquisition is similar. A rational agent immediately, directly comes to intend to do what she now sees conclusive reason to do now, but so far as intending to do something implies wanting to do it, she will do it voluntarily--because she wants to do it. Where the action is to be done later, she will now intend to do it then directly on seeing conclusive reasons to do it then, and will (typically) voluntarily do it when she thinks then is now.[27]

We claim that pragmatically justified basic preference-changes also occur directly. But how can one "directly" change one's preferences on pragmatic pretexts? It just spontaneously occurs on seeing one's predicament. If you see that x has properties you intrinsically prefer, you come to prefer x. Is it sensible to ask how? If you get good evidence for p, you believe p. Is there a problem about how you get yourself to believe it? If you see that it would serve your desires to have some desire, we say you will acquire it. Why should the question how arise here?

Perhaps for this reason: the pragmatic rationale for preferring x speaks not of the properties of x, but of the preference for x. So surely it justifies not an attitude to x, but to the preference for x.[28] A rational agent would respond to a good argument that p is true by believing p, to a good argument that x is preferable by preferring x. Commensurately, wouldn't she respond to a good argument that preferring x is preferable not by preferring x, but by preferring to prefer x? So how does she get from the second-order to the first-order attitude? But one can argue that p is true by arguing that the belief that p is a true belief; so can't one argue that x is preferable by arguing that the preference for x is preferable? This may seem a bad analogy, however. For p is true iff the belief that p is true. So to argue the one is to argue the other. But surely the preference for x can be preferable even if x is not; so arguing the former is not arguing the latter, and so is not arguing for a first-order attitude to x. We claim, however, that in some situations, PCSs, it is a reason to prefer x that preferring x is preferable; in PCSs, to argue that preferring x is preferable is to argue that x is preferable. Indeed, as we saw above in the discussion of objective value, PCSs prove that the value of x and the value of an attitude to x cannot always vary independently. (Compare Gauthier: PCSs prove that an action's rationality cannot always be independent of the rationality of the disposition to do it.) Thus in a PCS, it is rational to respond to the argument that preferring x is preferable by directly preferring x.

But perhaps we have only shown that it would be practically rational to acquire retaliatory preferences, not that retaliation is preferable. If so, to acquire such preferences, one must arrange to get them by a non-rational process, like reward-conditioning. And this may be costly. But we can handle this: just factor the costs of self-alteration into those involved in the risk of having to retaliate; it is rational if the EU of revising one's preferences by the means needed exceeds that of not.

---

27. See Terrance Tomkow, Against Representation (forthcoming, Cambridge University Press) on the relation between intentions and rational actions.

28. Thanks to Sobel for the issue.

7. Psychologies Perfect and Real; Contingent Constraints on Rational Self-Alterations; the Problem of Coherence

Many think the DD exists only in theory. Does anything like it occur for real agents? Sure. E.g., in our early character development, it is demanded that we not just act a certain way but be a certain way, acquire certain values. This is induced by situations where our values will be advanced by their own revision. The teen wants acceptance and learns that he will only get it if he becomes "cool", autonomous from the approval of others, "his own man."[29] But growing up is hard, raising the question, do we have fully rational psychologies, ones able to respond directly to the demands for change from the calculus of instrumental rationality? An empirical matter, and perhaps questionable, though I doubt adolescent anxieties call it into question. They may not reveal psychological barriers to direct shifts, but just represent the discomforts of the circumstances motivating them.

But now consider Kavka's Toxin Puzzle (TP)[30], which may reveal biological obstacles to preference change. You are offered a million dollars ($1M) to intend tonight to drink a temporarily nauseating poison tomorrow. You needn't drink tomorrow to collect, only intend by tonight to drink tomorrow. This is a PCS since it maximizes to intend to drink, but not to drink. (You prefer the $1M to avoiding nausea, but dislike nausea and would normally avoid its causes.) I say you should prefer to drink. You then would get the $1M tonight, but would and should drink anyway tomorrow. But can you get a preference to drink simply by seeing its use in getting the $1M?

Assume that whatever one does to oneself, the poison will nauseate as a fact of biochemistry. Doesn't this mean one must disprefer drinking when one is to drink? So how will one's preferences be well-ordered if one acquires a preference to drink? Won't one prefer and disprefer it--have incoherent and so irrational preferences? Or won't it be impossible directly to prefer it by tonight since it cannot be made non-nauseating?

But nausea is not something one logically must disprefer. Indeed, one hears many do not--bulemics, anorexics, the Ancients who took emetics after feasts to renew their appetites. They, we hear, are nonchalant about it, and while they may first only endure it for its instrumental value (in keeping thin, renewing appetite), they may later like it for itself (or not mind it) from association with what they like for itself (thinness, appetite). So he who prefers what will nauseate him need not have ill-ordered preferences.

---

29. For other examples, see my `Persons and the Satisfaction of Preferences: Problems in the Rational Kinematics of Values' (unpublished manuscript).

30. Kavka, `The Toxin Puzzle.'

Some philosophers, though, think some experiences inherently unpleasant--pain, extreme hunger and thirst, nausea--others, pleasant--the tastes of some foods, the sensations of sex--and that these ground all disvalue and value. Coming to prefer to retaliate seems like coming to like pain, and since pain is inherently unpleasant, surely one can't rationally, if at all, come to prefer it for itself, nor retaliation either. (This is an aesthetic version of the objection from objective value, considered above.)

But many attitudes seem to separate an experience's being pleasant and being preferable: the masochist prefers to be hurt, the martyr, to suffer for others, the religious self-flagellant, to feel a physical wretchedness befitting his inherent sinfulness. It seems necessary to their getting what they prefer that they be in unpleasant states. Maybe they only instrumentally prefer them for their conducing to other things; perhaps the masochist just wants (pathological) proof of love, the martyr, to save others from suffering, the self-flagellant, to atone. But this need not be so. The masochist may just like--intrinsically prefer--painful sensations; the martyr may find more nobility in saving others by suffering than by a less unpleasant expedient; the self-flagellant may think his sinfulness part of his identity before God, self-infliction of pain the only appropriate response. If so, our account can recognize, explain, and, sometimes, rationalize the phenomena: even if some experiences are irreformably pleasant/unpleasant, that is separate from whether they ought to be preferred, from whether they are preferable or unpreferable. So even if our findings of things unpleasant is unalterable, our preferences for things need not be.

Still, it is contingent whether one's psychology is capable of preference-change on seeing its instrumental value (just as is whether one is capable of belief-change upon new evidence). There may be physiological barriers; or maybe attitudes get ingrained. Perhaps for some of us it is physically/psychologically impossible for our preferences to change directly on that pretext, or even for any conditioning to effect the change. But since it makes sense and need not yield ill-ordered preferences, people so limited are not ideally rational. They do have an excuse: they cannot be otherwise; and to choose rationally given this, they must choose in light of it. Facing the TP, it would be irrational for them, knowing their limits, to contract to buy a house with the $1M they cannot get. But that says nothing against us on what they must do if ideally rational. One's current aversion to something need not be a conclusive reason to avoid it; only if the aversion is a preference rational to have. And preferences can rationalize their own abandonment, whether one's psychology allows this or not. Thus, that one now has a negative attitude to an action needn't be a conclusive objection to preferring to do it.

There are other possible worries for our proposal deriving from the fact that one's values are interconnected.[31] If one disprefers harms in general, one disprefers this and that harm, things causing harm, people who harm, etc. If one is averse to nausea, one is averse to its causes, perhaps to mention of it, etc. So to change one preference, mustn't one change many? And wouldn't the magnitude of the task obstruct change? Not necessarily. Suppose that, even considering all my preferences, it would maximize to change a given one and all of those associated with it. The advantage then justifies changing all of them, and their relatedness does not entail that this cannot occur directly. An analogy: I thought it would not rain today, but I was wrong. Many of my beliefs must change: I thought it would be dry but now I must believe it will

31. Kavka brought this and the next few issues to my attention.

be wet; I thought I wouldn't need my umbrella, but now, that I do, etc. But does that make it harder for me to believe it's raining? Surely not. All of my rain-relevant cognitive attitudes just change. The same, surely, for conative attitudes in connected value changes. Maybe I hated some people because I thought they stole something from me. Turns out it was all a mistake. Now I like them and all who sail with them; all my negative attitudes to them just become positive. Attitude connectedness may multiply the practical difficulties of change if one's physiology or psychological ingrainedness makes change hard in general. It may even make it practically impossible, but that doesn't make it irrational; it may require an expensive arranged process, though if the change still maximizes even given the expense, it is still rational, if it will not be direct.

But what if, considering all my preferences, changing a given one and their associates would not maximize? This holds if my strongest preference is to keep to a moral principle forbidding character corruption (on which, more below), or if my self-concept requires me to keep my preferences, i.e., if I prefer most not to revise myself, to be somehow true to my current self.[32] Now one may so prefer, but need not, I think, to be rational. And one is not then in a PCS. It is not rational to change--not because that is irrational in general, but because, given the preferences one happens to have, one does not most care that harms be minimal (as in the DD), or to get $1M (as in the TP), and so is not in a situation where, to cause the targets of one's strongest preferences (to maximize), one must change one's choice basis.

I say it is rational in the TP to prefer drinking the sickening toxin even with the $1M in hand. But what is the character of this preference? Does one now view illness as a desirable way of celebrating new wealth? Or is it only drinking this particular glass of toxin on this occasion that one prefers? More like the latter. For while in PCSs it maximizes to change one's preferences in some way, it generally maximizes to retain them, because having them tends to make their rational holders cause their targets. Thus in revising one's preferences so as to maximize on them, one should try so far as possible to preserve this tendency. One should acquire preferences maximization on which demands, as often as possible, behaviours just like those appropriate to the old preferences, and different only in the case where the PCS obliges one to intend an action not maximizing on the old. The revision must be the minimum mutilation required. Thus, in the DD, one should not go from hating harms to loving them, but to hating harms except when they comprise retaliations against deterrable attackers. When not dealing with such agents, one then behaves as before, minimizes harms. In changing one's preferences, one maximizes on the old (reduces the odds of there being any harms), and in maximizing on the new, one's actions still in effect maximize on the old, except in the special case of an attack (when one will cause the harms of a retaliation). And in the TP, one should not go from loving money and the avoidance of nausea to loving nausea, but to loving money and avoiding all nausea except that nausea a commitment to inducing which was a condition of getting $1M. Again, in changing one's preferences, one maximizes on the old--gets the $1M--but in normal situations, continues to avoid nausea, as if prosecuting one's old preferences; only in the one odd case, the drinking of this glass of poison, does one act differently.

---

32. Howard Sobel may suggest agents have such a self-concept in his `Maximizing, Optimizing, and Prospering.'

But preferences are by nature related to the qualities of things, and rationally must vary in a coherent way with variations in those qualities. And nausea is nausea; whatever properties of it I hate in this nausea are also in that nausea. So how can I coherently have one attitude to the first and a different one to the second? How do my new attitudes make sense? Answer: their objects are different. The first is an aversion to all nausea the intention to induce which is not rewarded by $1M, the second, an affinity for such nausea the intention to induce which is so rewarded. My attitude in each case is to a condition of which nausea is but a part; the other parts of the conditions differ (one contains an intention to drink and a reward--the condition is a "fat" one spanning the preceding night). Thus my attitudes to the two conditions can coherently differ; I can hate one package and like the other because they are different packages. My preferences are coherent. (If I had had to acquire both positive and negative attitudes to the same properties of the same conditions, rationally I couldn't. For incoherent preferences are irrational. I might be rational in arranging to change, but not in what I had become. Nor could I rationally act on such preferences, for one cannot maximize on incoherent ones. But this is not the situation in DDs and TPs.)

But I used to view all nausea alike, while now I see it as being of two types, that preceded by certain intentions and rewards, and that not. Do my preferences still covary coherently with properties of the conditions preferred? Yes. The different properties to which I now have different attitudes were always possible properties of possible conditions, only I used to be indifferent to the differences; now, I discriminate. But I am still coherent, still treat like cases alike. Offer me the TP deal again and I'll prefer to drink more toxin.

But before, I dispreferred all conditions of nausea; now I prefer some of them: aren't I being incoherent in my preferences? No. For my preferences to be coherent at a time, they need not be identical over time. I coherently dispreferred all nausea conditions before, and coherently prefer some nausea condition-types and not other different nausea condition-types now.

But isn't there something inherently unpreferable about being nauseated, this making my preferences incoherent with the properties of the thing preferred? Different question. Whether my preferences cohere with themselves--treat like cases alike--is separate from whether they track objective preferability, cohere with value facts in the world. And we already found reason to doubt both the existence of such things, and whether they would cause problems for our view.

8. Kavka's Objections

Kavka too entertained the rationality of direct preference-change in the DD, but rejected it.[33] His thinking: it would maximize on one's preferences for them to change, but since by their standards the action one is to come to prefer and intend is not preferable, they make direct revision rationally inappropriate and contingently impossible. Nor then, can one directly rationally intend to retaliate. But one can rationally arrange for retaliation to occur, even if one cannot rationally directly intend to retaliate. So it may be rational to delegate the decision to retaliate to one whose preferences embrace it, to a machine which would do it automatically, or even to a modified

33. Kavka, `Some Paradoxes.'

future version of oneself.[34] But however rational it may be to arrange self-modification, one is not rational in the process of being altered (for the above reason). And it is to arrange to be "corrupted". This can take three forms. (A) One arranges to acquire a propensity to act against one's preferences not to retaliate. Since a rational agent's actions follow his preferences, but a corrupted agent's do not, this is arranging to be irrational. (B) One arranges to add to one's initial preferences both a preference to retaliate, and a propensity to act on it. But this too is arranging to be irrational, for one's preferences in the main still dictate non-retaliation. But one arranges to have and act on a discordant preference to retaliate. (C) One arranges to acquire a revised set of preferences or beliefs, given which retaliation maximizes. So thorough is the revision that one would prefer all things considered to retaliate, making it rational relative to one's new attitudes; but not to one's old, so this too is a corruption.

So Kavka's views differ from mine in two ways. First, he thinks that, even for an ideally rational agent, that it would maximize for his preferences to change would not directly induce change; that would not be a justified automatic response to events. Rather, he must treat himself as an object or mechanism which he must arrange to get "rewired" with psychological, social, or physical "surgery". He might be rational in arranging to acquire different preferences, but would be non-rational or irrational in acquiring them. But I see no obstacle to them changing as a rational, direct response to events.

Second, Kavka sees the revisions as "corruptions"; one makes oneself irrational, even if, as in (C), one would not so judge oneself by one's new attitudes. He is clearly right on this in (A) and (B), for choosing against the balance of one's preferences is a corruption of one's practical reason; also about (C) if one revises one's beliefs on pragmatic pretexts, for this corrupts one's epistemic rationality. But there seems no justification for the charge in versions of (C) where one's beliefs stay epistemically rational, but where one's preferences are pragmatically revised. For one comes to prefer as one should in the circumstances, as measured by one's original standards. One then chooses as on ought given one's new preferences. To impugn the resultant attitudes and actions we need a standard of rationality by which one's later comportment fails. The only one offered is the standard of one's previous preferences; but as we have argued here, they are, in effect, self-invalidating as standards in PCSs. To be rational by their standards, one must supplant them with new ones.

9. Rational Self-Alteration and Moral Scruple

There may be other objections to the attitudes and actions we recommend. E.g., maybe they are immoral.[35] But an agent in the initial position of the DD undergoing these revisions might yet be rational. And if she began with moral values, had rationally to revise them, and if to be moral is to choose rationally from initially moral values, arguably she remains moral. She begins with correct values, intends to do what a moral agent would do in the circumstances, does what one

---

34. Ibid., 520-52. See also Kavka, `The Toxin Puzzle' and `Responses to the Paradox of Deterrence.'

35. Kavka, `Some Paradoxes,' 523-525.

would do, and has the character, the values, one would have in the circumstances, properly broadly conceived.

But Kavka focuses on a conception of morality which, if the agent accepts it, may make the intentions, choices, and values we advocate immoral and irrational. Suppose an agent in a DD began with the preferences (1) that harms be minimal, (2) never to prefer an action discordant with (1), (3) always to act as someone with (1) would, and (4) never to reduce the degree to which her character, intentions, and actions fit (1)-(3): then Intending and Acting would violate some of these preferences. And Kavka seems to think one must have them to be moral and rational. To have (1) is to take the right value attitude to harms; to have (2) is to embrace the Wrongful Intentions Principle (WIP: an intention is wrong if to perform a wrong action); (3), the Right-Good Principle (RGP: an action is right if a good person would do it); and (4), the Virtue-Preservation-Principle (VPP: never reduce your moral virtue).[36]

But such agents would not be in a DD, for it would not maximize for them to Intend; Intending and Acting would be non-maximizing (due to preferences (2) and (4), which require refraining from both), and so irrational by their own standards.[37] Moreover the first preference, that harms be minimal, since it requires having different preferences (for the harm-deterring effect), conflicts with the second, which forbids different preferences. And the third, to do what someone averse to harms would do, conflicts with the fourth, to keep one's virtue defined as having (1). For to minimize harms, one must prefer to retaliate so as to deter, but would find so preferring a corruption forbidden by (4). Thus a rational and moral agent cannot have those preferences, for they are incompatible in the DD; they are ill-ordered, their holder, irrational.

Kavka concluded from the DD that our schema for evaluating the morality of persons, intentions and actions was conflicted. I think it shows one of two other things. First, a rational morality must prioritize conformity to the above principles, and so a rational and moral agent must have the corresponding preferences coherently so ordered that she can maximize on them in PCSs. Note the parallels with our proposal on value realism. Even if the targets of those preferences are objectively valuable, in a morally plausible valuation they must have different values. Thus someone whose preferences tracked objective preferability would have a coherent preference ranking, maximization on it sometimes requiring sacrifice of some conditions for others. E.g., one's objectively correct preferences must not preclude acquiring a retaliatory preference in maximizing on a preference ranking in which minimizing harms was originally most preferred.

The other possibility: (2), (3) and (4) should not be read as preferences, but as principles, ways of choosing given preferences. And they should be construed as consequences of the view that a moral agent chooses instrumentally rationally from (initially) moral preferences. Thus she is now choosing morally correctly and is morally good just if she (1) began averse to harms, (2) would never intend wrongly, never intend to do something not dictated by the intentions she would have reasoning instrumentally from that aversion, (3) would always do what an agent complying with (1) and (2) would do, and (4) would never fail to have the dispositions dictated

---

36. Kavka, `Some Paradoxes.'

37. Lewis makes a similar point in his `Devil's Bargains,' 142.

by practical reasoning compliant with (1)-(3). There is no conflict between the initially correct attitude to harms and WIP, RGP, and VPP, since these just require what practical reasoning from an initial aversion to harms requires, and since that aversion is only initially correct; it becomes incorrect if compliance with those principles through rational choices requires abandoning it.[38]

10. Conclusion

An action is rational only if it expresses rational preferences, ones whose possession maximizes given one's initial preferences. Where preferences can affect outcomes, instrumental rationality requires one to adopt (or retain) whichever preferences would maximize. One may have to adopt new ones, ones whose possession advances the originals; one must then maximize on the new. All of this is justified by the normative principle standardly thought constitutive of instrumental rationality: "maximize." And nothing in the nature of preferences or the conditions on their rational revision opposes instrumentally justified changes in basic preferences. A rational psychology is plastic. How plastic depends on how efficacious various alterations to it are seen to be in maximizing on its current preferences.[39]

38. For more on this, see my `Kavka Revisited.'

39. For more on this, see my `Preference's Progress.'