

*Fear and Integrity*¹

FREDERICK KROON
University of Auckland
Private Bag 92019
Auckland, New Zealand

I Introduction

I'll begin this paper with an autobiographical example — an instance of a common enough kind of case involving agents who are faced with making a choice they strongly care about, but who have tendencies that incline them towards choosing an option they prefer not to choose. Later in the paper, I apply some of the general lessons learned from this case to a philosophically more familiar example of a hard-to-make choice, and to the well-known problem the example generates for the idea of rational agency: Gregory Kavka's toxin puzzle (Kavka 1983).

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 common among bungy jumpers).² What

1 I am grateful for helpful critical comments from an audience at a conference of the Australasian Association of Philosophy, held at the University of Waikato. Special thanks to three referees for the Journal, as well as to David Braddon-Mitchell, Stewart Candlish, and Richard L Epstein.

2 Of course, the desire is scarcely a motivating reason for jumping, since the fear is what makes jumping so hard in the first place. The point is rather that someone who is able to get to the point where she can perform a bungy jump hopes thereby to be on the way to becoming the kind of person who is able to rid herself of the unfortunate accompaniments of a fear of heights (being seemingly frozen in place, and so on) when exposed to heights in more usual situations.

I found remarkable was the fact that friends and colleagues to whom I mentioned the act thought it at best slightly bizarre. No one thought that there could be anything commendable about the way one succeeds in dealing with fear in such a situation. (Indeed, most were inclined to think the fear must have been exaggerated, since I succeeded in jumping; and all agreed that the trivial nature of the activity somehow trivialised the overcoming of the fear.)³ I disagree. In my view, the decision to jump in the face of the fear was praiseworthy in itself, but not in the way one might think. I didn't, for example, think it displayed 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.)

For many of us, bungee jumping is hard not because of what we believe will happen after we jump (a sequence of breathtaking yo-yoing movements in space that poses no real danger and is terrifying for at most a second or two — or so I had been led to believe) but because of what it takes to get to the point of jumping. If you have a great fear of heights, what is hard — *really* hard — is coming to a sincere and stable intention to jump. Contemplating jumping leaves a tight, visceral feeling of fear that makes forming such an intention well-nigh impossible. Trying to form the intention feels like trying to penetrate a thick transparent wall. One can see what it would be like on the other side and one can't sense any physical impediment to getting there, but for all that the barrier remains unyielding.

Or rather, it remains unyielding unless one finds it 'within oneself' to set the fear aside. But the agent who manages to set her fear aside is not *ipso facto* deserving of special praise. For it simply shows that the agent has found the will to intend to do what she fears doing, so that the desire to perform the jump does after all trump the fear. What it doesn't reveal is the nature of the trumping: in particular, it doesn't reveal that the trumping may show something positive about the agent, that it may show 'character.' It may show no such thing. Perhaps agents succeed in forming the intention by using a form of self-deception: for example, pretending to themselves that they are in a situation where jumping is the only thing that will save them from certain death, or where certain death is precisely what they want. Should they manage this, the way in which the agents overcame their fear in coming to their intention

3 They thought it would have been a different matter if the activity had itself been a worthy one. On a view of this kind, overcoming fear is commendable, and something to the moral credit of the agent, to the extent that the activity enabled by the overcoming of the fear is commendable.

to jump disqualifies their choice from being considered praiseworthy in itself, although of course we may praise the agents for other things (e.g., for having honed their imaginative capacities to such a point).

But sometimes the nature of the way in which the desire to perform the jump trumps the fear does show something positive about the agent. Forming the intention may reveal character (*rational*, indeed *moral*, character), not just will. Furthermore, there is nothing unfamiliar in such displays of character. People often act this way. Much of the present paper is concerned with describing the phenomenology and intentional structure of this way of overcoming fear. Towards the end of the paper, in section V, I finally extend this account to the philosophically more familiar case of Kavka's toxin puzzle, suggesting that fully rational agents can draw upon these same inner resources of character in their attempt to form the toxin-drinking intentions whose formation Kavka declared to be impossible for rational agents. While I certainly don't claim that the appeal to such resources of character can dispose of all cases of rational intentions that are supposedly inaccessible to rational agents, the argument of this section may help to show why the range of such cases is bound to be limited.⁴

II Making hard choices

To generalize the discussion somewhat, let a *hard* choice for an agent be the choice of an option open to the agent, a choice that she prefers to the choice of other options open to her, given her beliefs and commitments, but one that she finds it hard to make because of her having certain traits or dispositions (a fear of heights, say) that are impediments to making the choice but that she lacks the present capacity to eradicate.⁵ I shall assume that the options open to an agent may include forming future-directed intentions, and that the bungee jump case involves the hard choice of forming the stable and sincere future-directed intention to jump.⁶ The claim I have advertised is that the way we make hard choices sometimes reveals a quality of character deserving of praise.

4 Consider especially Kavka's own work on paradoxes of nuclear deterrence (Kavka, 1987). See also note 21 below.

5 I am excluding 'weakness of the will' from such a list of impediments, since that is a trait that is clearly under the agent's rational control.

6 See Bratman (1999) for the distinction between future-directed and present-directed intentions. On my account of the bungee jumping case, the nature of the would-be bungee jumper's fear is such that being able to jump requires having a sincere

One immediate problem is discerning how the hardness of the choice involved in forming an intention can possibly bear on the issue of character. The following, or something like it, may well strike us as a plausible account of the conditions under which a rational agent has, or forms, an intention:

[R1] *Ceteris paribus*,⁷ a rational agent forms the intention to perform some action X, where X is a course of action open to her, if and only if she recognizes that, overall, doing X best meets the desires she has, given the beliefs she has.

But [R1] licenses the following argument. In the case of the intention to do the bungy jump, recall that we are supposing that the agent is in a situation in which jumping has neither intrinsic value nor disvalue for her, but in which she believes it is likely to lead to the satisfaction of one very important instrumental goal: the agent's overcoming a general fear of heights. So the agent has a strong overall preference for performing the jump, based on her firm beliefs about the likely impact of this course of action. Assuming that the agent is rational, it follows from [R1] that she forms the intention to jump. If she has difficulties in doing so, it reflects negatively on her rationality, since by [R1] a rational agent simply chooses on the basis of her recognition of what is in her best interests.

Such a conclusion seems quite implausible. Rational agents may surely find it difficult to do what is in their best interests. As in the bungy jumping example, they may have to overcome the effects of dispositions that, in themselves, scarcely count as irrational (a fear of heights, for example, is on the whole a useful fear).⁸ In the case of a rational

future-directed intention to jump, one that the agent wants to be able to form now so that she can act on it when the time comes (perhaps very soon, perhaps some time from now). Spontaneously, yet intentionally, jumping is simply not an option for her.

7 'Ceteris paribus' covers the fact that theses like [R1] are intended to apply to typical agents in typical situations rather than all agents in all situations. I am putting aside the possibility of ties among actions the agent faces as well as the agent's possible use of psychological tricks (suppressing fear through hypnosis, say). In general, I am interested in the formation of intentions on the part of ordinary agents, who, presented with clear choices, have available to them only ordinary *internal* psychological resources. (The other theses to be discussed, [R2] to [R4], are similarly subject to such a 'ceteris paribus' condition.)

8 Until the fear become a *phobia* (sometimes characterized normatively as an *irrational* fear, one without supporting reasons), when it begins to impose an unbearable restriction on the choices available to those who have it.

agent with such dispositions, recognizing which course of action is in the agent's best interests does not suffice for her forming the intention to undertake that action, since she has to find a way to overcome the effects of such dispositions. These are cases for which [R1] simply fails to capture the phenomenology of choice.

An obvious first response to this objection is that this argument forgets the fear that the agent faces as she contemplates jumping, and it is this fear that tips the balance: it may well outweigh any intrinsic or instrumental benefits deriving from doing the jump, and in that case jumping would not be the rational course of action, so that she does not form the intention to jump. But such a response stands things on their head. For on our present understanding of [R1], it is hard to see how the fear can be factored in. The agent acknowledges that if she actually jumps any fear she may then feel would not only be misplaced (since she doesn't think that she would be in real danger) but would also be outweighed by exhilaration. As for the fear she *now* feels, while this is certainly a psychological obstacle to her forming the intention it is hard to see how it can coherently count as one of the costs to be incorporated into the calculation of whether to form the intention in the first place (as opposed to just being relevant to whether the agent *does* form the intention).

It seems, then, that [R1] fails: it fails to factor in the fear (the 'fear-factor' problem for [R1]). But what is the deeper source of this failure? Part of the answer is that [R1] holds that when the agent decides on her intentions she makes reflective reference to her desires as desires, with the agent reaching her choice by recognizing that the action contemplated has the attractive property of promising to satisfy those desires. In general, however, rational agents do not choose by determining what their desires are and then deciding how those desires, *as* desires, are best satisfied. Generally, the desire itself features only as part of the motivating reason for forming the intention, not as part of the justifying reason, and to this extent the desire figures in the 'background' and not the 'foreground'.⁹ We need an account that more adequately recognizes this motivating level of an agent's engagement with her reasons for forming the intention.

There is an obvious alternative to [R1] that assigns desires (and beliefs) to the 'background':

[R2] *Ceteris paribus*, a rational agent forms the intention to do X, where X is a course of action open to her, if and only if she

9 I owe the terminology to Pettit and Smith (1990, 568).

comes to an overall preference for doing X on the basis of considering the consequences of each of the actions open to her in light of the *content* of her beliefs and desires (rather than in light of the way in which performing an action promises to satisfy the desires she has, in light of the beliefs she has).

According to [R2], a rational agent typically forms an intention on the basis of some comparative first-order practical reasoning that makes appeal to what she believes and what she desires (for example, she might form the intention to take the bus rather than the train to work on the grounds that the trains, unlike the buses, are never on time, and that her boss will not value her if she is not on time). She doesn't use second-order reasoning about how her beliefs and desires *as* beliefs and desires are best accommodated by her taking the bus.

But while [R2] is an improvement on [R1], it still does not answer our earlier worry. It still implies that a rational agent will form the intention to perform a bungy jump simply on the basis of her judgment that jumping is her best course of action, where this is now construed as the first-order judgment that jumping will help her overcome her fear of heights and will not hurt her (and may even thrill her). Although it gives a better account of the nature of the agent's engagement with her reasons, it still fails to factor in the fear that makes it so hard for the agent to form the intention to jump, precisely because this fear doesn't seem to play a role as a motivating reason. The agent engages with the fear at a different level.

III Rational choice and the imagination

So something different is needed. In my view, [R2] wrongly assumes, with [R1], that an agent's adopting a certain intention is based on a broadly calculative comparison of different possible states of affairs. There is an emerging body of empirical research that suggests that this is not the way in which agents typically come to make decisions. Work by Antonio Damasio and others suggests that when we make practical decisions our reasoning is action-guiding only in cases where we imaginatively engage with the potential outcomes of contemplated actions. Indeed, agents who can't engage emotionally with the impact of their possible choices can easily strike us as irrational rather than rational. In particular, individuals with damage to their ventromedial prefrontal cortex are able to articulate reasons for undertaking various courses of action but are unable to use those reasons as their motivational basis for action. Instead, their actual behaviour will often seem erratic and

unplanned. What Damasio, Bechara, and others have concluded from studies involving such patients is that autonomic responses play a central role in decision-making (Bechara *et al* 1994). We appear to engage imaginatively with the potential consequences of various courses of action, thereby activating our emotional response mechanisms; the presence of ‘somatic markers’ that encode the results of this activation then helps to guide our actual behaviour. While the details of this account are still widely debated, its overall structure has now gained wide acceptance. On the picture proposed, such simulated emotions are a fundamental feature of our cognitive repertoire.¹⁰

This is not the only evidence for the role of our imaginative engagement with the potential consequences of undertaking different courses of action. Contrary to the relatively ‘bloodless’ picture suggested by [R1] and [R2], deciding what we should intend to do has its own phenomenology, one that appears to reflect our felt engagement with the way things would be if we were to undertake this or that course of action. Not surprisingly, it is sometimes hard to discern this engagement, since many of our choices do not involve very vivid desires, or desires that markedly affect bodily states. We should therefore not be surprised if a broadly calculative account like [R2] strikes as more or less right for many cases. Often enough, however, it is easy to discern our felt engagement with the consequences of potential courses of action. When we are trying to decide which route to take to our destination tomorrow, we may involuntarily wince as we imagine the real possibility of falling as a result of taking one route rather than another, and this reaction — one that is best understood in terms of our simulating the likely outcome of taking the offending route — is very much part of what makes us intend to take the less painful route.¹¹ Such examples are easily multiplied.

Given the evidence, then, a better model of the way in which properly constituted rational agents decide what they should intend to do assigns an important role to their ability to engage imaginatively with the consequences of intending to do one thing rather than another: given a possible course of action, such an agent *imagines* herself in a situation containing the consequences of her having chosen this action, engaging emotionally with the envisioned impact of her (simulated) choice — as

10 I don’t mean to imply that such simulated emotions are ipso facto full-fledged emotions, although Damasio *et al.* assume this. Gendler and Kovakovich 2006 have recently argued that it follows from the work of Damasio *et al.* that the apparent ‘emotions’ experienced in the course of our interaction with fiction are also genuine emotions (contra Ken Walton’s views in particular; cf. Walton 1990, ch. 7).

11 See, for example, Currie and Ravenscroft (2002, 19-20; 97-9).

it might be, feeling disdain, delight, fear, etc. Her final decision will then involve a comparative judgement among these simulated choices, one that takes on board both the likelihood that the choices will have this impact and her felt engagement with the impact.

As a first attempt, this yields something like the following:

[R3] *Ceteris paribus*, a rational agent forms the intention to do X, where X is a course of action open to her, if and only if she comes to a preference for doing X on the basis of imaginatively considering (for each of the actions open to her, and in light of her beliefs and desires) the consequences that would obtain if she were to choose to perform that action.

But note that [R3] still implies that a rational agent will form the intention to do a bungee jump simply on the basis of her judgement that jumping is her best course of action; her imaginatively considering the consequences of jumping should not make any substantial difference, since she strongly believes that she should not imagine feeling fear in the course of imagining herself jumping (if anything, she should simulate feeling exhilaration). Once again, there appears to be no way to factor in the fear that the agent feels in her attempt to form the intention to jump.

Only one more thing is needed to make room for the fear the agent feels, and that is something perhaps already implicit in the way we set up [R3]. Surely in imaginatively considering the consequences that would accrue if she were to choose to perform some action, the agent is in effect imagining herself in a situation in which she chooses to perform the action and faces the impact of the action, with her imagined choice and its impact then considered in light of her beliefs and desires. [R3] explicitly captures the latter aspect (the way the agent imagines the outcome of acting on the choice), but what [R3] doesn't explicitly capture is the former aspect (the fact that the agent is also imagining choosing to perform the action). Once we include that aspect, something else needs to be taken into account, namely the fact that the agent thereby also (implicitly) imagines having earlier formed the intention to perform this action — after all, the agent's consideration of the various courses of action open to her is taking place in the context of her attempt to determine which action she should intend to perform, and so we are implicitly assuming that, in the situation at hand, having earlier formed a particular intention is what underlies the agent's choice of a particular action. This suggests we replace [R3] with something like the following:

[R4] *Ceteris paribus*, a rational agent forms the intention to do X, where X is a course of action open to her, if and only if she comes to a preference for doing X on the basis of imagining, for each of the actions open to her, that she chooses to perform that action (and hence that *she has earlier formed the intention so to act*) , and then imaginatively considering, in light of her beliefs and desires, the consequences of her having made this choice, *and hence of her having formed the intention so to act*.¹²

As we will see, this new formulation is finally able to accommodate the fear-factor.

IV Rational choosing through simulating choice

[R4]’s model of how rational agents form intentions explicitly makes room for the impact of forming an intention. This gives it two important advantages over our previous models. In the first place, as promised it captures a clear sense in which forming the intention to perform a bungy jump is genuinely hard for a rational agent in the grip of a fear of heights. While [R1] – [R3] had difficulty capturing such a sense (the fear-factor problem), a straightforward application of [R4] suggests that when the agent simulates the choice situation she finds herself confronted by a serious gap between imagining she has formed the intention to do the jump and imagining the consequences of jumping: assuming she is a reliable simulator (take this to be part of what is required by rationality on model [R4]), she will imagine herself having formed the intention and then find herself practically unable to deliver on the intention because of her fear. (To the extent that she is a good simulator, this matches how matters would in all likelihood stand were she actually to go through a process that she is now merely imagining: ‘Right, I am resolved to jump, and here I go — no, I just can’t!’) [R4], therefore, properly factors in the fear that posed a problem for our earlier models of how rational agents form intentions.

12 To the extent that we can’t imaginatively contemplate the consequences of every possible course of action open to us, [R3] and [R4] may seem wildly unrealistic. If this is a problem, however, it is also a problem for [R1] and [R2], and is to be solved in the same way: many of the courses of action open to us we can reject out of hand because we know their cost from past experience; in the case of others, we understand their attraction from past experience. In implementing a recipe for intention-formation like [R4], we are allowed to take short-cuts.

By thus factoring in the fear, this way of applying [R4] seems to imply that the agent may well find it too hard to form the intention to jump. But a second advantage of [R4] is that it permits another way in which an agent can simulate her choices in order to arrive at a settled decision, and this alternative way finally secures us a feature promised earlier: it shows how the agent can overcome her fear in a manner that shows rational, and even moral, character. To see this, it is important to highlight what seems a problematic feature of [R4]. How — when the agent is trying to decide what to intend — can she then use the thought that she has successfully formed the relevant intention as part of the reasoning towards forming the intention, along the lines of [R4]? Doing so seems blatantly circular.

But this misunderstands the nature of the proposal. In ordinary cases, a rational agent who contemplates forming an intention need not attend to anything other than the likely consequences of acting on the intention. But matters are different in the case of hard choices because of the impediments that would-be intenders face. A would-be intender may need to find something else to motivate her. I emphasized above that, if a person has a fear of heights, her attempt to simulate a situation in which she has formed the intention to do a bungy jump is likely to produce keenly felt resistance ('I am resolved to jump — no, I can't!'). I now want to suggest that such resistance can itself become the creative spur to successful simulation, eventually leading to the formation of a genuine intention. While the entire procedure assumes a context in which the agent first simulates having formed the intention, there is nothing circular in the procedure.

On this alternative way of applying [R4], the agent places herself in the full imaginative context in which she has successfully formed the intention, with a view to seeing whether she can in the end live with the intention (that is, genuinely endorse it) despite the fear that prevents her at first from forming it. On the resulting picture of how the choice is finally made, 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 fifty metres below her. Imagine her 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 she looks down, imagining herself 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 brag-garts 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.

This little monologue shows the agent forming an intention by bootstrapping herself into it. The agent imagines the intention having been formed and then sees whether she can act on it in the context of this simulation. This may be hard, and she may have to keep on trying. But the dialectic of the monologue reveals why the bootstrapping approach gives the agent some hope. For her having formed the intention is part of the imagined set-up, and her reactions, in the scope of this imaginative act, will now take on board not just her fear on contemplating jumping, but also the fact that she has already formed the intention. That brings issues of character into the equation, for there is something to be valued in an agent who delivers on her intentions, despite the difficulty she faces in doing so.

We might say that the agent displays a certain fickleness — a lack of a certain kind of integrity (what I will call *deliberative integrity*) — when she finds herself in an imaginative set-up in which she has the intention to jump but then fails to act on her intention. In the situation envisaged above, the agent sees that she is indeed in danger of failing to show such integrity, and this leaves her with the option of either deciding that she can't act on the intention ('No, I can't do it!'), or that she can. Given the way the imaginative scenario unfolds, she eventually finds herself able to act on the intention partly because in imagining herself as having the intention she has come to see that the deliberative situation has changed: there is an important new end in view — her deliberative integrity — and when, in this imaginative context, she finds herself resolute and able to perform the jump in the face of her fear of jumping she gives expression to this hard-fought integrity.¹³ There is now a matching of action to intention that has rational and even moral significance, rather like the matching of action to values and plans famously lauded by Bernard Williams in his critique of utilitarianism. For not only is the capacity for being deliberatively steadfast a precondition

13 I don't mean to suggest that the agent jumps because she thereby fulfils an important desire, the desire to show deliberative integrity. Expressing such integrity is different from fulfilling the desire that one have or show such integrity, just as expressing friendship is different from fulfilling the desire that one show friendship. It may be that the agent's action only shows genuine, rationally worthy, character if she gets to the point where she can be said to express her integrity rather than merely to fulfil a desire for showing integrity. (I am here indebted to Christine Swanton.)

for a rational and perhaps moral way of life, but individual acts and states in which an agent expresses deliberative integrity may contribute significantly to her plan for living by her lights. After all, the goals that are met through such acts may be important ones, while meeting them may be difficult — it may seem easier simply to adjust one's goals. If so, the state and acts of the agent as she expresses deliberative integrity in the attempt to meet these goals should strike us as having rational, and (if the goals are morally worthy) even *moral*, value.¹⁴

Up to this point, of course, the agent's reasoning and her efforts at expressing deliberative integrity have occurred in the context of her simulation. My proposal is that the simulation recorded in the little monologue above is a precursor to her forming the intention (*really* forming it). When, in the course of her repeated simulations, she finally comes to a settled preference for acting on the intention and undertaking the jump, the agent recognizes that she is at the point where she can really live with the intention. And so she forms the intention. What happens next is crucial. If the agent forms the intention, but then doesn't after all act on it (despite everything, she remains glued to the platform when the time comes), it shows that she was a poor simulator. Such lack of self-knowledge would, not surprisingly, be a considerable blow to her self-esteem. But it would also reflect poorly on her rationality, and if, as I am supposing, the agent in question is truly a rational agent, we should suppose that she is going to act on her intention, in line with the (repeated) successful simulations that led her to adopt the intention. By thus doing what she earlier successfully simulated doing, the agent lets her behavior display the very feature of character (deliberative integrity to a praise worthy degree) that earlier produced her successful simulations.¹⁵

14 For Bernard Williams' notion of integrity, see his 'Integrity,' in Smart and Williams (1973, 108-17). For Williams, integrity is counted in terms of the commitments that a person identifies with most deeply. Such an account has been criticized for lacking the moral dimension of our common-sense notion of integrity (see, for example, Damian Cox, Marguerite La Caze, and Michael Levine, 1999). My own view is that the capacity for integrity so understood might be a kind of ground-level virtue, one that it is necessary to have if an agent is to lead a moral and rational life. The same may be true of the capacity for *deliberative* integrity, given the central role played by the integration of intention and action. Still, 'deliberative integrity' may not be quite the right term. In 'Useful Intentions' (ch. 12 of Sobel, 1994), Howard Sobel talks more neutrally of 'steadfastness' and 'resoluteness,' and suggests, in a way that comports well with the spirit of the present paper, that our common caring and valuing of such states shows that rational agents can be capable of intentions whose adoption makes actions rational that would otherwise be irrational.

15 When I put the point this way, I have in mind a suitably thin notion of 'charac-

Nothing in what I have said implies that this final outcome — successful actual formation of the intention to jump, followed by the actual jump — is mandatory for rational agents. For a possible outcome of the attempted simulation might be that the agent decides that she cannot after all live with the intention. Some agents will find that their visceral sensation of fear as they imagine acting on their intention to jump will outweigh any tendency to express deliberative integrity, and for them the result will be 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.

V The toxin puzzle

Forming the intention to do a bungy jump in the face of a strong fear of heights counts as a hard choice. Sometimes what makes the choice hard is not a psychological disposition (say, the tendency to feel near-incapacitating fear in certain situations) but something more akin to our own rational tendencies. A famous such example is Gregory's Kavka's toxin puzzle (Kavka, 1983), which involves a seemingly rational choice to form an intention whose execution is antecedently recognized as irrational. This is how Kavka describes the case:

You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, 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. (Kavka 1983, 33-4)

The only other conditions that the billionaire places on the offer are that you are not to make any side-bets, do anything that will cause you to become irrational, or arrange for any way to avoid the effects of the toxin. Assume that you have overwhelming reason to believe that the

ter.' The idea of deliberative integrity, and the thought that such integrity can be marshalled in the course of implementing one's design for living by one's lights, are almost formal in content, and have little to do with the thick notion of character that has recently become the subject of philosophical controversy (see, for example, Doris, 2002).

facts are as the billionaire states them (in particular, that he has access to a device that can tell whether you have formed the required intention), and that the contract is a binding one.¹⁶

Deciding on the intention to drink the toxin is a species of hard choice. Even if your first thought is to agree to the billionaire's proposal since being ill for one day is a small price to pay for a million dollars, you soon realise that you won't even have to become ill in order to win the money. All you have to do is to *intend* to drink the toxin, not to *actually* drink it. But since you also know that after midnight you won't have any reason to drink the toxin, and every reason not to, how can you even *intend* to drink the toxin? At that point, you would already have been paid, and drinking the toxin would only make you unnecessarily ill. As you contemplate being violently ill from drinking the toxin in the absence of any reason to drink it, the barrier to the formation of the intention begins to loom large.

Indeed, some philosophers, beginning with Kavka, have argued that no rational agent could form such an intention (despite it being rational to form the intention), since a rational agent intends to Φ only if she antecedently has good reason to Φ and since in the present case she has no such reason (Φ -ing merely produces needless pain).¹⁷ I disagree both with the starting premise and the conclusion. First of all, the premise embodies something like [R2]'s account of the basis on which a rational agent forms her intentions, and we have already seen reason to be suspicious of that account. Second, once we turn to the alternative account [R4], it becomes easier to see how a rational agent might form the intention to drink the toxin despite the apparent irrationality of what is involved in acting on the intention. A rational agent is able to form such an intention by bootstrapping herself into it, via imaginatively putting herself into a situation in which she has formed the intention and contemplates acting on it, with a view to seeing whether she might be able to live with 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:

16 I doubt that such a machine is really possible (the idea that we can form an intention by implementing a successful simulation experiment surely attests to this — such an experiment takes time, and it may not be clear just when the experiment is complete). Like David Gauthier (Gauthier 1998, p. 47), I think it is enough if the billionaire has access to a very astute judge of the real intentions of his fellows.

17 Kavka 1983, 35-6

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 firmly resolved to drink the toxin, having seen how important it is for me to have formed such an intention. 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 these reasons 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 braggarts who, after having mentally rehearsed their jumps, boast that *they*, at least, will have no trouble doing a bungee jump, and then when the time comes 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.

If all goes well, it is through rehearsing, perhaps repeatedly, some such argument that the agent will bootstrap herself into forming the intention. Once again, this works because of a feature we first recognized in the case of the decision to perform a bungee jump: the way in which the evaluative situation changed in the light of the intention having been formed. There is a new end in view, namely the agent's deliberative integrity, one that competes for attention with ends that pull the other way. In the toxin case, we have an agent who is able to let this new end find expression in her (simulating) drinking the toxin, an act she imagines herself performing even though she is under no illusion about the painful effects of this act. After perhaps repeated simulations of this kind, the agent finally finds herself able to live with the intention to drink the toxin, and so she finally forms it, ready for the billionaire's judgement on her new-found resolve.

Having thus formed the intention, the agent will of course act on it. Or rather, she will act on it if she has rehearsed well, and so can bring off what she has simulated bringing off — as in the bungee jumping case, something that should go without saying if the agent is fully rational. Her willingness to act on her intention is not something that is

separate from her forming the intention, but simply confirms an ability she already displayed in the forming of the intention, in the way she found herself repeatedly able to simulate its execution. (This is why it would be very wrong to respond to the agent's willingness to act on the intention by accusing her of being a plan-worshipper. From the agent's point of view, that rather turns things on their head. As she sees it, not acting on the intention after what she took to be a successful simulation experiment would display an embarrassing and alarming lack of self-knowledge.)¹⁸

As before, it is important not to exaggerate the claim of the argument I have just presented. The claim is not that every agent facing such a situation should be able to form the intention this way on pain of irrationality. Perhaps there is nothing that the agent can do that will alter her refusal to drink the toxin under these imagined conditions. Despite knowing how valuable forming the intention would be, and hence how valuable it would be to get to the point where she can drink the toxin, she is not able, in her attempts at simulating the choice situation, to let the value she places on her deliberative integrity trump the disaffection for the pain and discomfort that she contemplates she will feel on drinking the toxin. Although we as onlookers might think that such an agent places too high a premium for her own good on not feeling pain and discomfort, nothing I have said rules out the possibility that some rational agents will behave this way. Still, given what she deliberately foregoes, such an agent can scarcely be held up as the archetype of a rational agent, someone to be emulated in preference to the agent whose ability at successful simulation gains her a million dollars.¹⁹

18 Note also how strategically important such self-knowledge is in the agent's quest to win a million dollars. If she is the kind of agent who doesn't know her own mind in such hard cases, she will scarcely be able to convince the billionaire's intention-checking apparatus or judge that she has formed the intention, and so she misses out on the million dollars.

19 It might be thought to be in the spirit of the present account to claim that such agents are not fully rational, since they don't care enough about the importance of deliberative integrity. But that would be a mistake. Such agents might well care enough: once they have *genuinely* formed intentions, they might consistently act on them, and they might consider it important to do so. The fault of such agents, if it is a fault, is rather different. As in the bungee jumping case, it is that they can't bring to successful completion an experiment they consider important (that of *genuinely* forming a certain intention they deem it important to form), and all because they can't get to the point where they can stably express deliberative integrity in the context of an imaginative attempt to form the intention. This is not yet to impugn their rationality. (In more extreme cases, where the benefits are even higher relative to the costs, our judgement will no doubt be harsher. Suppose our billionaire promises a hundred million dollars to help the victims of a disaster — a

[R4] is not the only account to defend the claim that rational agents might be able to form the intention to drink the toxin. Perhaps the best known defence comes from David Gauthier, who employs the idea of ‘constrained maximizing,’ first introduced to deal with the case of bargains (Gauthier 1994).²⁰ Applying the idea to the toxin case, Gauthier argues that

[i]t is rational ... to form the intention to drink the toxin, and to drink it. More generally, it is rational to form an intention, if one reasonably expects at the time that forming and executing it will better realize one’s objectives than not forming it; and it is rational to execute an intention, if one reasonably expects at that time that one’s objectives will be better realized after executing it than they would have been had one not formed it. (Gauthier 1998, 50)

But Gauthier’s account is open to an obvious objection. The agent knows that after midnight she will have no other reason to drink the toxin than the fact that she earlier formed the intention, and that the package of intention plus its execution is better than the package of no-intention and so no execution. But even though it is true that she ‘reasonably expects ... that [her] objectives will be better realized after executing [the intention] than they would have been had [she] not formed it,’ the fact remains that on Gauthier’s account she also reasonably expects that her objectives will be even better realized if she doesn’t execute the intention after forming it. Michael Bratman has argued that the availability of this latter piece of reasoning shows that it is at best rational to adopt the intention in what he calls a *non-deliberation-based* way (for example, by way of self-hypnosis), and that, acquired this way, it is eminently rational to reconsider the intention when the time comes for its execution (Bratman 1999, p. 106). Bratman also thinks that if the agent is indeed

goal strongly endorsed by our agent — if the agent succeeds in forming the intention to drink a very mild toxin. Suppose she finds herself unable, in her simulation of the choice situation, to drink the toxin. Despite the utter urgency of succeeding — lives depend on it — she can’t bring herself to act on the intention since she knows that drinking the toxin will cause unnecessary discomfort and she can see no outweighing value it will serve. Not finding herself able to act on the intention in her simulation of the choice situation, she fails to form the intention, and fails thereby to ensure the arrival of life-saving medical help. Such behavior seems paradigmatically irrational.)

- 20 See also Harman 1998. Harman thinks ‘the issue is whether you can adopt a firm, strong-enough, long-enough-lasting intrinsic concern (desire or intention) to drink the toxin to celebrate receiving the money’ (Harman 1998, 88). Harman’s talk of celebration, and the use to which he puts the idea, seems misplaced, however. Surely what we should celebrate is our resolve as we drink the toxin, for that is what got us the money.

confined to a deliberation-based way of adopting intentions, it is clearly *not* rational for the agent to adopt the intention, since deliberation-based ways of adopting intentions must focus on the consequences of the actions that are the objects of intentions (Bratman 1998, 1999).

No such concerns can apply to the solution presented above. On that solution, the agent does deliberate, but not simply by focusing on the consequences of adopting this or that option. Instead, the agent engages in a bootstrapping form of deliberation where the very process of deliberation brings a new end into play — her deliberative integrity — which in turn is able to tip the balance in favour of adopting the intention. On my account, that is what makes all the difference.²¹

Received March 2006

References

- Bechara, A., A.R. Damasio, H. Damasio, and S.W. Anderson. 1994. 'Insensitivity to Future Consequences Following Damage to Human Prefrontal Cortex.' *Cognition* 50 (1994) 7-15.
- Bratman, M. 1998. 'Toxin, Temptation, and the Stability of Intention.' In Coleman and Morris 1998.
- _____. 1999. *Intention, Plans, and Practical Reason*. Stanford: CSLI Publications. (First published by Harvard in 1987.)

21 Kavka also argued that the toxin case was a special case of the more general phenomenon that '[o]ne cannot intend whatever one wants to intend any more than one can believe whatever one wants to believe' (Kavka 1983, 36). In earlier work, he had argued that there are cases in which there is great rational and moral benefit attached to forming the conditional intention to do great harm (as a threat to deter nuclear aggression, say), and that rational and moral agents are simply unable to form such extreme deterrent intentions of this kind. On the surface, the kind of strategy I have discussed is powerless to overturn this conclusion. Given what is envisaged in such intentions (say, the horror of a useless retaliatory nuclear strike should one's enemy launch a first strike), it is plausible to suppose that the only possible outcome of a simulation experiment in which a rational and moral agent imagines that the conditional intention is in place, and that the condition has been fulfilled, is her utter inability to form the intention — and this even when forming such an intention would be the moral and rational thing to do. If so, this suggests that Kavka was right about extreme deterrent intentions, even if he was wrong about toxin-drinking intentions. (In Kroon 1996, I had argued that Kavka was also wrong about extreme deterrent intentions, on grounds I now regard as contentious; for a rejoinder, see Greenspan 2000, especially footnote 24.)

- Coleman, J. and C. Morris. (eds.). 1998. *Rational Commitment and Social Justice: Essays for Gregory Kavka*. Cambridge: Cambridge University Press.
- Cox, D., M. La Caze, M.P. Levine. 1999. 'Should We Strive for Integrity?' *Journal of Value Inquiry* 33 (1999) 519-30.
- Currie, G. and I. Ravencroft. 2002. *Recreative Minds*. Oxford: Oxford University Press.
- Damasio, A.R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Harper Collins.
- _____. 1999. *The Feeling of What Happens*. New York: Harcourt.
- Doris, J. 2002. *Lack of Character*. Cambridge: Cambridge University Press.
- Gauthier, D. 1994. 'Assure and Threaten.' *Ethics* 104 (1994) 690-721.
- _____. 1998. 'Rethinking the Toxin Puzzle.' In Coleman and Morris 1998.
- Gendler, T. and K. Kovakovich. 2006. 'Genuine Rational Fictional Emotions.' In *Contemporary Debates in Aesthetics and the Philosophy of Art*, ed. M. Kieran. Oxford: Blackwell.
- Greenspan, P. 2000. 'Emotional Strategies and Rationality.' *Ethics* 110 (2000) 469-87.
- Harman, G. 1998. 'The Toxin Puzzle.' In Coleman and Morris 1998.
- Kavka, G.S. 1983. 'The Toxin Puzzle.' *Analysis* 43 (1983) 33-6.
- _____. 1987. *Moral Paradoxes of Nuclear Deterrence*. Cambridge: Cambridge University Press.
- Kroon, F. 1996. 'Deterrence and the Fragility of Rationality.' *Ethics* 106 (1996) 350-77.
- Pettit, P. and M. Smith. 1990. 'Backgrounding Desire.' *Philosophical Review* 99 (1990) 565-92.
- Smart, J.J.C. and B. Williams. 1973. *Utilitarianism: For and Against*. New York: Cambridge University Press.
- Sobel, J.H. 1994. *Taking Chances: Essays on Rational Choice*. New York: Cambridge University Press.
- Walton, K. 1990. *Mimesis as Make-Believe*. Cambridge, MA: Harvard University Press.

