Just Imagine

Frederick Kroon[1]

Introduction

Some time ago I did a bungy jump. Nothing remarkable in that (nor in the fact that I have, or had, a great fear of heights; the desire to overcome a fear of heights is not an uncommon reason people give for bungy jumping). What was a little bit dismaying was the lack of credit I received for what I had done. Friends and colleagues to whom I mentioned the act, including friends in ARF, thought it slightly bizarre. Certainly they didn't think that there was anything particularly praiseworthy about the act. I disagree. I thought the act praiseworthy but not in the way one might think. I didn't, for example, think it displayed great courage, since I didn't think I was in any real danger when taking the plunge (I knew the statistics, and rejected as urban myth, perhaps wrongly, the story about retinas detaching—indeed, I thought the way people harped on this showed a special kind of weakness). In general, one shouldn't praise people for their courage when the reason they have for thinking they were in danger is foolish or based on inexcusable ignorance. That wasn't my situation.

This paper has its roots in my interest in setting the record straight—my desire to say what was praiseworthy about the act, certainly in terms of rationality and even in terms of morality. But because that wouldn't be an academically proper subject for a paper, I am going to start again, this time with some paradoxes of rationality that were discussed in the 1980s and 1990s but that deserve a continued hearing. The connection to bungy jumping will emerge later in the paper.

The problem with some good intentions

Nuclear deterrence, so Gregory Kavka argued some time ago, involves us in a kind of paradox.[2] Suppose that the best way, taking account of the full range of possible consequences, for a certain state A to deter a genuine and extreme nuclear threat against A made by some enemy B is for A to announce a sincere intention to punish B with B's own nuclear demise—in particular, the nuclear destruction of much of B's population—should B act on this threat.[3] Assuming that the survival of A is a rationally and morally worthy goal, it seems clear that A should form this retaliatory intention, both from a rational and moral point of view. It seems equally clear, however, that A can't be a truly rational or moral

[1] Thanks in particular to Richard L. Epstein for numerous useful comments on an earlier draft.

[2] See Kavka, 1978, 1987.

[3] For convenience, I shall write "A" and "B" rather than "the appropriate authority in A/B".

© 2004 Frederick Kroon.

Bulletin of Advanced Reasoning and Knowledge 1 (2004) 61–77.

agent if A is able to form this intention. After all, the retaliatory intention in question demands that A be prepared to act in a way that A himself must regard as plainly irrational and immoral: It demands of A that he use his nuclear weapons to destroy much of B's population should B attack A, and A can see that such a retaliatory act would only achieve more useless death and destruction. (Remember that in the scenario that A is being asked to contemplate, deterrence has failed; A's population has been virtually destroyed, and the subsequent destruction of B's population would simply compound the tragedy.) In short, while forming such a deterrent retaliatory intention is the rational and moral thing to do (assuming that the agent faces such a threat to his survival), it seems that, paradoxically, any would-be intender must be irrational and immoral because of the awfulness and pointlessness of the harm he agrees to inflict should this intention fail to deter.

Kavka thought that the right solution to this so-called "paradox of deterrence" was that the rationality and morality of actions and of agents sometimes come apart: Forming the deterrent intention is the moral and rational *option* for the agent facing such a nuclear threat to his survival, but no truly rational, moral *agent* can adopt what is the rational and moral option in this case. Call this the *agent-irrationalist/agent-immoralist* solution to the paradox.

Although the case of nuclear deterrence, on its classical "mutually assured destruction" construal, has been the most widely discussed instance of this paradox, Kavka thought that the puzzle also extended to certain less apocalyptic scenarios involving threats likely to deter unwanted behaviour. Thus, consider a person's threat to leave her partner (whether a sexual partner or a business partner) if the latter continues his cheating ways. It may be clear that it wouldn't be in the agent's best interests to leave her partner even if he chooses to continue cheating—considerations of emotional, financial, and/or physical security may make this plain. It may also be clear, however, that issuing a credible threat to leave him should he continue to act this way would have an excellent chance of affecting his behaviour—if only the agent could manage to issue such a threat (assume that bluffing is out of the question). The problem, as before, is how a rational agent can form the sort of deterrent intention required, given that she believes that actually acting on the intention should her partner not change his behavior would only make things worse for her. Forming the intention would be the rational option, given its probable success in preventing the partner's cheating, but a rational agent seems once again to be prevented from forming the intention.

Deterrent intentions are conditional in nature. Kavka later decided that his puzzle also applied to certain cases involving unconditional intentions, thus showing (as he thought) that one cannot intend, conditionally or unconditionally, whatever one wants to intend.[4] This was the substance of his now famous toxin puzzle.

> You have just been approached by an eccentric billionaire. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day,

[4] Kavka, 1983.

> but will not threaten your life or have any lasting effects. The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you *intend* to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed.[5]

The only other conditions on the offer are that you are to make no bets, do anything that will cause you to become irrational, or arrange for any way to avoid the effects of the toxin. This is all part of the offer, and you are able to confirm with your daughter, a lawyer, and your spouse, a biochemist, that the offer is legally valid and based on accurate information about the toxin. You also know that the billionaire has a piece of equipment able to detect with near-certainty whether or not you have formed the intention.

The puzzle is this. Suppose you decide that being ill for one day is a small price to pay for a million dollars. Your first thought is therefore to agree to the proposal. It then occurs to you that you won't even have to become ill in order to win the money since you won't have to drink the toxin—all you have to do is to intend to drink the toxin, not to actually drink it. But how can you intend to drink the toxin if you also know that at midnight you won't have any reason to drink the toxin? At that point, you would already have been paid, and drinking the toxin would only make you unnecessarily ill. So the best you can do is to *pretend* to form the intention—you can't actually form the intention, since as a rational agent you can't see yourself acting on the intention. Unfortunately, the billionaire needs ultra-reliable evidence that you have actually formed the intention, and so your best is not good enough.

In this case too, Kavka thought, the agent can't rationally decide to form the intention because of what its execution is known to involve—an action that in the event would be against one's clearest interests—even though the action of forming the intention (as opposed to *acting* on it) would clearly carry enormous benefits for the agent. Kavka thought that the tension in this description should again be resolved in terms of a distinction between agent-rationality and option-rationality: no fully rational agent can intend to drink the toxin, yet intending to drink the toxin is the rational option for the agent. I'll again call this an *agent-irrationalist* view of the situation. (On the existing formulation, there is no corresponding paradox for morality, but we can readily manufacture one. Suppose that the billionaire makes it clear that he will donate millions of dollars to help with the eradication of a certain serious disease, but that the other conditions remain the same. The new version of the problem suggests that no ideally rational and moral agent is able to form the required toxin-drinking intention despite the fact that forming the intention is in the circumstances the morally right thing to do.)

Whatever we think of agent-irrationalism as an account of the paradox of nuclear deterrence, I don't accept its extension to the case of less apocalyptic

---

[5] Kavka, 1983, pp. 33–34.

deterrence scenarios, such as the threat to leave a cheating partner.[6] Nor do I accept its application to the toxin-drinking puzzle. In this paper I want to show how to be an agent-rationalist about such cases. The paper is structured as follows. In the next section I consider the sort of model of rational intention-formation that makes agent-irrationalism a tempting response to these puzzles, and then I identify what I take to be a crucial flaw in such a model. In its place I propose an alternative account that uses the notion of the imaginative preconstruction of an intention, and I argue that such an account is able to accommodate the rational formation of deterrent intentions by showing how the conditionally intended behaviour can count as irrational *apart* from the preconstruction, and rational on the *basis* of the preconstruction.

Toxin-drinking intentions seem to be a different matter, however; drinking toxin seems to be a pointless activity on any account of the formation of the intention. That is where the example of bungy jumping proves useful. After sketching an account of how the resolution to perform such a jump might be formed, I argue that the way in which the bungy jumping resolution is arrived at mirrors the way deterrent intentions are arrived at, and I then argue that the problematic toxin drinking case falls under the same account. In all these cases, I argue that there is nothing in the idea of rationality as such that prevents a rational agent from forming the intention.

Agent-irrationalism and deterrent threats

Let's return to the case of non-apocalytic deterrent intentions such as the woman's threat to leave her partner. How good is the argument for thinking that rational agents cannot form and sustain such deterrent intentions, despite the rationality of forming and sustaining the intentions? That is hard to say without knowing more about the notions of a rational option and a rational agent. The following account is supposed to be uncontentious and minimal. To describe an *agent* as rational is to characterize the agent as epistemically responsible: Such an agent responds to evidence in the right sort of way, believing propositions when the evidence supports them (but at any rate not when the evidence supports other incompatible propositions) and deciding how to act by taking proper account of his or her desires and beliefs regarding the likely outcome of actions. This is clearly a dispositional notion, for someone is correctly described as rational to the extent that he or she is disposed to function in this way, not just that perchance they always do function in this way. But note that the disposition is characterized in

[6] In Kroon, 1996 I defend the idea that being fully rational and moral is no bar to being able to form such deterrent intentions—fully rational and moral agents can intend to do things that, from their current perspective, they recognise as being deeply irrational and immoral. Perhaps the main problem with such a view is that this sounds more like a (conditional) intention to become something one is not, namely an irrational and immoral agent, rather than a (conditional) intention to do something that fits one's overall goals and commitments. Although I argue against this construal (on the grounds that such an agent possesses a form of deliberative integrity missing from other examples of agents' deliberately becoming irrational), I agree that the construal is a tempting one.

terms of a more local rationality: Options open to a person have the property of being rational if they involve the impact of the person's evidence in the right sort of way (when the option is believing a certain proposition), or if they reflect his or her beliefs and desires in the right sort of way (when the option is performing a certain action). The rational agent is one who is disposed to let his or her choices of options be determined by whether or not they have this property.

The proper characterization of the latter property is, of course, a contentious matter, with different theories defining the property in different ways. Thus, among theories of rational choice we have theories that recommend maximization, whether of evidential expected utility, causal expected utility, or some other agent-value, as well as theories that promote satisficing or some more extreme kind of suboptimizing.[7] In addition, there are theories that explicitly allow only for instrumental rationality, others that allow for more, in particular, by allowing for a rational evaluation of agents' goals. For present purposes, however, there is no need to choose. What is important for my purposes (and Kavka's) is that all these views agree that rationality is first and foremost a property of the options available to an agent, a property that applies to an action in virtue of certain independently specifiable features it has or constraints it satisfies. Within the context of such an approach, we can understand Kavka as claiming that there are imaginable situations involving the adoption of certain conditional and unconditional intentions where the best theories of rationality declare that the adoption of such an intention is the rational option for appropriately placed agents (say, because of the likely beneficial consequences of adopting and announcing this intention), even though it is not an option available to agents who count as ideally rational on such theories.

I am going to assume the distinction between agent- and option-rationality for the remainder of this paper. Now let us return to the question of why, in the imagined scenario, it is thought that a rational agent cannot form and sustain the conditional intention to leave a partner should he continue his cheating. In schematic form, the problem is this. Let P be a rational agent who strongly desires that some other agent Q not do C, and who recognizes that, in all likelihood, the (only) way to prevent Q from doing C is to form and announce the conditional intention that if C happens she will apply sanction E. Indeed, suppose that forming and announcing the intention, rather than not forming the intention, is clearly the rational option for P, given her evidence and values. But suppose that P also knows that, all else being equal, applying E if C happens would not be in her interests: even under C, not-E is better than E. Knowing this, it seems that P can't reason her way to the conclusion "I intend to do E if C happens," even though part of her wishes she was the kind of agent who could bring herself to accept this conclusion.[8]

[7] Recent theories of rational choice include Robert Nozick's "maximization of decision-value" account (Chapter 2 of Nozick, 1993), where the decision-value of an act is the summed value of various kinds of expected utilities of the act, each value weighted by the agent's confidence in being guided by that utility. In stark contrast to all such maximizing or "optimizing" theories, Michael Slote, 1989, presents a radical suboptimizing theory of rational choice.

In short, we seem to have the following inconsistent triad:

T (*T1*) P is (fully) rational, and hence chooses to perform any action if that action is the rational option for the agent.

(*T2*) It is clear to the agent that forming the conditional intention to do E if C should happen (that is, an intention with the content: "If C happens, do E") is the rational option for the agent

(*T3*) It is clear to P that, intention aside, if C should happen it would be against the balance of reasons for her to do E.

The agent-irrationalist sees this triad as inconsistent, and thinks the inconsistency should be resolved by insisting that what it is rational to do—in this case, forming the conditional intention—is not always something that a fully rational agent is able to do.

There are two main construals of the argument that (T) is inconsistent, and both rest on the idea that to form the conditional intention to do E if C requires the rational agent to have a rational preference for doing E to not-doing E on the supposition that C has happened. One thought is that (T3) entails that our rational agent P will not in fact be able to form the intention (contra (T1) and (T2)), because forming the intention requires P to assent to the conditional judgement "Supposing C has happened, the rational option for P is E," and according to (T3) P rejects this judgement.

I A rational agent's conditionally intending to undertake some action, say X, should some event D happen must depend on her recognizing that X is the rational option should D happen. Hence, contra (P), a rational agent cannot conditionally intend to do E should C happen, since she sees that doing E should C happen is not the rational option.[9]

(I) posits the tension in (T) as a simple consequence of what is involved in forming and justifying a (conditional) choice. But it thereby relies on a controversial assumption. It assumes that a rational agent can form conditional intentions only by using the following kind of matching deliberative process:

> Form the intention to do X should D happen (if and) only if doing X would be rational in the event of D's happening.

---

[8] *Part* of her. Of course, the agent might well realize that if she was the kind of agent able to form the intention then she would either not be fully rational or she would have values that she in fact rejects.

[9] At times, Kavka comes close to endorsing something like (I). Thus:

> It is part of the concept of rationally intending to do something, that the disposition to do the intended act be caused (or justified) in an appropriate way by the agent's view of reasons for doing the act. (Kavka, 1987, p. 292)

The words "the agent's view of reasons for doing the act" (my italics) sounds uncomfortably close to the kind of reflective account I reject in the text. It is likely, however, that Kavka's meaning is rather different, and that he means to endorse something closer to account (II) below.

But why grant this assumption? The only reason I can think of rests on a certain model of how decision theory is to be applied in ordinary non-conditional cases. On this reading, (I)'s claim that the conditional attractiveness of doing X is to be analyzed in terms of the agent's reflective assessment of X as conditionally rational is just a natural extension of the claim that the unconditional attractiveness of doing X is to be analyzed in terms of the agent's reflective assessment of X as unconditionally rational.

But if that is what lies behind (I), we have every reason to be suspicious. For in its unconditional form this gives the wrong picture of rational choice. In general it is not, and it certainly need not be, the case that rational agents choose by determining reflectively that their chosen option fits the demands of some canonical decision theory, where among other things this involves explicitly identifying one's desires as desires: items whose satisfaction counts in a way determined by the theory. All that rational decision theory demands is that the choices an agent makes systematically match the conclusions of whatever account of rationality is chosen as canonical. Rational decision theory need not in addition function as a kind of decision procedure.

(I) aims to establish agent-irrationality by claiming that, in choosing, a rational agent reflectively focuses on the question of the rationality of options: The agent "foregrounds" the fact that a (conditional) option open to him or her is or is not rational in (conditionally) choosing among the options he or she faces.[10] (Call this the "foregrounding model".) A natural alternative is one that "backgrounds" any appeal to the rationality of options and simply lets the question of what such an agent would decide in the circumstances be determined by how such an agent would evaluate the options he or she faces in the light of their beliefs and commitments, where the question of whether to form the conditional intention to do X should D happen reduces to whether the agent, on assuming D has happened, would choose X. It then becomes tempting to adopt something like the following view of conditional intentions. What makes it the case that a rational agent forms the conditional intention to do something X should D happen is that when such an agent considers a scenario in which D does happen, with a view to determining what to do in that (imagined) situation, his or her presently held beliefs and commitments incline him or her to choose option X. On this alternative "backgrounding" picture of the way the conditional intention is formed, the charge of inconsistency facing (P) can be put as follows.

II A rational agent can only intend to do something X should some event D happen if in conditionally choosing what to do on the assumption that D does happen he or she chooses X on the basis of presently held beliefs and commitments. It follows that he or she can't intend to do E should C happen, since to choose E conditionally on the assumption that C occurs would be to choose against the balance of reasons that as a rational agent he or she identifies with.

---

[10] See Philip Pettit and Michael Smith, 1990.

The idea is simple. Underlying (II) is the thought that in conditionally choosing, a rational agent goes through some such reasoning as this: "Suppose C has happened. Then it will be of no use to do E (say, to leave my cheating partner), since I thereby make a bad situation considerably worse. Hence I won't do E." Here the agent shows in her reasoning that she identifies with certain kinds of reasons that as a rational agent she also identifies with in her non-conditional choice; after all, a rational agent confronting C surely must reason in the following sort of way: "Unfortunately C has happened. Doing E will just make a bad situation considerably worse. Hence I won't do E." Unlike on the model presupposed by (I), the agent who conforms to (II) doesn't make her decision by identifying what would be the rational option for her to take should C happen, and then choosing on that basis what to do should C happen. Instead, she makes a (conditional) choice on the basis of good reasons (given her beliefs and commitments), and thus chooses in a way that plays out her rationality. It is this rationally made conditional choice, (II) claims, that shows why our rational agent can't form the conditional intention "I intend to do E if C."

Deterrent threats and their imaginative preconstruction

There is rather more to be said for (II) than (I), in my view. For one thing, it is based on a more plausible, because far less demanding, account of the way rational agents make decisions about what to believe and how to act. In particular, it doesn't demand that a rational agent have the reflective capacity to identify that an option is rational for the agent, but demands only that rational agents be appropriately responsive to good reasons (where what counts as a good reason will depend on the theory of rationality on offer). Despite this, however, I believe we should reject both accounts. While (II) demands far less of rational agents, it still demands too much. Importantly, its model of making (conditional) choices leaves out the impact of the conditional intention itself.

On the surface, that seems a strange complaint: The intention is the outcome, surely, of a bit of conditional reasoning; it can't be another bit of input into the conditional reasoning. But this misunderstands the complaint. For consider again the backgrounding model of the way we form conditional intentions. The agent supposes or imagines that the condition applies, and decides —in the scope of her imagining—how to respond. It is this process that is described in too impoverished a way by (II). For in imagining only that the condition applies, the agent forgets that if the formation of the intention was indeed successful then in imagining that the condition applies the agent should be imagining that the condition applies *in conjunction with the agent having issued a credible threat to do X should the condition apply*. To form the conditional intention in a way that doesn't beg the question against the possibility of forming such an intention, the agent has to consider the full imaginative context, and that imaginative context should allow for her having formed the intention.

Now of course this might seem an impossible task, for how, when the agent

is trying to decide what to intend conditionally, can she then use the thought that she has successfully formed the relevant intention as part of the reasoning towards forming the intention? Indeed, why not use the thought that she has not managed to form the decision, which would leave us with a stalemate? There seems something self-referentially incoherent about making any allowance for the conditional intention itself.

But I think it is none too hard to see how to form such an intention under the backgrounding model. The agent simply imagines the thick context with a view to seeing whether she can live with the conditional intention, useful as it is. Once she sees that she can live with it, she forms her conditional decision to apply the sanction. Perhaps she doesn't do this in one go. When she first tries, she might balk at applying the sanction: "No, I couldn't leave him; I would lose too much." But perhaps as she re-imagines the situation it becomes easier: "Wait. I am forgetting that he continued his cheating after all I did to show him how much I cared about his not doing it. I even threatened to leave him if he did continue his cheating, a threat whose consequences to me, should I act on it, he knew to be disastrous." After repeated contemplation of the imagined scenario, including repeated contemplation of the awfulness of her partner cheating after all she has done by way of her threat to warn him off this behaviour, it may become all too easy for the agent to fix on the conditional intention as one that she not only can live with but wants to live with. I'll call the process by which this is done the *imaginative preconstruction of the intention*.

In short, the agent might be able to bootstrap her way into forming the intention by way of such an imaginative preconstruction. That, I am proposing, is how it is done. But this can't, of course, be the whole story, for so far it is still not clear how a truly rational agent can, even within a sufficiently enriched imaginative context, decide to apply the sanction. For doesn't it remain the case that she sees that applying the sanction, namely her leaving her partner, is irrational because it is against her best interests? How does enriching the context help? But this remark once again works with the wrong model of conditional-intention formation. Its talk of identifying what is the rational thing to do in the context suggests something akin to the rejected foregrounding model on which (I) was based. The backgrounding model, supplemented in the way suggested, makes it easier to see how the agent might be able to form the conditional intention. I want to suggest that the agent's imaginative contemplation of her partner's continuing his cheating despite her intention being in force is likely to engage the agent emotionally: She will feel anger and resentment in a way that makes all the difference to her rationally deciding what to do in the scope of her imagining, and hence all the difference to whether she can bring off her imaginative preconstruction of the intention.

There is a contrary perception which sees the rational agent as inevitably calm and aloof, subject to the coldly calculative exercise of reason, and the angry agent as inevitably irrational because subject to quite another, irruptive, sort of motivation; but this contrary perception comes from a tradition that is now

generally regarded as wildly implausible.[11] Indeed, it is difficult to conceive of rational agents who lack an emotional life. Consider any decision theory on which the rationality of an agent's choice is a function of the satisfaction of her desires in light of her beliefs, whatever counts as an appropriate level of satisfaction and whatever else is involved. Now, desires impact on our emotions in at least two ways. First, many of our desires can only be characterized in emotion-attributing terms, and so too, therefore, must the rational status of actions based on such desires. Thus, we may act out of love for a person, behaving rationally to the extent that our action satisfies our desire for our loved one's well-being in light of our beliefs. Secondly, if a rational agent deems a certain choice of action the appropriate one to undertake, given her most fundamental desires, then she is not likely to take a neutral stance towards a contrary action on the part of another agent that debases these desires. Not only are emotions like resentment and anger not irrational in isolation; they may even, in a sense, be required emotions for rational agents if rational agents are to identify in the right sort of way with their desires. There is a substantial body of empirical evidence that confirms such a role for the emotions.[12]

It is this latter kind of emotional engagement that initially seems most relevant to the imaginative preconstruction of conditional deterrent intentions. But merely noting the case for emotional engagement of this kind doesn't greatly help the case for a preconstruction of the intentions in question, for the anger and resentment might be "required" emotions in a fairly thin sense: It might just be unnatural not to have them, but still leave the agent unable to seriously think about leaving her partner in the context of her imaginative engagement with the scenario of her partner's cheating. For as a rational agent, she must surely continue to see leaving as against her interests, no matter how angry she feels. She can't allow the anger to make a difference to how she evaluates the possible options of leaving and staying.

But this misunderstands the role that emotions like anger and resentment can play in such cases. If, in the agent's imagined scenario, they motivate her to leave, this is not likely to be explicable in terms of the agent's action merely being an emotional reaction to her partner's cheating. That would still leave the agent susceptible to the charge of irrationality ("What you did was to lash out in anger. You only hurt yourself that way, and are left looking foolish."). In my view, emotions like anger play a far more nuanced and complex role in this kind of situation. Assuming they succeed in motivating the agent to leave, they bring this off because they embody a shift in the agent's evaluative perspective.[13] Prior to,

[11] One radical criticism of the tradition came from Robert Solomon, 1976, who argued that emotions were just judgements. (As it stands, this view is clearly implausible, for one can make appropriate judgements—say, that one is in danger—without experiencing the corresponding emotion. Indeed, this seems to be the experience of those who suffer from certain brain traumas, and, lacking the appropriate emotional affect, find themselves incapable of being appropriately motivated by such judgements. See Antonio Damasio, 1994.)

[12] For a useful and influential account, see Damasio, 1994.

[13] I am here indebted to correspondence with Patricia Greenspan, and to Greenspan, 2000,

and apart from, her issuing the threat, the agent's interests were focussed on her well-being, something that she saw as likely to be compromised by her leaving. Still, she realised that there was a good chance of gaining a better level of well-being (better emotional security, say) if she were to issue her threat. Having made the threat, however, and having seen its failure, she now has to face the humiliation, if she were to stay, of backing down, and the indignity of remaining in a relationship where the hurt of the cheating has been compounded by the humiliation her partner has thus proved willing to inflict on her (remember that her partner hopes and expects that she will stay). Her anger is a complex reaction that shows that she implicitly understands all this, and thus shows that the game has now changed. There is a new end worth fighting for—her dignity—and this new end, which is as complex and emotion-involving an end as her love for her partner, is one that our agent gives expression to if she leaves. Her behaviour, should she leave, is rational in what some call an expressive sense, not in the sense that it is instrumentally useful to something else she values, such as greater security. Or rather, such behaviour should be seen as not inevitably irrational. In the scenario we are envisaging it is an option that doesn't contravene the agent's status as a rational agent, but that doesn't mean that the agent's leaving is required by her rationality. There is no way of arguing for this stronger claim, since an agent who faces adjudication between such competing ends is also faced with the fact that she may have competing standards of value, and competing ways of resolving any unclarity in what is to count as important to her or what is a tolerable disvalue. Despite this insult to her dignity, she may find herself unwilling to leave, without this impugning her rationality.[14]

That, I suggest, is how one should argue for the claim that the agent's leaving in this scenario fully accords with her status as rational agent.[15] Or rather this *imagined* scenario, for remember that we are here talking of the agent's preconstruction of her conditional intention or threat. We are not talking of any actual scenario, since the question of there being an actual scenario

---

although my emphasis is rather different from hers. I have been concerned with the way the intention might be formed, whereas Greenspan is more concerned with how the agent might seriously act on her threat should her partner continue his cheating. (To be fair, Greenspan clearly thinks that the possibility of such a scenario holds important lessons for understanding how the agent can seriously utter her deterrent threat, but she is not very explicit about the process involved.)

[14] Note that by being unwilling to leave she may well be foreclosing on something that is very desirable to her. She forecloses on doing something to ensure that her partner does not cheat, which in the scenario in question involves her uttering a sincere threat to leave him should he continue his cheating. In her case, unfortunately, uttering such a threat proves impossible. Note that this doesn't impugn her rationality since we can scarcely blame an agent for not doing something that is not a genuine option for her, given her beliefs and desires.

[15] For a very different account of such threats and their rationality, see Robert Frank, 1988. Frank thinks that the ability to utter credible deterrent threats of this type involves the emotions in a way that assigns emotions a strategic role. He thinks that this ensures that such threats are entirely rational. But he also thinks that the agent's behavior should the threat fail—the agent's leaving in response to her partner's continued cheating, in the present case—would be irrational. In my view, this is true only on certain very narrow accounts of self-interested rationality.

presupposes that the agent has indeed formed, and announced, her intention—and the problem that faced us was understanding how the intention could be formed in the first place. I have argued that it can be formed, and rationally so, on the basis of an imaginative preconstruction of the intention.

Good-but-hard non-conditional intentions: the lesson of bungy jumping

It is time to revisit what must surely strike us as a much harder problem: the problem of clearly desirable *non*-conditional intentions that involve the agent doing something blatantly irrational. Consider in particular Kavka's toxin-drinking example. The suggested solution to the corresponding problem for conditional intentions yields the following insight into why this problem is much harder. In the case of non-apocalyptic deterrent intentions we saw that a crucial role was played by the agent's passional reaction, in the imaginative preconstruction of the intention, to her partner's cheating: The cheating triggered her leaving. There is no such trigger in the case of non-conditional intentions, precisely because they are non-conditional: If the agent has the intention to do E, then there is no specific triggering condition C such that she really only intends to do E if C occurs (although it is, of course, true that the agent only intends to do E if various background conditions continue to hold). Even more so, it would seem, there is no circumstance C which, were it to happen, would somehow make the actual toxin-drinking a rational option for the agent (at any rate, an option that is not just irrational) rather than something that is pointless and painful. So the model used to explain the rational formation of non-apocalyptic deterrent intentions must fail us in the toxin-drinking case.

I disagree. Although it is true that there are striking differences between the cases and that it is genuinely harder to form toxin-drinking intentions, I think that it is nonetheless possible for rational agents to form such intentions via something like their imaginative preconstruction. But just how this is supposed to work is certainly not straightforward, since it seems obvious that no clear-thinking agent could ever bring himself to intend to do something that, in the event, would be both pointless and painful.

This is where the example of bungy jumping proves useful. Suppose that you want to perform a bungy jump, say as part of a process to help you overcome your inordinate fear of heights. (Assume that there is nothing else in it for you: no praise from friends, say, should you succeed.) Performing a jump first requires you to intend to perform the jump. But if you know that you have a great fear of heights, then resolving to jump is no simple matter. It is no use simply telling yourself that you are going to jump, since you know full well that things don't work this way (you would in all likelihood just freeze once you got to the end of the platform and looked down; and safety rules prevent you from take a running leap that requires no looking). There is a palpable fear that seems to limit agents in such cases, and which poses as much of a barrier to the act as the anticipated nausea of toxin-drinking is a barrier to toxin-drinking. So what can you do to form the required intention to jump?

In fact, to make this case as much as possible like the toxin case, suppose you know that what is important is that you form the resolution to jump in the face of your fear of heights (that is, by actually confronting this fear, for we are assuming that you are sufficiently rational to know that only by realistically confronting your fear can you form a serious intention). Suppose that, beyond this, the actual act of jumping serves no further purpose.

Now I take it that we don't think there is much of a puzzle here, because there clearly are agents who are able to form such an intention despite their fear. So how do they do it? The answer—at least one answer, but one that conceivably applies to many agents—makes appeal to a process that can be effective without being very direct. On this account, the deciding is not easy or instantaneous, but requires time and character. Imagine the following monologue taking place with hours to go before the actual jump as the agent keeps looking down at the water some 80 metres below him, taking in what is required for him to take the jump and thereby trying to see whether he can make the decision. (Imagine him roaming along the bridge.)

> Let me try to see if I can jump. Here I go, I am walking along the platform, having at last decided to jump and now I am going to jump. I am inching to the edge ready for the leap. [Pause, as he looks down, imagining himself about to leap.] No, I can't do it. This is awful.
>
> [Another pause] But wait! This is ridiculous. I am supposing that I have decided to jump, and here I am stuck to the platform. I am behaving like those pathetic braggards who, after having mentally rehearsed their jumps, boast that they, at least, will have no trouble doing a jump, and then when the time comes find themselves "glued" to the platform. That's not me. I must remember that I have decided to jump, and now I will jump. In fact, I now find myself more confirmed in my resolve than ever. I can live with my decision to jump. I am going to jump.

The dialectic of this little monologue ought to have a familiar ring to it, for this is just another instance of forming an intention through the agent's bootstrapping himself into forming the intention. The agent doesn't first consider his desires and then decide how his desires are best met in the light of his beliefs. Rather, he imagines the intention having been formed and then sees whether he can live with it.[16] This may be hard, and he may have to keep on trying. But part of the dialectic shows why the bootstrapping approach gives the agent some hope. For his having formed the intention is part of the imagined set-up, and his

---

[16] Although this paper has concentrated on certain somewhat curious, puzzling, cases of decision-making, the lessons are rather wider. Accounts of how (rational) agents make decisions should be far more sensitive to the role of both emotions and the imagination. (See also Tamar Szabó Gendler, 2003, p. 136, who comments that "without the capacity to feel something akin to real emotions in the case of merely imagined situations, we would be unable to engage in practical reasoning.")

reactions, in the scope of this imaginative act, will now take on board not just his fear of contemplating jumping, but also—when he attends to it—the fact that he has already formed the intention. That brings issues of character into the equation, for there is something flawed—rationally, certainly, but also morally—about an agent who, with no change in the relevant background conditions, fails to deliver on his intentions. The agent lacks a certain sort of integrity, *deliberative* integrity. When he sees that his imaginative set-up leaves him in danger of failing to show such integrity, he has the option of either deciding that he can't form the intention after all, or that he can live with the intention and thereby also its execution. Given the way the scenario unfolds, he is able to live with the intention partly because in imagining himself as having the intention he now sees that the deliberative situation has changed: There is a new end worth fighting for, his deliberative integrity, and in performing the jump in the face of his fear of jumping he gives expression to this integrity. As in the case of non-apocalyptic deterrent intentions, I don't mean to suggest that this outcome is mandatory for rational agents. For some agents, their visceral sensation of fear as they contemplate jumping will outweigh any tendency to express their integrity, and for them the result will be the decision that they can't live such an intention, that trying it out was an experiment that failed. All I have tried to show is that it is possible for such an experiment to succeed, and where it does succeed it shows something positive about the agent: possession of an ability that is rationally and morally enhancing.

Earlier I pointed to an obvious asymmetry between the conditional and the non-conditional case: in the conditional case, the imaginative preconstruction involves the agent in a passional reaction to her partner's disregard of something she deeply cares about. Her behaviour in leaving was an expression of her dignity—a value involving her relationship to others. But even though in this case the other-involving aspect dominates the way the agent sees the situation, we can also discern a self-involving aspect. Part of what explains her leaving her partner in her imagined scenario is the insult to a relationship she deeply cares about. But another part involves herself: She has resolved to leave him should he continue his cheating, and her not leaving despite his continued cheating would show a certain weakness of character: lack of deliberative integrity. Part of what she expresses in leaving is her integrity as an agent whose actions match her intentions. If this is right, the preconstruction model of the way we might form conditional deterrent intentions is just a more complex version of the model as it applies to the formation of certain difficult non-conditional intentions, involving a crucial other-involving element to go with a self-involving element that centres on deliberative integrity.

How to have good intentions (and earn a million dollars)

So much for bungy jumping. But how does any of this help with our toxin-drinking problem: the problem of how to form the intention to drink toxin, where having the intention would be enormously beneficial to the agent but where the

act itself is palpably pointless and painful, offering no rewards of any kind?[17] Agent-irrationalists think that no rational agent could form such an intention, at least assuming appropriate constraints (no side bets or hypnosis, for example). I disagree. At bottom, this is the same case as my bungy-jumping problem, and the same solution applies. An otherwise rational agent is able to form such an intention by bootstrapping herself into it, via an imaginative preconstruction of the intention.

Here is a sketch of how she might do it, rather slower and more careful this time than the sketch I provided in the bungy jumping case, since the case is stranger. She argues as follows:

> Sure, drinking the toxin is unpleasant, but I would gain incredibly by forming the intention to do so. So let me try to see if I can live with such an intention. Suppose, for the moment, that I have resolved to drink the toxin. Trouble is, I am aware throughout that after the money is deposited I don't need to drink the toxin to get the money. So in the scope of the supposition that I have formed the intention to drink the toxin, I am also able to reason that I should go back on that intention and not act on it when the time comes. But this knowledge surely destroys my ability to be genuine about such an intention: I can feign having the intention, but I can't be serious about it.
>
> But wait! That reasoning misses the point. I am supposing that I have resolved to drink the toxin. The fact that after the money is deposited I don't need to drink the toxin to get the money is scarcely enough to persuade me not to drink the toxin, for I was already aware of that fact when I formed the intention to drink it; built into the resolution is my awareness, rehearsed above, that I don't actually need to drink the toxin to get the money. In short, in forming the intention, what I intended was to drink the toxin in the face of precisely the sort of reflective rehearsal of reasons not to drink the toxin that I am presently imagining. To be persuaded to give in to this fact after having formed the intention —something I am now imagining as I imaginatively reflect on having formed the intention—would be to display a strange and pathetic fickleness, a deep inability to know my own mind when it comes to the crunch. I'd be like those pathetic braggards who, after having mentally rehearsed their jumps, boast that they, at least, will have no trouble doing a bungy jump, and then when the time comes

---

[17] I agree that it is difficult to see bungy jumping and toxin-drinking as being on a par. That is, I suspect, because we tend to see bungy-jumping as an activity with its own set of expected rewards (feelings of pride, and perhaps esteem from others, say). We don't think there is any special problem about how agents who are swayed by the thought of such rewards are able to form the intention to jump. Even though my version of the example tried to eliminate these features of the situation, I agree that it is hard to think them away. That may well explain our sense that it is not unduly hard to form such intentions.

> find themselves "glued" to the platform. That's not me. In fact, I now find myself more confirmed in my resolve than ever.
>
> So I can live with the intention. This being so, I will resolve to drink the toxin.

It is through rehearsing—perhaps repeatedly —some such argument that the agent can bootstrap herself into forming the intention. Or so I claim. (I suspect, in fact, that only in this way can she form such an intention.) Once again, if this works at all it is because of a feature we first recognised in the case of deterrent intentions: the way in which the evaluative situation has changed in the light of the intention having been formed. There is now a new end to consider, namely the agent's deliberative integrity, one that competes for attention with ends that pull the other way. In the toxin case, we are now envisaging an agent who is able to let this new end find expression in her act of drinking the toxin, an act she imagines herself performing even though she is under no illusion about the painful effects of this act.

I'll conclude by answering just one objection, but it is a pivotal one, and it will allow me to rehearse a theme already encountered when discussing deterrent intentions. Many will find absurd the conclusion that a truly rational agent could form such an intention in full knowledge of what it involves. They will pointedly ask whether it is also true that a fully rational agent could sincerely intend to cut off her arm, say, under the same conditions, knowing full well that cutting off her arm would cause only needless pain and suffering (after all, the one million dollars the agent so desperately wants is already secure by the time she has to act on the intention). It is difficult not to sympathize with this complaint, but I suggest that it may harbor a confusion. The claim is not that it is possible for any rational agent to form hard intentions like the toxin-drinking intention, but that an agent's ability to form such an intention in such a way doesn't contravene her status as a rational agent: how she behaves does not reflect an inherently irrational way of valuing the things she cares about.[18] I don't doubt for a moment that many, if not most of us, will find forming such an intention in such a way impossible. But then most of us apparently find it impossible to come to a firm resolve to perform a bungy jump in full knowledge of what that involves, and yet it shouldn't surprise us that some people are able to do just that with no thought as to any rewards (feelings of pride and the like) that the act might bring. One shouldn't hold it against an agent if she is the kind of person who is able to give expression to a kind of integrity that we also value, even if we ourselves wouldn't give expression to it in just this sort of case. One especially shouldn't hold it against such a person if her ability to give expression to her integrity gets her a million dollars.

---

[18] Not everyone is a "Gordon Liddy", the Watergate felon who reputedly trained himself to be tough-willed enough to put his hand over a flame at will until the skin blackens. (I don't claim that Liddy is the best example to use of someone who puts a tolerance for pain to use for perfectly rational ends!)

Bibliography

Damasio, Antonio R.
1994 *Descartes' Error: Emotion, Reason, and the Human Brain* Harper Collins.
Gendler, Tamar Szabó
2003 Pretense and belief. In *Imagination, Philosophy, and the Arts*, Matthew Kieran and Dominic McIver Lopes ed.s, Routledge, pp. 125-141.
Frank, Robert
1988 *Passions within Reason*: *The Strategic Role of the Emotions* W. W. Norton.
Greenspan, Patricia
2000 Emotional strategies and rationality *Ethics*, vol. 110, pp. 469-487.
Kavka, Gregory S.
1978 Some paradoxes of deterrence *Journal of Philosophy*, vol. 75, pp. 285-302.
1983 The toxin puzzle *Analysis*, vol. 43, pp. 33-36.
1987 *Moral Paradoxes of Nuclear Deterrence* Cambridge University Press.
Kroon, Frederick
1996 Deterrence and the fragility of rationality *Ethics*, vol. 106, pp. 350-377.
Nozick, Robert
1993 *The Nature of Rationality* Princeton University Press.
Pettit, Philip and Michael Smith
1990 Backgrounding desire *Philosophical Review*, vol. 99, pp. 565-592.
Slote, Michael
1989 *Beyond Optimizing*: *A Study of Rational Choice* Harvard University Press.
Solomon, Robert
1976 *The Passions*: *The Myth and Nature of Human Emotions* Doubleday.