

Rational to be Irrational?

James R. Shaw

Suppose I know what I want and I know how likely any action I perform is to get me what I want. Then decision theory—the formalized theory of means-ends rationality—tells me what I ought to do: perform the action which, on average, is most likely to get me the most of what I want. In technical terms, I should choose the action that maximizes *expected value*.

Suppose things are a little more complicated: now I face not one, but a series of choices (or the possibility of multiple choices) where coordination opens up special opportunities for the advancement of my interests. “No problem,” the standard line goes, “simply think of each group of choices collectively available to you as a *plan*. Plans as a whole can be evaluated in terms of how much, at the end of the day, they will get you of what you want. So choose the plan you could execute which, on average, is most likely to get you most of what you want. That is, choose the plan that maximizes expected value.”

I am side-stepping some complications here in order to get at a simple idea: theorists who think in the way I have glossed endorse two conceptions of rationality—what I will call *choice rationality* and *plan rationality*. Moreover it is often assumed that when we face a series of choices these two notions *coincide*. If I choose a rational plan and execute it, then each choice I make in carrying out the plan is a rational choice, holding my choices at other stages of the plan fixed. This is doubtless generally the case.¹ Here, though, I would like to present a peculiar case where the two notions come apart. Moreover, in some such cases a rational plan involving irrational choices, I contend, is to be favored over a plan consisting of only rational choices. My claims rely on intuitions about what it is rational, in some pre-theoretic sense, for someone to do in certain scenarios. But this is always necessary if we are to challenge various techniques which purport to model pre-theoretic rationality.

So to recapitulate, there are three operative notions of rationality: plan

¹One might have reservations, but see p.4 for the standard means of dispensing with apparent counterexamples.

rationality and choice rationality, as characterized by some version of utility-maximizing decision theory, and the pre-theoretic notion of rationality as applied to an agent at a time. The standard claim is that an ideally rational agent in the pre-theoretic sense opts only for rational plans and rational choices. I will argue that this claim cannot be true by showing that rational plans may involve irrational choices. Moreover, in some cases rational plans are to be favored over rational choices.

The case I will use to motivate my claims builds on the following thought experiment: Sadistic scientists are going to put you to sleep on Sunday and flip a coin. If the coin lands heads, you will be woken up and given the choice between one of two options, *A* or *B*. If you pick *A* you'll get \$1000, and if you pick *B* you'll end up with \$0. After your choice you'll be sent on your way. If, on the other hand the coin lands tails, then you'll be woken up, broached of the fact that the coin landed tails and faced with a choice between 'opting in' or 'opting out'. If you opt out you get nothing and are sent on your way. If you opt in you get \$10 and are sent on your way.

Realizing the case is too simple, the scientists add a catch. They will make a clone of you while you were sleeping whose fate hangs on the outcome of the coin toss and the choices you make. If the coin lands heads, the clone is woken up, told she is the clone and sent on her way. Likewise if the coin lands tails and you opt out. However, if the coin lands tails and you opt in, then then clone is woken up in an indistinguishable scenario from the one you would be woken up in if the coin landed heads: she is given a choice between the options *A* and *B*. The payoffs for the clone are, however, inverted and magnified. On *A* the clone gets nothing and on choosing *B* gets $\frac{1}{\epsilon} \times \2000 —more on this ϵ shortly.

The case is complicated because during the experiment one might lose information about who one is. Suppose, for example, you form the following, apparently cogent plan: "if I wake to the opt-out/opt-in dilemma I'll take the \$10, and if I wake up to the *A/B* dilemma, I'll go for *A* and the \$1000." But you pause—suppose you wake from your blissful slumber your find yourself in the *A/B* dilemma and are about to choose *A* as planned, but a nagging question pops into your mind: might you actually be the clone? Since you were willing to opt-in when queried, there is a $\frac{1}{2}$ chance that a clone would be created, and if it were, then it would be in the situation you are now in, thinking the thoughts you are now thinking. Given that in such a scenario you are in a state subjectively indistinguishable from that your clone would be in, how much credence should you give to being her?

One answer that immediately suggests itself is: about $\frac{1}{2}$. But that might be hasty. As Elga (2004) points out, we must be careful in how we allocate credences to subjectively indistinguishable states *across worlds*, lest we rationally mandate wholesale skepticism. Since the issue is a tangled one, I won't do anything like propose a general principle here. I will, however, venture the following: in this case we should assign *some* credence, however small, to being our clone.

To help see why this might hold, consider the consequences of barring trans-world allotment of credences to subjectively indistinguishable states in the following alternative case: scientists are putting you to sleep on Sunday and will clone you. They'll wake your clone in a red room and you in a blue room. Supposing the scientists are known to be completely trustworthy, then I take it that upon waking to a red room your clone should be close to certain that she is, indeed, your clone—it would seem highly imprudent for her to believe otherwise. Now suppose the case is slightly different: there is a miniscule—say $\frac{1}{100}$ —chance that scientists will wake you in a red room rather than a blue one. Now, if we were to bar the requisite trans-world assignment of credences to subjectively indistinguishable states, then were you to wake in a red room you should be certain that you were yourself rather than your clone. But then presumably your clone should believe the same thing. But it is implausible that such a slight difference in the case mandates a shift in the clone's beliefs from near complete certainty in her own identity to complete certainty in being you.²

So it looks like, in my original case, if you wake to an A/B dilemma on the hypothetical plan, you should give *some* credence to being your clone—let us denote the relevant degree of credence with the aforementioned “ ϵ ”. Then if one woke to the A/B dilemma, assuming the linearity of both your and your clone's utility function in dollars, the expected values of your choices would be as follows.

$$\begin{aligned} \text{EU}(A) &= [(1 - \epsilon) \times 1000] + (\epsilon \times 0) &< 1000 \\ \text{EU}(B) &= [(1 - \epsilon) \times 0] + (\epsilon \times \frac{1}{\epsilon} \times 2000) &= 2000 \end{aligned}$$

Hence, it would be irrational to go ahead with the plan to choose A .

²My argument here might have poor traction with those who think that one should have less than complete confidence in any proposition. Of course, those theorists already agree with the conclusion I need.

The plan given by “*A* if faced with *A/B*; Opt-in if faced with Opt-in/Opt-out” involves an irrational choice. It is often supposed that a (pre-theoretically) rational agent cannot, in full awareness of the facts, adopt such a plan. More on this supposition soon.

For now let us compare all the plans and their expected values. They are given as follows.

<i>Plan</i>	<i>Expected Value</i>
Opt-in; <i>A</i>	505
Opt-out; <i>A</i>	500
Opt-in; <i>B</i>	5
Opt-out; <i>B</i>	0

Reconsidering our options we see that, as is perhaps not unanticipated, every plan except one involves irrational choices. Opting out is always a choice in which one is sure to get \$0 over \$10—so the second and fourth plan go by the board. This leaves only the “Opt-in; *B*” plan, yielding a meager \$5 on average, as the rational plan to adopt.

Or at least this would be so if our original thesis of the harmony between plan-rationality and choice rationality held. “Opt-in; *B*” is the only plan consisting of entirely rational choices. But it is not, I think, the rational plan. Nor is it, I will claim, the rational course of action.

As I alluded to before, it is often supposed that a pre-theoretically rational agent cannot intend to execute a plan with irrational choices, while believing that they will not suffer some sort of memory loss or be given new reasons in the interim. This is what is supposed to explain why cases like “Opt-in; *A*” do not present counterexamples to the claim that a rational plan involves only rational choices: that plan is, in a sense, not a live option for any pre-theoretically rational agent. In such cases the boon of rationality temporarily becomes something of a prohibitive nuisance—barring the foolish, though otherwise lucrative path.

Now the claim that certain plans are off the table because they involve irrational choices has its plausibility in that an irrational choice made during the plan is a reliable indication of pre-theoretic irrationality in the planning stages. Take “Opt-in; *A*” in this regard. Certainly, someone executing the plan

in taking A is seen as reckless:³ they are making a foolhardy choice without a justification available for it. How could the person foreseeing such a choice, and *willing* it, not themselves be subject to the very pre-theoretic irrationality their irrational choice manifestly embodies? How could the *local* irrationality of their choice not infect the *global* rationality of their plan? If the rhetorical force of these questions is on the mark, then pre-theoretic rationality simply precludes the possibility of choosing—willing, intending—the relevant plan.

I am sympathetic with this line of argumentation, particularly with its application to “Opt-in; A ”. However, I think “Opt-out; A ” might present us with a rare case where this method does not have obvious application. For the agent opting for this latter plan has a peculiar claim available to them as they make their irrational move which the former does not: “I choose to forgo a present good without compensation as part of an extremely desirable plan which indispensably involves this choice.” This may sound confused, but I think it is not. In particular I would claim the following: that it is not *easy* to impugn this irrational chooser as irrational pre-theoretically in the way other irrational choosers (and planners) might be.

To help see this, consider the clearer case of Miles. Miles is faced with an *iterated* version of the circumstance in which the scientists place you: every day for ten days they will repeat the experiment with Miles as the main subject, each day creating a new clone of her for participation in their experiment. Miles, like you, is a sophisticated reasoner. He sees that if he plans to make the rational choice when faced with Opt-in/Opt-out, he’ll be unable to opt for A should he face the A/B dilemma. So Miles reasons as follows: “I’m almost certain to face the A/B choice at least once in the next ten days. If I can just comfortably opt for A at least that one time, I’ll be much better off than if I execute the ‘rational’ choice “Opt-in; B ” every day. So I’m just going to do it: if given the choice to opt-out, I will do so every time without fail. Since I’m a very good judge of whether or not I’ll follow through with my plan, I can comfortably and rationally choose A whenever I am given the opportunity.” Miles does precisely what he plans—even to the last day. He makes out with \$4000—\$3900 more than if he had stuck with “Opt-in; B ”.

Imagining the case is meant to make two claims plausible. First: behaving as Miles did would be psychologically feasible for many, otherwise competent

³This might not feel as powerful in the case I have given, but one can draw out the sentiment here by considering a case where instead of being rewarded on B , the clone is tortured on choosing A .

reasoners. Second, it is not obvious we could condemn persons like Miles as being *irrational* in a pre-theoretic sense. After all, Miles arguably made out with as much of what he wanted in the scenario he was in as he could have. It is difficult to convincingly say to him: “Miles, you *should* have opted-in when you had the chance” as he would presumably reply, apparently truthfully, “but if I were to have done so, I would have been forced to miss out on these wonderful bundles of cash.”

Miles’ justification for a given irrational choice is that its involved choice irrationality is, in some sense, *offset* by the virtues of the plan of which it is a part. This justification, if any good, is also available to him while in the experiment. “Why not just opt-in this once, while you have the chance to make an extra \$10?”—“Because I am committed to a fruitful plan which requires me to behave just like this.” It seems that we can condemn his action as exhibiting his irrationality (in the pre-theoretic sense) during the run of the experiment if we can look back with him after the experiment and proclaim likewise. But this is not easy to do: how can one argue with such resounding, and predictable success?

Recall the idea raised earlier to maintain the thesis that the fully rational agent chooses the plan which maximizes expected value, and that plan has choices each of which maximizes expected value (other choices fixed) in turn. That idea was secured by supposing that plans involving irrational choices in normal cases simply couldn’t be intended by the ideally pre-theoretically rational agent. What we have seen is that the application of this idea to the plan Miles forms is far from clear.

If we cannot condemn Miles as irrational in the pre-theoretic sense, and if my claim about the psychological availability of the plan “Opt-out; A” is right, then that plan is the *most rational plan* among those available to Miles: it is the plan available to pre-theoretically rational agents which, if performed, stands to get them the most of what they want.⁴ At certain junctures in executing the plan, however, Miles will be making straightforwardly *irrational choices*: choices which, taken in isolation and holding other choices fixed, will not get Miles the most of what he wants on average. This is the sense in which plan rationality and choice rationality may come apart. Moreover, in this case it

⁴What about the plan of always opting-in and choosing A? Recall that though this has a higher expected value than Miles’ plan, that it appears much clearer in such cases that we can pronounce those who perform the relevant irrational choices, the plan as a whole, and those who intend it, as pre-theoretically irrational.

seems that plan rationality wins out over choice rationality as a candidate for a systematized clarification of our pre-theoretic notion of means-ends rationality. After all, anyone who could, like Miles, psychologically manage to plan and execute the plan rational option would doubtless take it over choice rational options in Miles' circumstance.

Note, of course, that this is not simply a case where it is rational to choose to make oneself irrational—as the famous case where one has the option to take a pill making oneself temporarily irrational so that one will not succumb to torture. In those cases the 'irrational choices' in one's plan are not choices one *intends to perform* as part of the plan. Rather they are choices one—in a state of induced diminished responsibility—will be prone to make. The case I have presented seems to be in one in which an agent is rationally required, in full possession of her faculties of reason and choice, to opt for an individual irrational decision.

Before I conclude I'd like to point to two applications of the claims argued for here. The first is to the possibility of certain kinds of credible threats. Suppose an agent Lucy suspects that Charlie will slight her. Lucy says: "if you slight me, I'll make a small personal sacrifice to see you really suffer." Suppose Lucy does not stand to gain by setting a precedent here, nor by seeing her enemies suffer. Then if Lucy is rational, the act she claims she will perform is an irrational choice: she would forgo a good for no ostensible compensation. Cases like these are ones in which, assuming the agent's rationality, a threat does not seem *credible*. Lucy's language is just 'cheap talk.'

But suppose circumstances are slightly different. Lucy is a bad liar with a dead giveaway 'tell' everyone knows: when she lies, her right eye twitches. Suppose further that Lucy is otherwise pre-theoretically rational. Lucy now issues her threat—no twitch in sight. Charlie is astonished, and perplexed, but one thing is sure: for him that is the end of the matter. Slighting Lucy is out of the question. If the lack of the twitch is what it looks to be, Lucy has indeed planned to ruin Charlie at her expense. We can suppose this to be as good as evidence as there could be that were Charlie to have slighted her, Lucy would indeed have gone ahead and taken a loss to carry out her plan. When asked about her intentions Lucy said "it would have been unfortunate for me to bring Charlie down if he had slighted me, but it would have to have been done." Lucy opted for a plan which which by the lights of most theorists shows her to exhibit (perhaps mildly) irrational behavior. I think reflection on cases like Miles' can increase the sense that this is not as obvious as it is sometimes made out to be.

A second application of my claims might place the much discussed ‘toxin puzzle’ of Kavka (1983) in a different light. This is the case where an agent is to be substantially rewarded on Monday for forming the intention on Sunday to drink a painful emetic on Tuesday. In discussions of the case it is sometimes stipulated, but sometimes presumed, that we cannot form the relevant intention or would find it difficult to do so, as we know or suspect we will not drink the poison. Perhaps many people could not drink the poison or would find it difficult. But some might be able to. Indeed, the more one reflects on Miles’ plight, the more it seems possible. And indeed, if it is possible, then we again have a divergence in plan rationality and choice rationality to ponder. Moreover, if the case of Miles is any indication, at the end of the day it might make most sense to just drink up.

References

- A. Elga (2004). ‘Defeating Dr. Evil with Self-Locating Belief’. *Philosophy and Phenomenological Research* **69**(2):383–396.
- G. Kavka (1983). ‘The toxin puzzle’. *Analysis* **43**(1):33.