Imaginative Motivation

FREDERICK KROON
University of Auckland

This article argues for a certain picture of the rational formation of conditional intentions, in particular deterrent intentions, that stands in sharp contrast to accounts on which rational agents are often not able to form such intentions because of what these enjoin should their conditions be realized. By considering the case of worthwhile but hard-to-form 'non-apocalyptic' deterrent intentions (the threat to leave a cheating partner, say), the article argues that rational agents may be able to form such intentions by first simulating psychological states in which they have successfully formed them and then bootstrapping themselves into *actually* forming them. The article also discusses certain limits imposed by this model. In particular, given the special nature of 'apocalyptic' deterrent intentions (e.g. the ones supposedly involved in nuclear deterrence), there is good reason to think that these must remain inaccessible to fully rational and moral agents.

1. INTRODUCTION

Consider the case of very useful, but potentially very dangerous, threat-behaviour, say the case of the deterrent threats involved in maintaining a policy of nuclear deterrence. Gregory Kavka argued some time ago that such 'apocalyptic' threats give rise to a paradox.[1] In brief, while forming such a deterrent retaliatory intention may well be the rational and moral thing to do in view of the threat to the agent's survival, it seems that any would-be intender must, paradoxically, be irrational and immoral because of the awfulness and pointlessness of the harm such an agent agrees to inflict should this intention fail to deter.[2] Kavka

[1] See Gregory Kavka, 'Some Paradoxes of Deterrence', *Journal of Philosophy* 75 (1978), pp. 285–302, and *Moral Paradoxes of Nuclear Deterrence* (Cambridge, 1987). Kavka in fact thought that nuclear deterrence gave rise to a number of different paradoxes of deterrence, but I will consider only what I take to be the most serious such paradox, one involving a tension between agent-rationality and option-rationality. (Daniel Farrell argues that this is in fact the only case that constitutes something akin to a paradox; see his 'On Some Alleged Paradoxes of Deterrence', *Pacific Philosophical Quarterly* 73 (1992), pp. 114–36.) The expression 'apocalyptic threat' is Gauthier's (from his 'Assure and Threaten', *Ethics* 104 (1994), pp. 690–721). Gauthier applies it to any deterrent threat 'that, should it fail, would require [the agent who made the threat] to bring utter disaster on her head' (p. 719). Note that Gauthier talks of apocalyptic *threats*, to emphasize the importance of the intention's being made known to the threatened party, but I will use 'intention' and 'threat' interchangeably.

[2] David Gauthier's 'Deterrence, Maximization, and Rationality', *Ethics* 94 (1984), pp. 474–95, is an influential defence of the claim that such deterrent intentions may be entirely rational. Gauthier also argued that if such intentions failed to deter it would be rational to act on them as well (on the grounds that the rationality of *forming* an intention implies the rationality of *acting* on the intention, absent a change in the background conditions). For criticism of Gauthier's argument, see Michael Bratman, *Intention, Plans,*

doi:10.1017/S095382080900346X
Utilitas Vol. 21, No. 2, June 2009
Printed in the United Kingdom

thought that the right solution to this paradox of deterrence was that the rationality and morality of actions and of agents sometimes come apart. He thought that forming such an apocalyptic intention is the moral and rational *option* for the agent facing such a nuclear threat to his survival, but that no truly rational, moral *agent* can adopt what is the rational and moral option in this case. Kavka, we might say, was an *agent-irrationalist* and *agent-immoralist* but an *option-rationalist* and *option-moralist* about such intentions.

Although the case of nuclear deterrence, on its classic 'mutually assured destruction' construal, has been the most widely discussed instance of this paradox, the puzzle also extends to certain less apocalyptic scenarios involving threats likely to deter unwanted behaviour. Thus consider a person's threat to leave her partner, in a relationship she deeply cares about, if her partner continues to deceive her. Assume that the two agents believe each other to be rational, and that each understands the choice facing the other agent. It may be clear to the first agent that, all things being equal, it wouldn't be in her own or her partner's best interests if she were to leave him, even if he chooses to continue to deceive her (considerations of emotional, financial and 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 a sincere threat to this effect (assume that bluffing is out of the question). The problem, as before, is how the 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 behaviour would only make things worse for her. Forming the intention would be the rational and, indeed, moral option, given its probable success in preventing the partner's deception, although actually forming the intention is prima facie impossible, given her status as a rational agent. (*De dicto* impossible, not *de re*. The claim is not that forming the intention is impossible for X, where X is a rational agent, but that 'X [the agent in question] is rational' and 'X formed the deterrent intention' are incompatible. The first construal implies that forming the intention is not an option genuinely open to the agent, in which case there would be no paradox needing resolution.)

Cases of this kind have seemed especially embarrassing for a consequentialist approach to rationality and morality, since they don't

and Practical Reason (Stanford, 1999), pp. 105–6. A rather different picture emerged in Gauthier's 'Assure and Threaten', which defends a more complex account of the conditions under which commitment behaviour counts as rational. I describe and criticize Gauthier's account in section 2.

suggest simply that the (probable) rightness of an action and an agent's having a right motive (by usual standards of right motive) come apart in certain familiar situations, but that these are cases in which they *must* come apart – indeed, that agents must be prepared to equip themselves with motives that are rationally and even morally wrong by usual standards of wrong motive if right action is to result. In my view, however, consequentialism has nothing to fear from such arguments, at least in the case of non-apocalyptic threats. Beginning in the next section with a critical discussion of two of the most prominent arguments for agent-irrationalism about deterrent threats (both by theorists who are in other ways friendly to consequentialist ideas – Gregory Kavka and David Gauthier), I describe an alternative account that uses the idea of the imaginative *pre*formation of a conditional intention and explains how such a preformation can lead to the *actual* formation of the intention.[3] I then argue that this account is able to accommodate the case of non-apocalyptic threats by showing how the conditionally intended behaviour can count as irrational *apart* from the preformation, and rational on the *basis* of the preformation. The final section returns us to the difficult case of apocalyptic threats.

2. AGENT-IRRATIONALISM AND DETERRENT THREATS

How good is the argument for thinking that rational agents cannot form and sustain such deterrent intentions as the threat to inflict massive harm if one's enemy strikes or the threat to leave one's partner if he continues cheating? In schematic form, the problem confronting such agents is this. Suppose that P – a rational agent – strongly desires that some other agent (Q) not do C, and that she 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. 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', despite recognizing that forming this intention is the rational thing to do.

We seem to have the following problematic triad (T):

(T1) P is a (fully) rational agent who has the opportunity to form the conditional intention to do E if C should happen.

[3] In 'Deterrence and the Fragility of Rationality', *Ethics* 106 (1996), pp. 350–77, I present an earlier version of such an account, applied only to the case of apocalyptic deterrent intentions, and arguing (wrongly, as I now think) that such 'intentions' can be adopted by rational and moral agents in the full knowledge that what they conditionally enjoin is irrational and/or wrong.

(T2) It is clear to P that it would be rational for her to form this intention (since forming this intention is an option that, given her beliefs, best fits her overall desires, both intrinsic and instrumental).

(T3) Given the likely impact of doing E should C occur (given P's beliefs and desires), it is clear to P that if she forms the intention under the conditions under which it is rational for her to form the intention, then if C should occur it would nonetheless be against the balance of reasons for her to do E.

The agent-irrationalist thinks that it follows from (T1) and (T3) that P avoids forming the intention in question despite the rationality of forming the intention, but that inconsistency is avoided because it doesn't follow from (T1) and (T2) that P chooses to form the intention. According to the agent-irrationalist, it is not part of the notion of agent-rationality that a rational agent always chooses the rational option facing the agent, the option that best comports with her overall beliefs and desires.

But what precisely is the argument for the claim that (T1) and (T3) entail that P does not form the conditional intention to do E if C should happen? The most likely candidate involves a certain picture of how conditional intentions are formed by rational agents, one based on the notion of conditional choice. There seems to be a blatant tension between the claim that our rational agent P forms the intention to do E if C should happen, and the fact that she recognizes that it would be against the balance of reasons for her to choose E if C should happen. This suggests the following necessary condition on the formation of conditional intentions:

(I) *A rational agent can only conditionally intend to do something X should a condition Y obtain, if in conditionally choosing what to do on the assumption that Y obtains, she determines that the choice of X is at least as well supported (given her presently held beliefs and desires) as its relevant, admissible alternatives.*[4]

The idea underlying (I) is simple. In order to form the intention to do X if Y should happen, an agent must be able to argue as follows: 'Suppose Y has happened. Then it is reasonable to do X', where the agent's reasons show that, given her beliefs and desires, the choice of X is at least well supported as its relevant, admissible alternatives. (If there are competing reasonable courses of action, the agent will somehow opt for one of them.) It follows that if doing X on the condition that Y obtains

[4] See P. Pettit and M. Smith, 'Backgrounding Desire', *Philosophical Review* 99 (1990), pp. 565–92, for the difference between an account on which belief and/or desire are *foregrounded* (in the sense that the agent reasons by focusing on the fact that these are her beliefs and desires) and an account on which they are *backgrounded* (in the sense that the agent reasons by focusing on the content of these beliefs and desires). I have in mind the backgrounding way of understanding the condition.

is seen as *unreasonable* because not as well supported as other options open to her, then she will not conditionally choose X, and so she will not form the conditional intention to do X if Y should obtain. The conclusion that our rational agent P will not form the conditional intention to do E if C should happen, even though she has the opportunity to do so, now follows immediately from (T1), (T3) and (I).

(I) provides us with what we might call a pure 'act-focused' necessary condition on the formation of conditional intentions. This condition is widely accepted, and it clearly played a role in Kavka's own argument for the paradoxical nature of deterrent intentions.[5] (I) leaves us with a prima facie puzzle, however. In implying that P, a rational agent, cannot form the intention to do E should C happen because P would not choose E conditional on C having happened, (I) leaves out of consideration any role that the conditional intention itself might play in an argument about whether to choose E. According to the model of rational choice underlying (I), the agent supposes that the condition applies and then decides – in the scope of her supposition – how to respond. But in supposing *only* that the condition applies, the agent forgets that if the formation of the intention was indeed successful then she should be supposing that the condition applies *in conjunction with the agent having issued a credible threat to do E should the condition apply*. To form the conditional intention in a way that takes account of all the relevant facts, the agent would have to consider the full context underlying her conditional choice, and that context should allow for her having formed the intention, if that is indeed what she ends up doing. The agent shouldn't use a procedure that renders (potential) relevant facts simply invisible. The omission is a significant one in the case of intentions like deterrent threats, since these have what Kavka called 'autonomous effects' – effects that are independent of the intended act's actually being performed.

This objection, which I'll call the 'intentional effects' objection, may strike one as peculiar. The obvious response is that if the agent *does* form the intention, then the full context underlying her conditional choice will include her having formed the intention, and if she does *not* form the intention then the full context will include her *not* having

[5] According to Kavka, '[i]t 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' ('Some Paradoxes of Deterrence', p. 292). See also Michael Bratman's account of the 'rationality of an agent for her deliberative intentions' in *Intention, Plans, and Practical Reason*. Bratman's ahistorical and historical principles both contain the condition that the agent in intending 'reasonably supposes that [the object of the intention] is at least as well supported by his reasons for action as its relevant, admissible alternatives' (pp. 84–5). Although Bratman doesn't discuss conditional intentions as such, there is every reason to suppose he would take them to fall under an appropriate extension of this condition.

formed the intention. What we need to know is which antecedent is true, given that she is a rational agent, and for that we seem to need an account of deliberative rationality that makes no allusion to the intention itself but only to the content of the course of action it conditionally prescribes. Anything else would be circular.

But such a response depends from the outset on the truth of the act-focused picture of the formation of conditional intentions. If a rational agent had access to a deliberative procedure that was able to incorporate the benefits of forming the intention independently of the benefits of the conditionally enjoined course of action, there would be no such circularity. In the next section, I shall explore a certain descendent of the conditional-choice account that doesn't fall prey to the 'intentional effects' objection. But first I want to consider another way of incorporating the benefits of forming the intention, due to David Gauthier.

What we might call 'intention-focused' accounts of conditional intention-formation emphasize deliberative procedures that focus on the benefits of forming the intention, and not just on the benefits (or absence thereof) of acting on the intention. Perhaps the most sophisticated intention-focused account on the market – an *impure* version, since it is partly act-focused – is David Gauthier's 'constrained maximizing' theory of choice, first developed in 'Assure and Threaten'. Like the straightforward maximizer (someone whose deliberative procedure Gauthier classes as self-defeating), Gauthier's constrained maximizer is concerned to further her ends, but her deliberative procedures differ from those of the straightforward maximizer in a crucial way. In considering a course of action, she considers whether it is

> conducive to [her] life going as well as possible, where a course of action is distinguished and demarcated by its intentional structure. One acts rationally in doing what, among those actions intentionally compatible with one's previous behavior, will lead to one's life going best, *provided one expects to do better than one would have done had one not performed any potentially intentionally restrictive acts that have proved relevant to one's choice*. ('Assure and Threaten', p. 717; my emphasis)

Consider 'potentially intentionally restrictive' acts such as forming, and then announcing, conditional intentions that are meant to assure or threaten (assurances and threats). In such cases, so Gauthier argues, it is the entire package – intention plus execution – that should be considered, since it is the entire package that can result in one's life going better than *not* performing the 'potentially intentionally restrictive' act of forming the conditional intention. Thus consider your assurance to a neighbour that you will help him harvest his crops next week if he helps you with your crops this week. It is rational to make such a commitment since this commitment is likely to result in your

life going as well as possible; better, certainly, than if you don't give him this assurance, since that would mean that you both lose out. But once the assurance is given, you should act on it since your life will go better acting on it than if you had not formed the commitment at all.

The threat case is different, however. Gauthier writes:

> One may offer as a reason for carrying out an assurance, that one's life will go better than if one had not made the assurance, but one cannot offer a parallel reason for carrying out a threat. Without such a reason, one would act irrationally in doing what one did not expect would thenceforth make one's life go best, and so one would act irrationally in carrying out the threat. And if one did not expect to have such a reason, one could not rationally do what one realized might intentionally restrict one to acts that would be irrational without it. As a rational agent I am able to offer sincere assurances, but it seems that I am unable to issue sincere threats. ('Assure and Threaten', p. 713)

Note that Gauthier is here denying the rationality of issuing threats when this is done on a case-by-case basis. He thinks that the same scepticism shouldn't extend to the case of an agent who has embedded her threat in a general *policy* of issuing and executing threats, for having the policy in place may help to make her life go better than if she had not formed such a policy (Gauthier mentions commercial enterprises in this connection).[6] But not even this manoeuvre, he thinks, can save the case of apocalyptic threats. Gauthier takes it that, at the point where such a threat has failed,

> she would expect her life to go less well, were she to enforce her threat, than it would have gone, had she not embarked on any policy of issuing and enforcing threats. And so she would not consider it rational to enforce her apocalyptic threat. ('Assure and Threaten', p. 719)

I am here only interested in attempts by rational agents to issue threats on an ad hoc, case-by-case basis (a one-off threat to deter a cheating partner from cheating, for example), so Gauthier's claim that rational agents can issue threats as a matter of policy is of no help to us. But is Gauthier right to think that ad hoc attempts of this kind must fail, even if attempts based on a general threat policy can succeed? In my view, no. Although Gauthier recommends an intention-focused deliberative procedure that is able to give due weight to some of the benefits of forming the intention, a close inspection of his argument shows that Gauthier's way of describing and motivating his account remains subject to the 'intentional effects' objection that we earlier lodged against the conditional-choice picture of the formation of conditional intentions.

[6] 'Assure and Threaten', sect. IX.

To see this, note that Gauthier accepts something like the following account:

(II) *A rational agent can only conditionally intend to do something X should a condition Y obtain, if she expects that, were she to form the conditional intention and then find that Y obtains, her life would go at least as well doing X as it would have done if she had not formed the intention in the first place.*

Now consider our earlier triad (T1)–(T3) again. Although I doubt that Gauthier accepts (T2), since he seems to draw a tight connection between an option's being rational for an agent and the agent's being rationally able to choose the option, it is evident that Gauthier accepts at least (T1) and (T3), subject only to the condition that in (T3) we interpret 'it would be against the balance of reasons for the agent to do E if C should occur' to mean 'the agent's life would go less well doing E if C should occur than it would have done if she had not formed the intention in the first place'. With (T3) so construed, it is clear that Gauthier thinks we can use (T1), (T3) and (II) to infer that a rational agent cannot sincerely issue the threat that she will do E if C should happen, even though it is an option open to her.

To assess Gauthier's agent-irrationalist argument, we must first disambiguate. The right-hand side of (II) can be taken either in an *engaged* or a *non-engaged* way. On the non-engaged construal, she only takes account of her present desires, not those she sees she might acquire were she to form the intention. On such a construal the agent expects that acting on her threat (were she to issue it) would make her life go worse than not having issued the threat. Now, Gauthier may well be right about the agent's expectation, but it is far from clear that this has much to do with whether the agent can, or should, form the intention. To draw any conclusions about rationality, a better perspective might be an *engaged* way of taking on board such scenarios, a way in which an agent considers such a scenario from the perspective of someone who has formed the intention (where having the intention in place is recognized as having significant benefits) and who then engages with the consequences of having formed it. And so construed, it is not in the least clear that the agent who threatens to leave her cheating partner, for example, would argue the way Gauthier presents her as arguing. If, on supposing that she has issued her threat, the agent contemplates what she would do should the threat fail, she may well find herself thinking that she would follow through with the sanction, that her partner's disregard of her threat, despite her clear demonstration of how much she cares about his heeding it, would be an insult that she would not care to live with. And from that perspective, Gauthier's question of whether her life would go worse acting on the intention than if she had not formed the intention at all might well

strike her as quite beside the point. Note that she wouldn't have this kind of reaction unless she contemplated matters from the perspective of having issued the threat; prior to issuing the threat, the agent would prefer her relationship not to be compromised, even with her partner's continuing his cheating.

As far as I can see, Gauthier would deny that such an engaged perspective is of any relevance when the agent comes to decide whether she should issue her threat. For him, the agent's supposition that she has issued the threat is a supposition that has the agent acting irrationally. Gauthier has in mind a *non*-engaged way of understanding what the agent expects should she have formed the intention and then find the condition of the intention satisfied – a way that focuses on her present desires and not on those she might acquire were she to form the intention. The agent is asked to compare acting on the intention (with the intention in place) with not acting on it (because the intention is not in place), and any such comparison can only be valid if it rests on desires and commitments that stay constant. As a result, (II) has rendered any changes in perspective that might arise from the agent's actually forming the intention invisible to the deliberative procedure used to rationalize forming the intention. Hence Gauthier's argument for the irrationality of agents who form such intentions remains subject to the very objection – the 'intentional effects' objection – that I earlier raised against the argument for agent-irrationalism based on the conditional-choice account of the formation of such intentions.

3. DETERRENT THREATS AND THEIR IMAGINATIVE PREFORMATION

Like the conditional-choicer before him, Gauthier may think that the intentional effects objection is simply misplaced because it generates a vicious circle. How, when the aim is to give an account of how a rational agent might form a conditional intention, can it possibly make sense to include in one's account the agent's appeal to the intention having been formed, and the way this impacts on the way she might then evaluate her executing the intention? (Why not include the intention's having been *rejected*?)

But such a response is based on a misunderstanding. Consider the following variation on the standard conditional-choice model. First of all, an agent front-loads the contemplated intention 'I intend to do X if Y' into her deliberative procedure – the intention doesn't simply appear as the end-result. Second, the agent bootstraps her way into forming the intention if she prefers the imagined consequences of having the intention in place, including the situation in which Y occurs and she must apply X because of the intention being in place, to the imagined

consequences of the intention not being in place.[7] Note that in the case of conditional intentions that lack autonomous effects, this process reduces to something like the usual conditional-choice account. This is because those are cases where placing oneself in the full imaginative context in which the intention 'I intend to do X if Y' has been formed (in order to see if one prefers this situation to one involving other courses of action one might take) doesn't require one to assess the impact of the intention itself. All the agent has to do is to assess the probability-weighted benefits of doing X if Y should happen (benefits that are independent of the intention being in place) and then compare these to the benefits of alternative options, just as the agent would on the conditional-choice account.

But matters are different if intentions have autonomous effects. Take deterrent intentions again. On the alternative picture I am recommending, the agent places herself in the full imaginative context in which she has successfully formed the conditional intention to apply sanction E if condition C should occur, with a view to seeing whether she can live with the intention, useful as it is and given her various beliefs, desires and commitments. (The notion of 'living' with the intention is supposed to be neutral among competing accounts of the notion of a life going well, or at least as well as possible giving one's goals and commitments.) Once she sees that she can live with it, and sees that she prefers living with it to living without it, she is at the point where she can actually form the intention.

The crucial point of difference with ordinary intentions is that the intention's autonomous effects can make a significant difference to her assessment of whether she can live with the intention. They can do this in two ways. First, and most obviously, the usefulness of the intention, based on the usefulness of its deterrent effects, can make it worthwhile for the agent to *try* to find a place for the intention among her other commitments. Second, the fact that the agent expects the intention to be useful can indirectly create a new reason for the agent to act on the intention should deterrence fail (action that would be against the wishes of those she is trying to deter, thereby reinforcing the effectiveness and hence usefulness of the intention), which in turn can make it easier for the agent to live with the intention. To see this second point, take the case of the deceiving partner again. Given the nature of the sanction involved in the threat, when the agent contemplates the situation of her partner deceiving her despite the threat, she may

[7] In 'Fear and Integrity', *Canadian Journal of Philosophy* 38 (2008), pp. 31–49, I suggest how such an account might be extended to *unconditional* (future-directed) intentions, including the kind of problematic unconditional intentions that feature in Kavka's well-known Toxin Puzzle ('The Toxin Puzzle', *Analysis* 43 (1983), pp. 33–6).

well 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 somewhat easier: 'Wait, I am forgetting that he continued his deception after all *I* did to show him how much I cared about his not doing it. I even threatened to leave him if he continued, and he knew how awful it would be for me to leave him.' After repeated contemplation of the imagined scenario, including repeated contemplation of the awfulness of her partner's deception after all she has done by way of her threat to warn him off such 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. If so, she will have come to the point where she is finally able to form the intention.

I'll call the imaginative process by which the agent considers the impact of having formed the intention, in order to decide whether she can live with the intention, the *imaginative preformation of the intention*. (Think of it as a simulation experiment, one that succeeds to the extent that the agent finds herself imaginatively able to live stably with having formed the intention.) On this model, *actually* forming the intention is the result of the successful imaginative preformation of the intention, with the agent bootstrapping her way into forming the intention when her imaginative preformation of the intention has been successful so that she is confident that her actual behaviour will match her imaginative preformation. That, I am proposing, is how it is done, or how it *can* be done. But this cannot, 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 against her best interests? How does enriching the context help?

But this remark forgets the distinctive manner in which the agent is able to engage imaginatively with the scenario of the partner's continued cheating once the intention is imagined as being in place. For the agent is now not just asking 'What should I do if the threat is in place and my partner nonetheless continues his cheating?' but rather 'What should I do, now that my partner has continued his cheating despite my threat to leave him if he continues?' The difference between these two questions reflects the difference between what I earlier called a non-engaged and an engaged perspective. When the agent faces the question as formulated in the second way, her imaginative contemplation of her partner's continuing to deceive her, despite her intention being in force, is likely to engage her emotionally: she will feel anger and resentment in a way that makes all the difference to her deciding what to do in the scope of her imagining, and hence all the difference to whether she can bring off her imaginative preformation of the intention.

Note that this is an entirely *natural* way of involving emotions in our rational lives. If a rational agent considers courses of action that she or others might undertake she is likely to feel pleased or happy at the prospect of the satisfaction of any desires she has (indeed, this emotional accompaniment may be an important part of being motivated to undertake a course of action). By the same token, she is not likely to take a neutral stance towards a contrary action on the part of another agent that debases these desires.[8] Hence emotions like resentment and anger may, in a sense, be required emotions for rational agents if rational agents are to identify in the right sort of way with their desires.

Still, merely noting the case for emotional engagement of this kind doesn't greatly help the case for a preformation 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 rationally to contemplate leaving her partner in the context of her imaginative engagement with the scenario of her partner's deception. 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 *re*action to her partner's deception. That would still leave the agent susceptible to the charge of irrationality ('What you did was to lash out in anger, and you only hurt yourself that way'). Emotions like anger play a far more nuanced and complex role in this kind of situation. As Patricia Greenspan has pointed out, they can embody a crucial shift in the agent's evaluative perspective.[9] Prior to, and apart from, her issuing the threat, the agent's interests were focussed on her well-being, something that she

[8] For a general argument for the central importance of the emotions in our rational lives, see Michael Stocker (with Elizabeth Hegeman), *Valuing Emotions* (Cambridge, 1996). The motivational importance of emotions in decision-making is also underscored in important empirical work done by Antonio Damasio and his co-workers. See, for example, Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (New York, 1994) and *The Feeling of What Happens* (New York, 1999); and Bechara et al., 'Insensitivity to Future Consequences Following Damage to Human Prefrontal Cortex', *Cognition* 50 (1994), pp. 7–15.

[9] I am here indebted to Greenspan's 'Emotional Strategies and Rationality', *Ethics* 110 (2000), pp. 469–87, although my emphasis is somewhat different. I have been concerned with the way the intention might be formed, whereas Greenspan seems more concerned with how the agent might bring herself to act on her threat through a rational shift in evaluative perspective. See also Bennett Helm, *Emotional Reason* (Cambridge, 2001), which attempts to bridge the cognitive-conative divide by, in part, construing emotions as themselves evaluative in nature.

saw as likely to be compromised by her leaving. Still, she realized 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 deception 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 as an agent who refuses to acquiesce in such humiliation – and this new end, which is as emotion-involving an end as her dedication to her partner, is one that our agent gives expression to if she leaves (or rather, as she *imagines* herself leaving; remember that this is all occurring in the context of her imaginative preformation of her threat). 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.[10]

That, I suggest, is how one should argue for the claim that the way in which the agent sees herself as resolutely prepared to leave in this imaginative preformation of her threat fully accords with her status as rational agent.[11] Once she has bootstrapped herself into actually forming the intention, she can be sincere in declaring her threat to her partner, who in turn will understand that she means what she says since he believes her to be rational and therefore to know her own mind.

4. THREATS AND QUASI-THREATS

I have defended the possibility of agent-rationalism for a class of threats – non-apocalyptic threats – where others defend agent-irrationalism. But what about apocalyptic threats, in particular nuclear deterrent threats? This kind of case is more tricky. The agent's preformation of the intention will focus on the deep anger that the agent feels on contemplating an enemy's striking in the face of the agent's efforts to deter the enemy from striking (efforts that include

[10] For the notion of expressive reasons for action, see, for example, Joseph Raz, *The Authority of Law: Essays on Law and Morality* (Oxford: Clarendon, 1979), pp. 253–8. What is important in the present use of this idea is that the expressive reason for acting depends for its existence and force on the formation of the intention. The reason was not available for incorporating into intention-independent deliberation about whether to perform the act.

[11] For a very different account of such threats and their rationality, see Robert Frank's *Passions within Reason: The Strategic Role of the Emotions* (New York, 1988).

threatening retaliation, where retaliation is against the agent's own interests, not just the interests of the enemy).[12] It is certainly harder to see how a rational agent can live with an extreme apocalyptic intention of this kind. An 'intention' on the part of such an agent to inflict great and useless harm if attacked sounds more like an intention made by a wholly different agent, an agent with vengeful, even suicidal, proclivities and desires that are simply incompatible with certain of her core desires. Genuinely intending to do a certain thing must surely involve goals that are appropriately integrated with goals and commitments that one presently identifies with, not goals that should strike the agent as intolerable from the perspective of presently held goals and commitments. 'Intentions' of the latter kind sound more like predictions about oneself.[13] Take the case of Dr Robert Banner aka the Hulk, from the Marvel comic series, TV series, and 2003 movie [*The Incredible*] *Hulk*. Because of the effects of irradiation by gamma rays, Banner becomes the Hulk when he is provoked: someone who is super-strong, but seething with rage. Suppose Banner, in trying to deter Y from happening, announces that he 'intends' to inflict great harm on P if Y should happen, on the grounds that he knows he will become the Hulk if Y happens, and so will inflict great harm on P. Banner's claim that this is what he *intends* to do will surely strike most of us as contentious; the action he is talking about seems more appropriately described as the action of his alter-ego the Hulk.

Given what is contemplated if a nuclear threat fails, the same sort of thing might be said about nuclear intentions. No fully rational and moral agent could intend to impose such an awful retaliatory sanction; an intention of this kind requires a degree of corruption that shows such an agent to be less than fully rational and/or moral.[14] But the same cannot be said about the sort of non-apocalyptic threat discussed earlier, since this kind of case does not involve the acquiring of desires that are incompatible with desires and commitments that the agent presently identifies with. Although the agent is dedicated

[12] This description will be contentious if the envisaged scenario is a survivable nuclear war (a *near*-apocalyptic scenario). For the agent issuing the threat may then have as one of her rational and moral goals the conditional goal of ensuring that the attacker doesn't survive intact, to avoid the agent's nation being placed in bondage to a wholly alien way of life (cf. Greenspan, 'Emotional Strategies and Rationality', p. 484 n. 24).

[13] For an excellent discussion of the distinction, see Gilbert Harman, *Reasoning, Meaning and Mind* (Oxford, 1999), ch. 2.

[14] David Lewis once argued that real world deterrers (at least those in the U.S.) were a 'strange' mixture of good and evil and of the rational and irrational. See his 'Devil's Bargains and the Real World', *The Security Gamble: Deterrence Dilemmas in the Nuclear Age*, ed. David MacLean (New York, 1984). In conversation, he took this to show that an agent-irrationalist and agent-immoralist view of nuclear threats sets the standards for rationality and morality too high. Lewis's view suggests another way of defending agent-rationalism, although not one I am inclined to accept.

to her partner, she does not, prior to forming the intention, have the unconditional desire that she stay with her partner; what she desires is that she stay with her partner, all things being equal. When, in the course of the preformation of the intention, the agent sees herself as prepared to leave after her partner's continued cheating, this is the result of all things *not* remaining equal, and is the natural outgrowth of her present desires – her preparedness falls out of both her dedication to her partner and her own sense of dignity.

(Note that matters may well be different if the agent puts a very high value on co-dependency and if her central goals include protecting the security that her relationship with her partner brings. For then we may no longer have compatibility between her present desires and a desire to leave in response to continued cheating. So minded, our agent may well not be able to form the sincere intention to leave her partner, since she doesn't see herself as someone who would be prepared to leave under these circumstances. This should remind us that the present way of classifying cases simply locates apocalyptic threats at the extreme end of a continuum, with certain non-apocalyptic threats being closer to the extreme than others. Agent-irrationalism seems the appropriate response to both types of cases.)

The approach I am here adopting yields a final benefit. Consider otherwise rational agents who know enough about themselves (perhaps through trying a preformation of the relevant deterrent intention) to know that their own settled desires would, or might well, undergo a more or less radical shift under certain extreme forms of provocation – provocation that includes a calculated disregard of attempts at deterrence. At the point where emotion takes over and their actions subvert their own strong, settled preferences, these agents would not be acting rationally by their own lights. But simply being *liable* to show such behaviour in extreme circumstances, and being aware of being liable, is surely not enough to make such agents irrational.[15] Take the agent who can't imagine leaving her cheating partner. She may still have deterrent success in warning her partner that she will, or might well, leave him if he continues his cheating, for both the agent and her partner may understand that, even though she cannot find it in herself to form the conditional intention to leave him, the agent may well become irrational by her own lights, and leave him.[16] Such an agent

[15] Just as 'X is fragile' does not mean 'X will break when struck, no matter what the possible circumstances', so 'P is rational' does not mean 'X chooses rationally, no matter what the possible circumstances'. Knowing which possible circumstances are relevant is, of course, a difficult matter.

[16] Even if the agent does end up leaving her partner, this is still not enough to show that the agent is irrational. It may just signal that the agent's preferences have undergone a sharp, unanticipated shift. She may now see her leaving as something she wants to do to

has not deterred by means of a genuine deterrent intention or threat, but only by means of what we might call a 'quasi-threat' – a sincere utterance that announces a conditional, and emotionally intelligible, course of action, is designed to deter, but falls short of stating a genuine deterrent intention or threat. (Note that this seems to rule out the sort of 'threat' that Dr Banner, anticipating his transformation into the Hulk, might utter. Lacking the element of emotional intelligibility, his 'threats' are just reliable predictions.)

Apocalyptic 'threats' such as those involved in a classic policy of nuclear deterrence are also best seen as quasi-threats. Assuming that the agent's enemy believes the agent to be fully rational and moral, and so believes that the agent can't *intend* to impose the awful sanction that figures in the agent's threat-like utterance, the enemy may nonetheless still be deterred by the agent's utterance from striking at the agent. For the agent's utterance may be a believable quasi-threat. Both parties may realize that the agent's anger and grief when confronted by the awfulness of a strike that the agent tried so hard to prevent might well make for a radical but emotionally explicable shift in perspective, rendering the agent prepared to undertake actions that, seen from the perspective of the agent's current goals and commitments, must count as deeply irrational and immoral. As before, I doubt that there is anything in the concepts of rationality and morality that prevents a fully rational and moral agent from being *liable* to such drastic changes under certain unusual hypothetical circumstances – especially if the agent is not only aware that she is liable, but is also willing to use her awareness for deterrent purposes to ensure that the circumstances remain hypothetical. If the enemy understands all this, he will be prepared to take the agent's threat-like utterance entirely seriously, without treating it as a serious statement of intent.[17]

f.kroon@auckland.ac.nz

show her disaffection and anger – her action may thus count as expressively rational. By contrast, if the agent leaves her partner but then thoroughly *regrets* taking this course of action because she continues to identify with her original desires, then we would say that she acted irrationally.

[17] I am grateful for helpful critical comments from many colleagues, especially David Braddon-Mitchell, Stewart Candlish, Richard L. Epstein, David Lumsden and Jonathan McKeown-Green.