THE JOURNAL OF PHILOSOPHY

VOLUME XC, NO. 4, APRIL 1993

PERSONS AND THE SATISFACTION OF PREFERENCES: PROBLEMS IN THE RATIONAL KINEMATICS OF VALUES*

"IF you loved me less, I could love you more." "You'd win more if you didn't want to win so badly." "They can tell you have no class; they won't respect you unless they see that you don't really want their respect" or "that you really love the blues (feminism, the impressionists, possible-worlds realism), which you despise." "I'd drop my gun but I can tell you aren't really willing to drop yours." "If you weren't so fond of money I'd feel more comfortable about loaning you some." "She's the perfect woman; I could love her if only I weren't so vexed by her addiction to tartan." Sometimes it seems the biggest obstacles to our getting what we want are just that we want it, or that we do not want something else. Thus, that most irritating advice: "Don't just *act* differently; *be* different," which is to say, "have different values."

I. INTRODUCTION

Reified into philosophy, we have the paradoxes of instrumental rationality, situations where it advantages you, given your current preferences, to *intend* to do an action it does not advantage you to *do*. (Call these *paradoxical choice situations*—PCSs.) You have reason to intend in the advantage of intending, reason not to intend in the disadvantage of acting as intended. Should you intend? If the benefits of intending outweigh the costs of doing, surely you should. But if rationally to intend an action, you must prefer to do it, then you must change your preferences so that you *can* then intend. To get what you want, you must change what you want. Some examples:

* For helpful discussion, my thanks to Terry Tomkow, Stephen Monk, Tori McGeer, Douglas Butler, and especially Bob Martin, with whom a conversation on terminology much altered my thinking. Thanks also to commentators on a draft read at Dalhousie University: David Braybrooke, Richmond Campbell, Wayne Fenske, Ariella Pahlke, Sue Sherwin, and Sheldon Wein.

0022-362X/93/9004/163–80

"I'll cooperate with you when you want to cooperate with me" (the *prisoner's dilemma*—PD).[1] "You say you'll retaliate if I attack, but I know you won't; you love life too much" (the *deterrence paradox*—DP).[2] "I'll give you a million dollars tonight if you come to want, by tonight, to drink this temporarily nauseating poison tomorrow" (the *toxin puzzle*—TP).[3] "I'll put a million dollars in box *A* if you will take only box *A* when I offer you box *A* and box *B*, which will contain a thousand dollars" (*Newcomb's problem*—NP).[4] These have in common with the above predicaments of ordinary life that to get what you want, you must alter your wants, that is, you must come to value something (morality, revenge, nausea, fiscal moderation) to which you are now averse (for you are exploitative, or altruistic, or you hate nausea, or you are greedy).

That you would get what you now want if you wanted differently is a prima facie reason to change your wants. Indeed, I think it *is* rational to change your preferences if this would cause the conditions targeted in the originals. Now, it is not news that preferences can rationally change. Everyone grants that, if you want this (e.g., to get into graduate school), and see that having this other thing (e.g., getting good grades) is needed to get the first, it is rational to come to want the second (good grades); also, if you like that kind of thing (e.g., Fellini movies), and see that this is such a thing (a Fellini movie), then you ought to come to like this. What *is* news is that it is rational to acquire a preference for x not because x is causally needed for or logically part of something else you already want, y, but because *preferring* x is needed to get you y. Here, intuition balks. Normally, if one sees that doing something would cause what one wants, one directly, automatically prefers to do it. Thus, one can see wanting to drink, and actually drinking, a temporarily nauseating poison tonight if drinking it will get one a million dollars tomorrow. But can one directly and automatically want by tonight to drink tomorrow where one need not drink to collect, merely want to drink? Surely one cannot so easily come to want to do something one dislikes doing and need not do to get what one wants?

Still, I believe it *is* rational to change one's preferences in PCSs, and that they *can* be changed. There are no arguments that chang-

[1] For the PD, see David Gauthier, *Morals by Agreement* (New York: Oxford, 1986).

[2] Gregory Kavka, "Some Paradoxes of Deterrence," this JOURNAL, LXXV, 6 (June 1978): 285–302.

[3] Kavka, "The Toxin Puzzle," *Analysis,* XLIII (1983): 33–6.

[4] For this version of NP (with a pre-interactive opportunity to amend one's character), see Gauthier, "In the Neighborhood of the Newcomb-Predictor (Reflections on Rationality)," *Proceedings of the Aristotelian Society,* LXXXIX (1988/89): 179–94.

ing them is *impossible.* There are some—bad—arguments for them not being under direct rational control: e.g., the argument above objects that one's current preferences, which instrumentally rationalize an aversion to a certain action, will block automatic acquisition of preferences that would rationalize performing it, even if *acquiring* them would advance one's original preferences. But this overlooks that it is only rational for one's current preferences to base an aversion to an action if it is rational to continue to have those preferences; and it is this prior question I address. I claim that one should revise one's preferences—and that they would automatically so change—if one sees that this would cause the conditions targeted in one's original preferences. But even if my preferences would not *directly* change when I see that this would help satisfy them (much as my beliefs automatically change to fit new evidence), I may yet be able to *arrange* their change—by taking a pill, undergoing reward conditioning or hypnotherapy, hanging out with the wrong or the right people. It could then be rational for me to change them *by* an arrangement.

But even if *arranged* change is *possible,* there are many objections to its *rationality.* I shall try to meet some of them; but mostly I want to explore the issues they raise. The objections:[5]

(1) It is not *always* rational to change one's preferences in PCSs: suppose that to cause what you prefer, you must disprefer it—"I'll give you what you now love if you will hate it by when I give it." In acquiring the new preference, you would also be arranging that it *not* be satisfied; and this seems to violate the rational obligation to maximize one's expected utility, to satisfy one's preferences. So you should not change.

(2) Satisfaction involves not just a condition's obtaining, but also one's preferring it. Thus, one cannot cause a preference's satisfaction by ceasing to have it by when its target condition obtains, for then it no longer exists to be satisfied. Thus, losing a preference cannot be a means to its satisfaction. So you should not change.

(3) Rational agents must satisfy *their* preferences. But agents *are* their current preferences. If something has different preferences from those you now have, it *cannot* be you. Thus, you cannot cause *your* preferences' satisfaction by changing them; *you* would no longer exist, so nothing could then count as satisfying *your* preferences. So you should not change.

The issues? First, must rational agents advance their current *and* future preferences (i.e., choose "prudently"), or only their current ones? (1) implies the former. But why care about preferences you do

[5] From Bob Bright, David Zimmerman, and from reflection on Derek Parfit on personal identity, respectively.

not yet have, especially if it would prevent your serving ones you do? Second, if utility is what you get when a condition obtains which satisfies a preference you have when it obtains, can you rationally have and advance goals whose attainment could not raise your utility? (2) says "no." But surely you can rationally have and advance goals involving your own death (e.g., providing for your family when you die), and so the impossibility of getting utility from their attainment (for the dead have neither preferences nor utility). Third, if, by (3), I *am* my present cares, I cannot care differently and still be me. Yet surely persons can improve themselves—e.g., improve the moral quality of their aims—and surely not every character change is suicide. So what must persons be to explain this?

These objections reveal a conflict between three theories of practical reason. In the first, it is rational to cause whatever one prefers; in the second, only to cause the utility of attaining ends one will want when attained, and one must have attainable ends; in the third, only to cause the utility of attaining ends one currently wants and will still want when attained. Only in the first can one change a want to satisfy *it*. In all three, one can change some wants to satisfy others, but in the second, it is only rational to change one where that will cause a higher utility over one's life given all the preferences one will ever have, not just those one had when contemplating the change; while in the third, it is only rational to change one if that will raise one's utility by the preferences one *now* has.

These theories also have different implications for the nature of persons and their identity—at least on an assumption I shall defend. The assumption: a person is a psychologically dynamical kind of thing; she can have different preferences at different times on certain conditions, namely, when her later preferences rationally evolved from her earlier ones (rather than being due to, say, nonvoluntary chemical brainwashing or other trauma, when we might say a new person has been created, the old, washed—or bashed—away). On this assumption, in the first two theories of rationality, persons can update their goals without loss of identity, though the pretexts for doing so are different in each theory; e.g., in the second, it cannot be rational to revise a goal as a means of advancing *it*. But in the third, persons are much more permanent characters: for it is never rational to change oneself if that will not raise one's utility by the preferences one *now* has.

I shall argue that only the first theory of practical reason is right on the issues. And we shall see that the rationality of preference revision and the natures of rationality and persons are essentially connected. Reflecting on the first will prove to be a method to understanding the second and third.

II. WHAT TO DO IN PARADOXICAL CHOICE SITUATIONS

First, the proposal that occasions the objections. On the standard theory of rational choice, it is rational to maximize one's expected utility given one's preferences. But in PCSs, one would do better (would maximize on one's current values) if one chose on some other basis, e.g., if one had and chose from (maximized on) different preferences. I claim that a choice is rational only if it maximizes on rational preferences, ones rational to have. A currently held set of preferences, *P,* is rational only if there is no other set, *P**, the having of which is more likely than having *P* to cause *P*'s target conditions in the order preferred in *P*. So a current preference set is rational just if "self-maximizing," if having it maximizes by its own measure compared with having any other. Thus, we apply the maximization test not just to the choice of means to ends, but also to the choice of ends (given current ends); e.g., consider

Case 1: I have many marbles, you, none. You want as many of my mine as possible. I'm tired of defending them so I offer to give you 5 if you come to want exactly 5 (so that you would pass on chances to steal more); that way, you are pacified, I can relax. Decline and you get 0. It maximizes on the desire for as many marbles as possible (*P*) to abandon it and instead desire exactly 5 (*P**); so it is rational to change your preferences. This will cause a better result by your original preferences. (Note: you will also attain an outcome—getting 5 marbles—you will then want by when it obtains.) Likewise, it is rational in the PD to come to prefer to cooperate with just those whom this preference change would make cooperate with you; in the DP, to prefer to retaliate against those deterrable by one's so preferring; in the TP, to prefer to drink; in NP, to prefer to leave the extra one thousand dollars.[6]

III. OBJECTION (1): REVISIONS TOWARD INUTILITY

Case 2: Same story except you can have 5 marbles just if you come to hate marbles. It maximizes by your preference for as many as

[6] This resolves PCSs in ways thought impossible on the standard theory (this impossibility often taken to prove the theory self-defeating and so false): it maximizes to change one's preferences, then maximizes (since one now prefers differently) to cooperate in the PD, to retaliate in the DP, to drink in the TP, and to be a one-boxer in NP. We can thus rationalize as maximizing the actions Gauthier sought to rationalize, without need of his reconception of rationality. (He thinks it maximizing and so rational to adopt maximization-*constraining* dispositions, then rational to act *non*maximizingly because rational to act on dispositions rational to adopt. Many find the last bit implausible). Rather, even maximizers can be reliable cooperators, threateners, and compromisers. There are also implications for the rationality of plan following. For criticisms of Gauthier, see my "Retaliation Rationalized: Gauthier's Solution to the Deterrence Dilemma," *Pacific Philosophical Quarterly,* LXXI (1991): 9–32. For more on the possibility and rationality of preference revision, see my "Preference-Revision and the Paradoxes of Instrumental Rationality," *Canadian Journal of Philosophy,* XXII (1992): 503–30.

> possible to come to hate them, for then you will at least get 5. But by then, you'll not want them, though I will force them on you. In changing your preferences, you would cause a better result by your original preferences (for as many marbles as possible), but would get an outcome you would then wish *not* to get; you would get 5 marbles, but prefer 0!

Your choice: (a) not getting what you want (many marbles) because I shall not give it to you (because you still want it); (b) not getting what you want because, though I shall give you what you first wanted (many marbles), by then, you will not want it (for you will have come to want 0 to make me give you many). You may have either your current or your later preferences unsatisfied. Surely there is no reason to favor either. Since revising one's preferences is not (uniquely) rational here, my thesis that it is always rational to revise preferences if that will best satisfy them is too strong.

But even here, by changing what you want, you cause what you first wanted. And when choosing from those wants, it is rational to do what will serve them.

> *Objection:* this involves choosing to be frustrated, to get a low utility by one's foreseeable preferences; but a *maximizer* should make his utility as high as possible.
> *Reply:* it will not be low by your original preferences, those defining all you care about when now choosing among preferences.
> *Objection:* if preferences are only rational if self-maximizing, surely this holds for prospective preferences, too. But then, (i) if having the preference for 0 marbles will make you get 5, it is not self-maximizing; it *prevents* its own target. Thus, surely, (ii) a rational agent will adopt preferences maximizing on his originals *except* where that would not maximize on the *new* ones.
> *Reply* to (ii): rational agents must always choose from (maximize on) the values they *have.* It maximizes on *those* to revise them even in case 2. And even if the *new* values may prove non-self-maximizing from *their vantage,* that is irrelevant to whether to adopt them *now* (from the *previous* vantage).
> *Reply* to (i): (i) has a false premise, that the new values are not self-maximizing. For *their* revision would not advance their own satisfaction (I shall still force marbles on you). Thus, they are not even irrational from *their* vantage. True, one cannot satisfy them. But values are not rational just if satisfiable; they are rational just if having them would best satisfy those one has when choosing whether or not to have new ones.

It might have been that the new values *would* be advanced by *their* revision, e.g., if the agent who now prefers 0 marbles later met someone who would refrain from forcing yet more marbles on him if he came to prefer that they be forced on him. The new preference

for 0 marbles would *then* be irrational (for revising it would maximize)—but not *until* then, not from the vantage of his first values in case 2's initial position. So it is not as if he would be rationally (and paradoxically) acquiring irrational values. He would, rather, still be rationally (because maximizingly by his original values) revising his originals; and the new values would still be prospectively rational (because adopting them would maximize on the originals). Once he has the new ones it would be rational to change yet again, for that would then be self-maximizing. But this just means he should change again, not that he should not have the first time.

I say preference revision is rational even if it frustrates foreseen preferences; one need only aim at what one *now* prefers. But surely if I foresee preferring something to which I am now averse, that changes what is rational for me now. Now I want to spend all my money on movies. When I am fifty I shall wish I had saved it and bought a house. Should I not temper my current desire for movies with my future desire for a house? Call choosing now in light of both current *and* foreseen preferences, "being prudent." Is that not always rational? So is it not irrational to acquire a preference one knows will be caused by its acquisition to be dissatisfied?

But prudence *cannot always* be rational. For if it is, one must treat one's foreseen preferences as if they are, like one's current ones, relevant to current choices. I now prefer movies (A) to a house (B); but I shall later prefer *B* to *A*. I cannot have both. How am I now to choose rationally by both preferences? That would be choosing as if I preferred *A* to *B and B* to *A*. But one cannot maximize on ill-ordered preferences.[7] Perhaps one need only be prudent if that would not conflict with maximizing on current preferences. (But why should it matter to me now that I later get what I do not now care about? Not, for instance, because I would get utility from it, for it is not now utility I find worth caring about.) Our cases, however, are precisely ones where, to get what one now wants, one must frustrate a later want. So this weaker prudence is not violated in adopting self-defeating preferences.

IV. OBJECTION (2): A BIPARTITE STRUCTURE FOR PREFERENCES?

Surely a rational agent, in wanting *x*, must also want still to want *x* by when she gets *x*. For must she not maximize her utility, try to cause conditions satisfying her preferences? And must one not *have* a preference if a condition is to satisfy it, if one is to get utility from

[7] Parfit, in *Reasons and Persons* (New York: Oxford, 1984), pp. 155–6, almost makes this point. For more on this critique of prudence and on its differences from Parfit's, see my "Rationality and Prudence (Or: One's Life Going As Well As Possible For One; The Very Idea)," presented to the Canadian Philosophical Association, May 1991.

its obtaining? But then one can *never* advance a preference by abandoning it.

Unfortunately, the notion that rationally to prefer x one must also prefer still to prefer x by when it obtains (or that one rationally must preserve the preference for that time) has absurd consequences. First, if one had both some preference and the preference to keep it, one could not get into a PCS where it maximizes on one's preferences to change them. One's preference that one's preferences not change makes changing them antimaximizing (if it is at least as strong as one's other preferences). But PCSs *are* possible.

Second, if every folk preference for an outcome also involves a preference still to prefer it upon the former's satisfaction, preferences for events after one's death would be irrational. For one would, in effect, be preferring a condition entailing that one is dead (e.g., that one's family inherit one's fortune) *and* alive (that one is still around to prefer that they inherit). Thus, preferences for postmortem conditions must be ill-ordered (except, possibly, for Cartesian dualists)—a *very* odd consequence.

Third, no action could be rationally justified as maximizing on a preference concerning a postmortem condition. For since the preference for that condition and the preference to still so prefer postmortem cannot be jointly satisfied, nothing could count as advancing their joint targets. But many of our preferences can only be rationalized as ones for things that cause postmortem conditions. For example, one prefers to buy life insurance in order to provide for one's family should one die. But if preferences for postmortem events cannot rationalize ones for premortem events, the latter cannot be rationalized. Why buy life insurance if not to benefit survivors? To cause something enjoyable while one lives, like one's family's security? But that is just their security from poverty after one's death. So one cannot rationally prefer their security without rationally preferring that they not be poor after one's death. But on (2)'s premises, one cannot. The justificational impotence of preferences for postmortem conditions, then, would render irrational those preferences which for many of us define our strongest duties. So there had better be a way to have justifying preferences for conditions that will only obtain after one no longer prefers them, or we are crazy.

Worse, we could not ascribe such preferences to agents (though they seem to self-ascribe them, e.g., with wills). For a preference is *defined* by what actions would express it: rational actions reveal preferences by maximizing one's expected utility, by increasing the probability of certain outcomes; one "has a preference" for the outcomes made more probable by a rational action. This action would

cause that condition, *x*, so a rational agent who knows that and so acts must prefer *x*. But since no actions could rationally serve or express preferences for postmortem conditions, no action could reveal that one had them. Indeed, one *could not* have them. For preferences are *reasons,* things possibly relevant to the rationality of choices; choices are actions, so preferences must be able to be reasons for actions. As such, they *justify* actions. Thus, preferences are the same just if, necessarily, they justify the same actions in the same circumstances; actions are the same just if, necessarily, they make the same causal difference to the likelihood of events. If an event is impossible (e.g., one dying and not dying, which is, by (2), what one prefers if one prefers a postmortem condition, since one prefers that something happen when one is dead, and prefers still to be around so to prefer when it happens), nothing could increase its likelihood. So no action could be justified as doing that. So no preference could justify an action as doing that. So, qua reason for choice, there can be no preference for a logically impossible event; any preference of mine qua reason for *action* must be one *some possible* action of mine could advance.[8]

I take it that these consequences of objection (2)—that there are no PCSs, that preferences for postmortem events cannot rationalize actions, that one cannot exhibit or rationally have preferences for postmortem events—are false, and that it refutes a theory of rationality if they follow from it, gives plausibility to one if they do not. So what has gone wrong?

Objection (2) reads the rational duty to maximize *expected* utility as the duty to maximize *utility.* Since utility = a condition's obtaining while preferred, rational agents must cause preferred conditions to obtain while preferred. (2) also assumes that utility = preference

[8] This may seem false. Can I not prefer that the Jays beat the *A*s without my help? But surely no action of mine can increase the chance of this preference's satisfaction: if I help them, they will not have won without my help; if I do not, I will have done nothing. And cannot I prefer that President Kennedy not have been shot? But surely no action now could make his shooting less likely? (Thanks to Sue Sherwin and Sheldon Wein for these examples.) I am not so sure, however, that my actions cannot affect these conditions. Maybe I could time-travel and warn Kennedy. And maybe the Jays want me to finish this paper, will play better if they hear it is going well; then I could make sure I do not help them by pouring coffee on my word processor. But we need not crush every example. For there are other affective attitudes than preferences qua reasons, e.g., regrets. The former are often distinguished in this way: the target of a preference must be *possible* in the circumstances. You cannot have reason-giving preferences about the past, the contradictory, the known-to-be-contrary-to-fact—see Myles Brand, *Intending and Acting: Toward a Naturalized Action Theory* (Cambridge: MIT, 1984). So if something seems like a preference, but cannot be by our criterion, we can plead that it must be some other affective attitude.

satisfaction. But if utility = a condition's obtaining while preferred, this means a condition cannot now satisfy a preference no longer held. This fits a view about the "reward condition" for preferences: preference satisfaction involves a pleasant feeling (or state) caused by (or consisting in, or supervening on) the known current obtaining of a condition one currently desires. And surely one cannot cause the reward just by causing a state of affairs; one must also have an attitude to it which, when its satisfaction condition obtains, causes reward.

It might be objected that this is just sophomore hedonism. In having and acting on a preference, agents do not necessarily seek pleasurable feelings; they merely seek to procure what they prefer. Sometimes this is a feeling, but other times just a satisfaction condition—satisfaction in the logical sense, not the emotional or phenomenological; satisfaction as the obtaining of a condition that would make true the proposition to which one's preference is an attitude. But this is not the real problem. For even if feelings are not necessary to every reward condition, one's having a preference when a condition obtains may be necessary to its *being* a satisfaction condition. Otherwise, what preference does it *satisfy?*

Well, why not an ex-preference? Because if, per (2), preference satisfaction = utility, this would mean that my utility rises if any condition I ever preferred obtains. But when the target condition now obtains for a preference I only had as a child (e.g., I now, as an adult, inherit my uncle's baseball cards, which I only wanted as a child), surely my utility does not rise. It would only if I still wanted that condition. So while this proposal lets a preference be satisfied when it (or even its holder) has expired (as in executing wills), it entails absurdities about utility. My utility does not rise from the obtaining of a condition I no longer prefer—especially if I am dead.[9]

Here then is what must be the correct reply to (2): first, preference satisfaction and utility are not equivalent. Having utility entails that a preference is satisfied, but that a preference is satisfied does not itself entail having utility. Thus, preferences can be satisfied by the obtaining of their target conditions even if the preferences or their holders have expired by then. It is just that satisfaction then does not yield one utility.

But this may seem no help. For (2) sees rationality as obliging the maximizing of utility *simpliciter.* And a condition will yield *utility* only if one will prefer it when it obtains. So it is irrational to prefer

[9] See Bob Martin, "Harming the Dead" (Dalhousie University, 1990).

or cause conditions one will not prefer or be alive for when they obtain. We are only justified in doing things likely to cause what we currently prefer *and* our concurrent preferring of it; and we can only rationally prefer conditions whose obtaining entails our then preferring them. (2) is wrong, however, in its reading of the standard theory of rationality, which obliges maximizing *expected* utility. That obliges maximizing not *utility,* but *probable satisfaction.* One has a certain *utility* if one now prefers in some degree a currently obtaining condition (and, perhaps, knows it obtains). One's actions have a certain *expected utility* if they make probable in some degree something one now prefers. Rational choices maximize the product of the current preferability and the probability of conditions, maximize expected utility, not utility. Thus, one can rationally now have and choose from preferences one will not have upon their satisfaction. For one *prefers* a condition just if one will do what one thinks most likely to cause it if one is able and rational. The rational *expression* of a preference is doing what seems necessary to cause its target. Rationality *justifies* one in doing what (one thinks) will cause a condition *now* preferred. One need not prefer it *when the condition obtains;* rather, an action is made rational by its being thought to help cause (possibly much later) what one now prefers to be caused (however much later). Since rational agents aim at expected utility or satisfaction, losing a preference can rationally help satisfy it. One will then get no utility from its satisfaction, but no matter; one's only rational obligation is to seek its satisfaction, not utility from its satisfaction. Thus, one can rationally have and choose from preferences for postmortem conditions, and can rationally renounce preferences in order to secure their satisfaction.

V. THREE CONCEPTIONS OF RATIONALITY

The standard theory is one thing; but these objections can be taken to speak for different theories of rationality altogether. I now consider these and defend the received one against them. First, some terms: 'one's having a certain level of utility' = one's knowing some conditions now obtain plus one's now preferring them in some degree (or order). 'A preference is satisfied' when a condition obtains that anyone ever preferred, 'concurrently satisfied' if the condition obtains *while* preferred. 'The expected utility of an action' = the sum of the products of multiplying the utility of each possible condition (as measured by current preferences) by its probability of obtaining given the action. 'One prefers a condition' just if one of one's possible rational actions would raise its likelihood. 'One now prefers x to y' just if, given a choice, one would rationally act to make condition x more likely than y.

On the received theory, it is rational to: maximize one's individual expected utility, maximize the probability and current preferability of conditions given choices (MIEU). But here are two other theories of rational choice. First: maximize one's expectation of individual utility, maximize the probability and *con*current preferability of conditions (MEIU). Second: maximize one's expectation of individual utility by the rigid measure of current preferences, maximize the probability and concurrent preferability of conditions by *current* preferences (MEIUCP).

MEIU and MEIUCP are recommended by the dogma that rational agents care about and aim for *utility,* the enjoyment of conditions, not just *satisfaction,* the obtaining (whenever) of conditions now preferred. But they differ on what utility is worth caring about, and so on which choices are rational. MEIU obliges prudence: rational agents aim at satisfying their current *and* future preferences concurrently with having them. It sees no rational difference between them, so one must concurrently maximize on both. This can mean sacrificing the concurrent satisfaction of a current preference to that of a future one. The view that there is no rational difference between one's present preferences, future ones, and those of others present or future, is utilitarianism: maximize aggregate utility, concurrently maximize on everyone's present and future preferences. MEIU is stoicist: one should only prefer what one can get, for having unsatisfiable preferences reduces one's possible utility. MEIUCP also says one must aim for utility, but defines the only kind worth caring about as that from the concurrent satisfaction of one's current preferences. As in MEIU, one only gets utility from satisfying a preference had *when* satisfied, not from satisfying ex- or others' preferences, nor from satisfaction postmortem. But neither is there utility worth wanting now from satisfying preferences one does not *yet* have. MEIU requires one to treat one's future preferences as if they were current; MEIUCP, that one *not.*

We can now see each objection as touting one of these other theories. (1) says it can be irrational to acquire a preference one knows will thereby be caused to be unsatisfied, this confirming MEIU: it is only rational to change if that maximizes one's *total* utility over present and *future* preferences. But the future preference caused by its own acquisition to be unsatisfied may be stronger than whatever current one is caused to be concurrently satisfied by the change (whatever 'stronger' means in what is in effect an interpersonal—current versus future self—utility comparison). Since the change then yields a net decrease in one's *total* utility, it is irrational by MEIU. (2) asserts MEIUCP: one cannot rationally arrange to lose

a preference as a way of raising one's utility by *its* measure; for if it does not survive, no future condition will give one the utility of *its* concurrent satisfaction.

By MEIU and MEIUCP, one cannot rationally have or advance preferences for postmortem conditions, for one can get no utility from such conditions. Both theories explain why one cannot be happy when dead, though only by making it impossible (rationally) to prefer conditions entailing one's death.

What decides the correct theory? All seek to describe means-end choosing and so must cohere with the concept of means-ends rationality. It has three parts: first, when a rational person chooses between x and y (where only they are at stake), if he prefers x he must choose x; second, given a choice between actions likely and unlikely to cause what he prefers, he must choose the former; third, when choosing among actions different in the probability and desirability of their consequences, he must combine parts one and two—he may only take great, expensive risks (expensive by what, among what he prefers, he might lose) for strongly preferred conditions, only refuse small, inexpensive risks for weakly preferred ones, etc. Methodologically, to show that rationality involves more than these principles, one must deduce the extra from them and from beliefs about the circumstances of choice. Why 'deduce'? Because if you do not deduce the extra from the idea of means/ends reasoning, you will end with a concept of more than *instrumental* rationality, of rationality given preferences. So, if you think instrumentally rational agents must now advance not just their current preferences, but also their future ones or other people's, you must show that they could not obey these principles without advancing such preferences.

Why is *this* the theoretical kernel? Because having a preference *is* being disposed to choose like this. (See above on how preferences qua reasons must be revealable in choices.) To prefer x *just is* to be inclined to chose x or an action making x more likely, given a choice. If one knowingly made something else, y, more likely, one must really have preferred y. The objects of possible preferences are just the conditions such choices target. So a theory of rationality violating this kernel is crippled as a theory of means-ends reasoning, for it is sometimes unable to ascribe ends as reasons for choices to agents. Thus, it is a theoretical advantage of MIEU that agents who choose by MEIU or MEIUCP will sometimes violate part two of the kernel, fail to do things likely to cause conditions currently preferred, e.g., where this requires preference revision or (less contentiously) the causing of postmortem events.

The correct theory must also fit the data about possible prefer-

ences, rational choices, and utility. Again, MIEU has the edge. First, it explains why the dead cannot be happy: their (past) preferences can be satisfied postmortem, but no utility then results, for utility is the known satisfaction of a current preference by the current obtaining of its target. Dead people do not prefer, so they cannot have utility. Neither does one's utility increase now simply by satisfaction of ex-, future, or other people's preferences (unless one now prefers their satisfaction). Second, it can describe prudence and altruism without conflating preferences with different provenances. The prudent are MIEUers who now prefer that their future preferences be satisfied; altruists are MIEUers who now prefer that others' be satisfied. Third, it explains why projects like rationalizing duties to others and to one's future self are nontrivial: it is *hard* to show that either must be preferred simply in being instrumentally (means-ends) rational in the circumstances; one must show that it advances current preferences to advance future ones or those of others. Fourth, it explains why some preferences are not satisfiable by their own revision: some are preferences to prefer some condition while it obtains.

So, to be rational is to MIEU, to maximize the probability and *current* desirability of conditions by one's choices, to make most likely that the future will be how one *now* wants it. It is not also to make likely that one will prefer conditions upon their obtaining, that one will be happy (get utility) then. Preferences for postmortem conditions are not irrational or nonjustifying just because one will not exist to prefer and appreciate such conditions. That would only make irrational the preferences expressly to exist and to appreciate those conditions *as of* when they obtain. And when these preferences target postmortem conditions they are crazy in their own right; they are preferences to be alive while dead, to be happy while being nonexistent. They are, in short, ill-ordered. But many preferences are just for conditions, not for their obtaining concurrently with some preferences. So one *can* change *them* to maximize on them.

I have also independently confirmed the rationality of preference change. For what would rationality have to be like if such a change was *never* rational? Since being practically rational is just aiming to cause currently desired ends in the order preferred, for it never to be rational to change preferences, it must never advance them to change them, i.e., it must be logically necessary to every objective that one prefer it when attained. Only this would make it necessary that losing those preferences is irrational. That would make it impossible rationally to prefer postmortem events. But that *is* possible. So

it must not be rationally obligatory that one prefer one's objectives when attained. Thus, since to be rational is to advance one's ends, and since changing them can do this, it must then be rational to change them.[10] This must be right if rational preferences for, and actions causing, postmortem events are possible, as they are. I have deduced and tested a consequence from the theory, and it survived. I have also shown that the real aim of all rational choice must be satisfaction, not utility.

VI. REPLY TO OBJECTION (3): PERSONS AND THEIR PREFERENCES

There remains the objection that people *are* their current preferences; change them and one ceases to be. So one cannot satisfy a preference of *oneself* by ceasing to have it. To meet this I must discuss personal identity. But it proved not generally necessary to a preference's being rational, advanceable, or satisfiable that one will exist when its target obtains (unless it is a preference for a condition *in which* one exists). Thus, some preferences for postmortem conditions could rationalize suicide: if I have a contagious disease and can only save others by immolating my diseased body, suicide is rational if I prefer to save them, even though their being saved is a postmortem condition I shall not exist to savor. But your preference for marbles is *not* advanced by your cessation; dead people cannot "have" marbles. So changing a preference had better not mean your not existing as the changee.

Consider, then, the problem of personal identity. By the principle of the indiscernibility of identicals, A is B just if they have the same properties. But early person A and later person B have different ones, so A cannot be B. There are three theories of how A could yet be B: first, if A and B share some essential, unchanged thing—same body, memories or soul;[11] second, if A is uniquely spatio-temporally contiguous with things contiguous with things . . . contiguous with B;[12] third, if A is uniquely psychologically related to B.[13] In the first, A is B in spite of changes because there have been no changes of essence, and people are their essences, not their accidents; in the second, because 'A' and 'B' refer to the same set of spatio-temporally contiguous stages. The earlier stages have different properties

[10] Assuming (as I believe) that they *can* change (see my papers in fn. 6); otherwise, the ought-implies-can principle might exempt one from a rational obligation to change.

[11] Richard Swinburn seems to hold the sameness of essence theory. A is B if they have the same soul.

[12] Thus Bernard Williams: one *is* one's brain-in-body.

[13] Locke and Parfit (though Parfit thinks R-relatedness is not personal survival, but something else, the thing really worth caring about; however, I shall take B's being uniquely R-related to A for an *analysis* of A's survival as B).

than, and so are different stages than, the later. But persons are sets of spatio-temporal parts, of stages related by spatio-temporal contiguity. Like his physical parts, a person's temporal parts can have different properties, yet be parts of the same person. In the third theory, *A* is *B* because *A* has some appropriate psychological relation, "*R*-relatedness," to *B*, and people are *R*-related psychological stages. Trouble only arises for us if preference changes violate *R*-relatedness, and if it is needed for identity (e.g., if sameness of body is not enough). Suppose it necessary. Do we violate it?

One might argue that no matter how radical someone's preference changes, the rest of his psychology may stay sufficiently *R*-related. He might yet have the same memories, beliefs, etc. But this may be impossible: to have certain preferences may just *be* to see and remember things in certain lights, to be disposed to judge in certain ways. Thus, people with different values may remember things in different lights, as pleasant, awful, neutral. And they may believe different things—about what is right, wrong, rational, relevant to choice problems, to what to believe. Of course, they may undergo slight preference changes with little change in their total psychology. Still, the identity problem arises even for slight changes. Besides, someone might find it rationally necessary to undergo massive preference change, and so massive global psychological change.

A and *B* having different preferences *can* be reason for thinking that *A* is not *B*. But it is not decisive. Consider the deterrence paradox: you are Christ; you prefer that harms be minimized. I am the evil empire; I shall attack your people unless you would retaliate. You could not rationally do that if you kept preferring harm minimization; better that you not add retaliation's harms. But you could scare me off and prevent all harms by preferring to harm if attacked, by acquiring Dr. Strangelove's preferences. But he and Christ are as different as people can be. Someone with his values would, prima facie, not be Christ. Christ loves people, Dr. Strangelove, violence on provocation. Surely your having Strangelovean values proves you are not Christ.

It *is* prima facie evidence. And there *are* ways for you to acquire such values and cease to be Christ; e.g., in a non-PCS, against your will your brain is neurochemically wiped of what made it instantiate Christ's psychology, reconditioned into Strangelove's. You are not Christ here. But you are in the first case. What's different?

In the first case, your values change through rational reflection in an expression of your original values; the change serves them; they change through an activity characteristic of persons, the exercise of their rationality. Such preference change is akin to changes in one's

beliefs upon getting new evidence, and to changes in one's desires upon seeing that something has a property one already prefers in things, or a power to cause what one already prefers. These, surely, are paradigms of *R*-relations between mental stages. For one's beliefs to change in ways that preserve one's psychological identity is surely for them to change responsively to evidence. If they cannot change in *this* way and one's later and earlier psychologies be *R*-related, I do not know how they *ever* could. *R*-relatedness is just relatedness of psychological states by processes that distinguish a psychology as such, and of these, rational changes of such states are paradigms. It is as much a personalizing and person-individuating trait of you that you are rational—disposed to change in circumscribed ways in response to your perceived environment—as that you are someone who (*now*) hates harms. For persons, unlike instants, dates, starting times, and appointed hours, are inherently dynamical objects. Their nature is that they may change. They are dynamical as believers in how they respond with beliefs to evidence, and as valuers as we normally understand them, i.e., in how they respond with desires to the recognition that things have traits they already desire, or have powers to cause such things. But they are also dynamical in what values they will acquire when they see what they must value to advance their current values. One way for an agent's later beliefs to *R*-relate to his earlier ones is for the earlier to justify the later (by making their truth probable) and to cause the later's formation *because* they justify them. A similar causal/justificational relation can *R*-relate one's earlier and later desires, though there are three kinds of desire justification. Later desires can be justified as being ones for things with traits already desired in things earlier, for things with the power to cause conditions earlier preferred, or as being ones the having of which can cause conditions earlier preferred.

R-relatedness is sometimes read as resemblance.[14] But this is unnecessary. My belief that *p* on some evidence *e* is very unlike my later belief that not-*p* on new evidence *e**. What *R*-relates them is the rational evolution of the one from the other caused by new evidence, not their resemblance.

So instrumentally rationalized preference changes are paradigms of *R*-relations—at least, on MIEU: rational agents are expected satisfaction maximizers; they try to make probable what they prefer when they choose. Since changing preferences can maximize on their originals, their preferences can rationally change. If, as surely

they are, persons are dynamical—if they are sets of R-related psychological stages—this is a paradigm of identity preservation by R-relatedness. Unlike homeostats, people can alter their "set points"; indeed, sometimes they *must* to remain persons. A psychologically frozen person would be *broken* in a way threatening her very personhood.

Now, MEIU and MEIUCP also admit rational changes of preference, though, as we saw, different ones than MIEU. So they would count some MIEU changes as changes of person, not just of preferences. But these theories imply falsehoods about persons: that they cannot now care about what happens after they die, nor reasonably now act differently because of the postmortem effects of different actions. False too, then, are their implications on personal identity.

Rational psychodynamics are paradigms of R-relatedness; forced changes not based in a psychology's autonomous self-updatings by rational/causal evolution, candidates for its violation. But are there not identity-preserving R-relations that are not rational evolutions? For example, surely persons after the onset of psychoses and neuroses need not be different persons? And surely if I develop an irrational belief or preference, I do not cease to be me? If a person has a complete breakdown of personality, of the rational integration of his psychological stages and personality traits, we have no person any longer. But mild irrationality is no threat to identity. Where to draw the line is inherently indeterminate, needing stipulation on pragmatic and "forensic" grounds—not part of the objective metaphysics of persons. But all theories of personal identity go soft on matters of degree. The issue is, what kind of thing is it differences in which, however divided into degree and kind, individuate persons? I claim persons are rationally self-updating psychologies, and that instrumentally justified changes in a person's preferences are paradigms of identity-preserving psychodynamics. Thus, B is the same person as A so far as B's psychological states rationally evolved from A's.

DUNCAN MACINTOSH

Dalhousie University