Practical Reason, Instrumental Irrationality, and Time

(Forthcoming in *Philosophical Studies*)

Manuel Vargas

Standard models of practical rationality face a puzzle that has gone unnoticed: given a modest assumption about the nature of deliberation, we are apparently frequently briefly irrational. In what follows, I explain the problem, consider what is wrong with several possible solutions, and propose an account that does not generate the objectionable result.

Consider the following model of practical rationality that relies on standard instrumental reasoning, where— allowing for consistent tense changes and the like— I(e) is read "Intends end e" and B(e->m) is read "believes that end e implies the necessary means m" and I(m) means "intends to m":

I(e)

B(e->m)

I(m)

Call this Case 1.

Case 1 can be distinguished from at least two different forms of practical rationality. One form is for an agent to intend not to undertake the means he or she views as necessary for the end (I(~m)). For example, if I intend to go to my office on campus today, and I believe that leaving my house is necessary for me to go to the office, but I also intend *not* to leave my house, this is irrational. Another form of irrationality occurs when the agent simply fails to intend the necessary means (~I(m)). For instance, if I intend to call my brother right now, and I believe that picking up a phone is necessary for calling him, it would be irrational for me to fail to intend to pick up a phone. It is irrationality of this sort— irrationality in failing to undertake the means— that is my focus.

Irrationality of the sort under consideration can be modeled in the following way:

I(e)

B(e->m)

~I(m)

Call this Case 2. What makes Case 2 a standard model of this form of irrationality is that the agent fails to intend the means required for the intended end.[1]

To forestall complications that are orthogonal to this paper, two qualifications must be added. First, we will assume that the I(e) was not acquired irrationally. This rules out complications brought on by irrationally acquired intentions or plans "bootstrapping" agents into rationality when downstream intentions are rationally acquired.[2] Second, unless otherwise stated the assumption is that the considered agents do not or will not have reason to reconsider or reject I(e). Since a plausible principle of instrumental rationality only requires that agents who have I(e) and B(e->m) either acquire I(m) *or* give up I(e),[3] this assumption allows us to ignore contexts described by the second half of this disjunct.

Now consider the following plausible assumption:

[1] Though both include qualifications that are either not relevant in the present context or addressed below, see Christine Korsgaard "The Normativity of Instrumental Reason" in Cullity and Gaut, eds. *Ethics and Practical Reason* (Oxford: Oxford, 1997), pp. 236-9. and R. Jay Wallace "Normativity, Commitment, and Instrumental Reasoning" <www.philosophersimprint.org/001003> *Philosophers' Imprint* 1:3 (Dec. 2001), p. 24-5.

[2] See Michael Bratman, *Intentions, Plans, and Practical Reasoning* (Cambridge: Harvard, 1987), pp. 24-27, 86-87.

[3] See Wallace "Normativity, Commitment, and Instrumental Reasoning" p. 17.

(RT): Practical reasoning takes time.

For our purposes, we can suppose that the scope of reasoning includes the acquisition of an intention as a suitable consequence of a piece of practical deliberation. If RT is true, then an agent who fulfills the description given in Case 2 might be treated as practically irrational because the agent has not had enough time to complete the reasoning and acquire I(e). That is, the agent might be one who has not yet (but perhaps will) acquire the I(m). Thus, given RT and the model of practical irrationality described by Case 2, people are likely frequently irrational when engaged in practical reasoning. The duration of the purported irrationality varies depending on the nature of the intention. For present-directed intentions, it will generally be momentary. For future-directed intentions, it may be much longer.[4]

Call the account whose description generates the problem *The Simple View*. On this view every time an agent forms an I(e) where the relevant I(m) does not already exist, there is a period of irrationality that follows until I(m) is acquired. Since deliberators often lack the intention to pursue a means necessary for some end prior to concluding that a means is necessary, it likely that deliberators — ourselves included— are very frequently (though perhaps often only briefly) practically irrational.

For many, this consequence of *The Simple View* calls for a repair. As we will see, such a repair is not as straightforward as it may initially seem. However, given the prevalence of models of irrationality that lead to the problem, some attempt to respond to it is in order. Still, there may be some for whom the possibility that we are subject to the widespread irrationality entailed by *The Simple View* and RT is not a reason to reject the model of irrationality described in Case 2. If

[4] On the distinction between future- and present-directed intentions, see Bratman, *Intentions, Plans, and Practical Reasoning*, p.4.

this is a case of one philosopher's interesting conclusion being another's *reductio ad absurdum*, then what follows is only for those who reject the "interesting" conclusion.

One potential solution is *The Fix is In*. On The Fix is In, we should hold that the predicate 'practically rational' and its various forms do not apply to cases where an agent has not had adequate time to acquire I(m). On this view, an agent might be rational until he or she acquires B(e->m). At that point, the proverbial clock starts running for the agent to acquire I(m). During this time, the agent is neither rational nor irrational, but in a state whose resolution will settle the agent's rationality with respect to the considered case. Once adequate time has elapsed, depending on the outcome of the agent's reasoning, he or she is either instrumentally rational or irrational. Until that time, however, it would be a mistake to describe the agent as rational or irrational, at least with respect to the considered piece of instrumental reason.

There are two major drawbacks to this approach.

First, the strategy seems to resolve the puzzle by disregarding the intuitions that give rise to it. On this approach, we simply do away with the apparent irrationality by fiat. This will strike some as an unacceptable resolution, for it drives a wedge between our philosophical accounts of practical rationality and the intuitions that provide the framework for our philosophical reflection. In particular, the stipulated solution appears *ad hoc*. The proposed solution amounts to an unprincipled attempt to rule out the problem by fiat, akin to eliminating poverty by declaring that we will not call the impoverished "poor." Moreover, on The Fix is In we are lead to the unhappy result "solving" one counterintuitive result (that we are frequently briefly irrational) by providing a further counterintuitive solution (one that requires giving up on the agent's being rational) solely for the sake of eliminating the prior counterintuitive result. Consider that (excluding various marginal or non-standard cases), ascriptions of instrumental rationality are "permissive." That is, if you are rational that just means you are not irrational. Normally, when an agent learns that he or she is not irrational, this is cause for relief because it implies that the agent is rational.[5] Thus, the challenge is to explain how an agent that should be treated as rational could come out that way given various proposal about instrumental rationality in light of RT. If the trouble with The Simple View is that it entails that the agent is irrational, The Fix is In does little better because the agent still does not come out rational (even though the agent is not irrational, either).[6]

A second drawback of The Fix is In is that it creates the difficult burden of specifying what "adequate time to acquire I(m)" must mean. If "adequate time" is indexed to non-individual characteristics such as species- or kind-typical speed of intention acquisition, we might imagine that slower intenders (and perhaps on occasion, faster intenders) could justifiably object to standards of rationality disconnected from the specifics of their particular deliberative context, including their self. On the other hand, if "adequate time" is indexed to individuals, a wide range of possible factors (including death and deliberative incompetence!) might make *no* amount of time sufficient to acquire I(m). Irrationality would threaten to provide its own exculpation.[7]

A more promising solution is *Time Enough*. According to Time Enough, an agent is irrational at a time when, by the agent's own lights, it is too late to act on I(m), were the agent to I(m) at that time.[8] On this proposal, a practically rational agent could intend some end, believe

[5] An alternative way to put the point that follows is this: To the extent to which one has not violated norms that prohibit rationality, this is to an agent's rational credit. So, on a permissive view, an agent ordinarily gets credit unless he or she is irrational. The problem with The Fix is In is that on an ordinary (i.e., permissive) understanding of rationality, the agent is failing to get due rational credit in the considered case. Claiming that the agent is neither rational nor irrational is no advance, because this has the effect of making the agent is *ineligible* for rational credit when what the agent should be getting is rational credit.

[6] My thanks to an anonymous reviewer for emphasizing the point about the permissiveness of rationality ascriptions, and for encouraging me to clarify this part of the paper.

[7] Note that this difficulty holds even if one rejects the permissive interpretation of rationality.

[8] This seems to be the spirit of Bratman's discussion of "filling in" plans over time. See his *Intentions, Plans, and Practical Reasoning*, p. 31. However, Bratman does not specify that the agent must view him or herself as having

that a means is required, and not intend the means for years— just so long as it is that case that by the agent's lights, there is still adequate time to form I(m) and act on it. Time Enough allows us to deal with examples like the following case:

> David intends to die with a copy of Plato's *Phaedo* in his hands. It is not a mere wish or hope, but something he firmly intends to do (perhaps because he promised his first Plato teacher on his deathbed that he would do so). Since David made the promise as a young man and expected to live a long life, he did not consider how we would go about keeping his promise. David lives out almost his entire life without acquiring a copy of *Phaedo*, or knowing anyone who has one, or ever being around one. But, in his old age he recalls his promise. Since he has also come to think that he will die in the forseeable future, he views this as a reason to consider how to ensure he has a copy of *Phaedo* near to him. After some deliberating, he decides that due to the circumstances in his life, it is necessary to order one from Amazon.com and have it shipped with standard delivery. After looking at the clock and considering things for a few minutes, he decides he will submit the order within the hour.

Though David's promise might seem a bit unusual, he is surely not irrational with respect to it. He is responsive to the rational pressure he is under, and forms intentions accordingly.[9] To its credit, Time Enough has no difficulty accounting for the gap in time between when David decides an Amazon.com order is necessary to achieve his end and formation of the intention to submit the order. Since David believed there was time enough to act on the intention he believed to be necessary for the ends, and since he formed the requisite intention, he is (at least through the end of the example) immune to criticism regarding his practical rationality concerning the promise.

sufficient time to act on I(m), were he to I(m). Bratman only requires that the agent must view him or herself as having sufficient time to intend the necessary means. This difference is immaterial for present purposes.

[9] For future-directed plans the (practical) rational pressure to fill in the details depends on a wide range of things. In this case, what (practical) rational pressure David is under is partially a function of the rationality of his I(e), but also partly a function of certainty about his impending end, and his views about how distant it might be.

David could have false beliefs about whether or not he will soon die. Or, those beliefs might be unjustified. Or, they may have been arrived at through a deviant process. This range of possibilities may reflect defects in, among other things, David's theoretical rationality. They do not impugn his practical rationality, however. Even if he dies five minutes after ordering the book, this would not (by itself) show that he had been irrational— what reasons there are for an agent to do something may sometimes diverge from what reasons an agent *has*.[10] Similarly, even if he died five minutes before he was to order the book, this too would not besmirch his practical rationality, for he acted appropriately by his lights.[11]

Time Enough is an improvement over the Simple View, and less obviously problem-ridden than The Fix Is In. Its chief difficulty concerns how it handles a particular kind of case of present-directed intention.

Consider the following scenario:

> Malik is not a regular video game player, but he is over at Kiyoshi's house and invited to play a typically violent "first-person shooter" video game. Malik acquiesces, and Kiyoshi explains how the game works. Kiyoshi highlights the fact that a particularly fast and powerful creature in the game (the "Alien") can only be defeated by pressing a particular button when it appears. The problem is, the Alien is very fast, and just which button does the trick is not hinted at until some time late in the video game, a point well after Malik is likely to be engrossed. At t1, Malik begins to play the game, with the intention of defeating the Alien. At

[10] Though he ultimately argues against there being anything like external reasons or external rationality, the distinction is suggested in Bernard Williams' influential "Internal and External Reasons" reprinted in *Moral Luck* (Cambridge: Cambridge, 1981), p. 101. There are various other ways of carving up similar terrain, e.g., distinguishing between subjective reasons and objective reasons. Regardless of how feels about the possibility of external reasons, I take it this distinction marks a difference worth keeping track of, even if it does not track a distinction between kinds of reasons.

[11] This example suggests we may frequently misattribute practical irrationality. Suppose I announce my intention to be in shape by the time I retire. When I have not started exercising my grossly corpulent physique within at least six months of retirement, people will naturally begin to think that I am practically irrational with respect to this aim. However, I may simply (mistakenly) believe that it will only take two weeks to go from corpulence to septuagenarian health. The more an agent's action-relevant beliefs depart from the beliefs we expect agents to have, the more likely we are to (mis)attribute irrationality to the agent.

> t2, in the middle of appalling simulated carnage, Malik learns that the color of the Alien indicates what button is the one that triggers the Alien-terminating weapon. However, since he has not yet seen the alien, he does not yet know what button to push. At t3, the Alien appears. It takes a split second for Malik to recognize that Alien for what it is, but once he does (at t4), he puts things together and comes to believe that pressing the green button is the only way to defeat the Alien. So, Malik intends to defeat the Alien, and believes that pressing the green button is necessary to defeat the Alien. But, because practical reasoning is the sort of thing that takes time —even if only milliseconds — at t4 he does not yet have the intention to press the green button. However, only milliseconds later at t5, the Alien disappears. Malik realizes the Alien is gone for good, and that if he were to intend to press the green button now, it would be too late to defeat the boss.

In this example, was Malik ever practically irrational? I think the answer must be no and it would be a mark against any account if it said otherwise.

Start with t4. At t4, Malik recognizes the Alien and this engages his intention to defeat the Alien (I(e)). Let us suppose that nearly instantaneously, he determines that pushing the green button is necessary for defeating the Alien (B(e->m)). We can also suppose that by his lights, it is not too late to intend to push the green button I(m). On Time Enough Malik is not yet irrational because by his own lights there is still time enough for him to act on the intention to push the green button, were he to intend to do so. At t5, however, time runs out— even by Malik's own lights. If at t6 Malik were to go on and press the green button as a means of defeating the Alien (this sort of intention inertia is common among video gamers), then we might rightly call him irrational. At t5 —that is, before he goes on to press the green button— it does not yet seem appropriate to call him irrational. Yet, Malik satisfies Time Enough's conditions for irrationality: He is a Case 2 agent (I(e) & B(e->m) & ~I(m)) who also believes that were he to I(m), there would not be enough time for him to act on I(m).

One might argue that at t5, by the agent's lights it is no longer rational to I(e), and so the example violates one of constraints we accepted earlier (that I(e) is rational). Or, one might argue that when the Alien disappears, Malik no longer I(e), or even B(e->m). These replies are not sufficient, however.

Given the truth of RT, there is a certain degree of inertia that rationality must have in an agent.[12] For example, suppose that it takes practical reasoning even mere tens of milliseconds to do its work (e.g., realizing that pursuing I(e) is now impossible) after Malik recognizes the Alien is gone. During that time when reasoning is doing its work, the rationality of I(e) has not yet been overturned. (It may help to recall that the sense of practical rationality under discussion is circumscribed by what reasons the agent has, not by what reasons there are.) Thus, at t5 it is not yet the case that it is no longer practically rational for Malik to I(e).

Similar remarks hold for the issues of whether Malik still intends to defeat the Alien at t5, and whether he believes intending to push the green button (even *right now*) is necessary to defeat the Alien. Even if it only takes tens of milliseconds for Malik's practical reasoning to determine that it is no longer appropriate to I(e) or B(e->m), during that time when practical reason is at work, Malik still I(e) and B(e->m). We are fast, but not instantaneously responsive to changes in our environment.

Thus, the problematic moment for Time Enough is when the future catches up to the present, when (1) Malik's rational I(e) requires a present-directed intention, (2) he believes that e requires pushing the green button now, but (3) he has not yet formed the intention to e, and (4) were he to intend e now, he would not have enough time to do act on that intention.

[12] I take it that the considerations that follow are part of what motivates the afore-mentioned "permissive" understanding of rationality.

The source of the problem is that it would be inappropriate to describe Malik or any other agent as irrational on account of the agent lacking the opportunity for practical reason to do its work. This is true for acquiring the intention to m and for treating the considered end as irrational. (In what follows, I focus on the former case, though of course these points also cover the latter case.)

The obvious solution is to stipulate that an agent counts as rational (and is not irrational) if the failure to acquire I(m) is the consequence of the agent lacking the opportunity to do so. Call this emendation of Time Enough the *Quick Fix* solution. Like The Fix is In, it too proposes a stipulative solution to the problem. However, Quick Fix differs from The Fix is In in several important respects. First, it does not require that we reject the "permissive" account of rationality. In doing so, it does not appear to solve one counterintuitive result by positing a further counterintuitive result. Second (and relatedly), it avoids the *ad hoc* objection because the solution can be independently motivated. Finally, especially when coupled with Time Enough, there appear to be sufficient resources for explaining what will count as an opportunity to acquire I(m).

Regarding the first difference (compatibility with a permissive notion of rationality), Quick Fix need not hold that Malik is neither instrumentally rational nor instrumentally irrational. Instead, we can hold that Malik is not irrational (i.e., is rational) through t5. So, Quick Fix need not be saddled with The Fix is In's rejection of a permissive notion of rationality, and thus does not incur objections about counterintuitiveness on this account.

Neither does Quick Fix raise worries that its proposal is *ad hoc*. That is, there is a principled explanation for why agents under conditions akin to those of Malik at t5 are not to count as irrational. The principle turns on something like the thought that "ought implies can."

Unless the agent had suitable time to acquire I(m), we lack legitimate grounds for criticizing the agent's rationality. Surely having superhuman, or perhaps even physically impossible capacities cannot be a prerequisite to avoiding irrationality in these contexts. If these were prerequisites, what could be the grounds for such a demand? In the absence of any suitable explanation, it makes sense to hold that an agent cannot be irrational unless the agent has had an opportunity to acquire the relevant intention required by principles of instrumental reason.

What then of the crucial third difference, some account of what "opportunity to acquire I(m)" amounts to? Here, the account ought to be understood in terms of subjective features of the agent; suitable opportunity is relative to how long it takes for that agent to form the relevant intention. However, this answer may raise several concerns that parallel worries that the Fix is In initially raised.

First, one might worry that very slow deliberators are permitted unreasonably long periods of time in which they do not count as irrational. Suppose that Bertha is an *exceptionally* slow deliberator— it typically takes her about two minutes to make the obvious conclusion to I(m) when she I(e), and also B(e->m). Even in this extreme case, it does not seem that it would be appropriate to call Bertha irrational before she has had an opportunity, an opportunity relative to his cognitive powers, to acquire the requisite intention. We might remark to others or ourselves that Bertha is slow or dim-witted— but her rationality ought not be treated as a function of speed.

These remarks might raise a second worry, concerning what is embedded in the notion of opportunity, and whether it requires a certain metaphysically robust picture of human agency and its place in the natural order. These concerns are connected to issue in contentious debates about free will, and it would be impossible to draw any demonstrably justified conclusion in this paper.

So, these worries should be bracketed. Suffice to say that if practical rationality does require opportunities incompatible with determinism or a physical causal order, this would be an important discovery.

A different line of objection is this: tying a notion of suitable time to the particular facts about an agent's intention acquisition permits irrational sources of delay in acquisition of the intention to m. For instance, if Bertha is taking a long time to acquire I(m) because deliberation about the considered issue inspires fear, or if Malik takes a long time to I(m) because he is so engrossed in the video game, these seem to be factors that mitigate the rationality of the agent. On the face of it, "interfering" psychological states such as fear and distraction should be capable counting against the agent.

Delays in acquiring the intention to m (whether it is a future- or present-directed intention) on account of interference will count against the agent's practical rationality when those interferences are suitably infected by irrationality.[13] If Bertha's deliberation-freezing fear is suitably infected by irrationality, then it makes the delay brought on by its interference irrational. Nonetheless, not all interference is obviously infected by irrationality. If, for example, Malik's decision to play the game and his tendency to become engrossed in it is not infected by irrationality— perhaps it is in keeping with his not unreasonable aims for occasional leisure and stress-release— then interference of this sort does not infect the practical rationality of Malik.

One might dispute whether psychological states such as engrossment, or fearfulness can be practically irrational. I am inclined to think they can be, but the account works either way: if interfering states or their origins are themselves never subject to practical irrationality, then

[13] This is connected to the "no bootstrapping" qualification I mentioned at the start of the paper. What counts as "suitably infected" is hard to say. It is reasonable to think that at some point downstream intentions lose the infection of irrationality that the original ends might have had. When and how that happens is doubtlessly complicated.

agents whose intention acquisition is delayed by non-irrational interfering states cannot, at least for that reason count as practically irrational, either. This result would be in keeping with the somewhat revisionary nature of such views about the rationality of those mental states. If, by contrast, we accept that some interfering states or their origins can count as irrational, in those instances we can appropriately describe the agent as irrational.

In short, there is Time Enough to ensure a Quick Fix so that our practical reasoning itself does not so frequently entail momentary irrationality.[14]

[14] Thanks to Agnieszka Jaworska, an anonymous reviewer for *Philosophical Studies*, and especially Ken Stalzer and Yonatan Shemmer for comments and discussions on earlier versions of this paper. Thanks also to Michael Bratman and Alan Hájek for helpful conversations about these issues. Work on this paper was done with the financial support of the University of San Francisco Fleishhacker Family Endowment and the California Institute of Technology.