NOÛS 37:1 (2003) 25–51

*Trust and Diachronic Agency**

EDWARD S. HINCHMAN
Claremont McKenna College

> It is but giving your little private convulsive self a rest . . .
> —William James, *The Varieties of Religious Experience*

On your way out, you're standing, key in hand, at the door. You've just locked that door. Or so you think. This is what memory clearly tells you. But have you really? You give the handle a turn. Yes, locked. The thought occurs to you, however, that double-checking hasn't improved your epistemic predicament. If it made sense to doubt then, it makes equal sense to doubt all over again. A feel for the absurdity of triple-checking suffices to send you on your way.

At the produce stand you pause. You'd intended to buy eggplant, not okra, but now you're unsure. Why not choose okra? You feel certain that nothing relevant to your deliberation has changed. Yet you see no harm in quickly rehearsing the considerations pro and con. So you redeliberate and affirm your intention to buy eggplant. The thought occurs to you, however, that redeliberating hasn't changed your practical-reflective predicament. If it made sense to redeliberate then, it makes equal sense to redeliberate all over again.

These failures of self-trust have a peculiar feature. Lacking a basis in any consideration apart from the thought that it couldn't hurt to double-check, there is no way to redress them rationally. A baseless worry admits of no rational redress—you just have to wait for it to subside. If you'd not been paying attention as you locked the door, double-checking would give you a memory belief that would assuage the temptation to triple-check. But you already have that memory belief. If at the produce stand you believed that you'd overlooked some matter relevant to your decision to buy eggplant, you could reopen your deliberation with some idea how to get it closed

again. But you don't believe you've overlooked anything. Still, in each case you do feel the worry.

Is such a worry irrational? Your memory *might conceivably* be inaccurate; it *might conceivably* be the case that you misdeliberated. How can it be irrational to check, if checking is easy and involves no prohibitive cost?[1] Yet if you don't trust yourself at least to the extent of being disposed to accept the deliverances of memory or deliberation when you have no basis for questioning them, the life of reason will be inaccessible to you. Have you not *therefore* a right to trust yourself? The matter is not so simple. If the attitude in question is *trust*, then an entitlement to it must be earned. Trust is the appropriate response not to your desire or need for the succor you'll derive from a life of reason but to your ongoing status as trustworthy in living that life. If you are not trustworthy, or if you have good evidence that you are not, then it seems you *should not* trust yourself. If you are trustworthy then it seems, by parallel reasoning, you should. These "should"s seem to express norms of rationality.

In this paper I let pass your doorside predicament and treat only your predicament at the produce stand. I focus exclusively on practical-reflective pathologies of trust. I'll argue that the diachronic exercise of practical reason in general presupposes an absence of the sort of pathology you manifest at the produce stand. The diachronic exercise of practical reason is in that respect based on trust—typically on self-trust, but when self-trust fails, I'll argue, it will have to be based on trust in others.[2]

It is revealing that self-trust runs the same risk of pathology as trust in others. In each case, I'll argue, reasonable trust presupposes trustworthiness in the trusted. But reasonable mistrust equally presupposes untrustworthiness in the mistrusted, or evidence thereof. That's why the exercise of practical reason is incompatible with baselessly failing to trust your earlier intention-forming self. Yet earlier selves are not inherently more trustworthy than other people. If it is ever reasonable to act on the basis of self-trust, it must sometimes be reasonable to act on the basis of trust in others.

I

Say you've decided to buy eggplant, not okra, but there's still time to reconsider. Under what conditions would reconsidering be reasonable? We can distinguish at least four kinds of condition. It would be reasonable to reconsider

(a) if you realized that you'd overlooked some matter relevant to your deliberation—for example, that one of your guests is allergic to eggplant;
(b) if you realized that you'd based your decision on faulty reasoning—for example, undercounting the number of guests who might prefer

okra or inaccurately weighing the force of that consideration in light of your desire not to offend them;

(c) if you realized that your desires or values had changed in relevant ways since you made the decision—for example, the passion to take revenge on eggplant-loathers that you now see controlled your deliberation;

(d) if you realized that you'd based your decision on a false belief concerning the circumstance in which you would act on it—the belief, for example, that by the time your busy schedule let you squeeze in a trip to the market there'd be any healthy-looking eggplants left.

Why is it unreasonable to reconsider merely in order to assuage a nagging doubt that you misdeliberated? Well, it might not be unreasonable. How confident are you that none of these conditions holds? If you are not confident, by all means think again. If you really aren't confident you locked that door, by all means double-check. When you lack the short-term memory belief, you'd be irrational not to double-check. But pathological door double-checkers don't lack this belief. They remember locking the door yet mistrust their memory.[3] In the practical case matters are somewhat more complicated. What you mistrust is not your memory but your decision: you mistrust the earlier self of yours that formed the intention to buy eggplant. But your mistrust is just as pathological if you cannot cite—at least gesture at—some respect in which the decision is likely to prove ill taken. If you cannot at least gesture at some respect in which your memory belief is likely to prove ill grounded, your mistrust of it is pathological. If you cannot at least gesture at some respect in which your decision is likely to prove ill taken, your mistrust of it is likewise pathological. (I'll explain the nature of the pathology in section III.)

Why not view such double-checkers and redeliberators as merely mistakenly vigilant beyond the point up to which vigilance is rational? You need to move on from that doorstep; you need to buy something for dinner. You can make a pragmatic argument against continuing to check and re-check, against continuing to redeliberate: there's a cost, after all, to not getting on with your life. Why speak of pathologies when we can speak of errors?

Consider more closely your tactic in redeliberating. One problem with viewing this as a mere deliberative error is that we can imagine cases in which the costs of redeliberating will *not* tip the scales against the repeated perceived benefits of repeatedly scratching the itch of self-doubt. What if you have all afternoon to get your shopping done and really don't mind spending half of it at the produce stand? A purely cost-benefit analysis may recommend redeliberating far beyond any point at which we could recognize you as an agent pursuing her or his ends.[4] Is the problem merely that you suffer from this itch? Norms of agency require nothing as strong as the absence of pointless self-doubt. They require merely that we not *give in* to

pointless self-doubt. Your problem as your self-orientation strands you there lies not in what you feel, nor even in your assessment of the importance of what you feel. (The feeling may be powerful, and redeliberating may bring welcome—albeit short-lived—relief.) Your problem lies in the fact that you've let—that you keep letting—the issue of what to do in light of what you feel arise in the first place. As we may imagine your preferences with the afternoon stretching out before you, this is not an issue that you *can* resolve through deliberative means.[5] Until the costs and benefits change, you can resolve it only if you stop deliberating. You must stop deliberating, but not because you deliberatively conclude that further deliberation would be pointless. Deliberation itself cannot resolve this deliberative quandary. You must simply *stop*.

II

How shall we characterize the basis of this rational requirement? In theoretical terms as the thesis that it is of the very nature of personal agency that you have a reason to avoid such pathology? Or in practical terms as the thesis that you thereby most effectively pursue your rational aim? While I'll not attempt to resolve the dispute between proponents of these theses, I'll use the distinction between them to structure my attempt to explain the nature of the pathology.

J. David Velleman argues for an instance of the first thesis by characterizing intentions as self-regarding expectations: intending to φ at t is expecting that you will φ at t as a result of this very expectation.[6] Velleman furthermore argues that it is constitutive of personal agency that an agent is motivated to know what he is doing. As he puts it, "what you should add to subjects of motivation, in order to create agents, is the higher-order aim of knowing what they're doing."[7] Thus it is constitutive of your status as agent that you have a *pro tanto* reason to follow through on any intention that you form. The reason derives, as follows, from your possession of the higher-order aim. On the hypothesis that φing is what you expect to be doing at t, if when t rolls around you aren't φing, you won't, by hypothesis, know what you're doing. But that would violate the higher-order aim constitutive of your agency. So you have a reason to follow through on your intention and φ at t. While following through on an intention does therefore promote an end, the end is not one that you could abandon while remaining an agent. Your reason to follow through on your intention would have this theoretical basis: you thereby satisfy an aim constitutive of personal agency.

Now I don't see how Velleman's can be the whole story of personal agency—or even, as he presents it, of autonomy. A subject of motivation with the higher-order aim of knowing what he's doing is not yet, as I see it, a subject capable of governing himself: he is not yet capable of autonomy. Others can govern you without your trusting them, but you cannot govern

yourself without self-trust. And self-governance cannot be rational, and thus cannot yield autonomy in a fully normative sense, unless the governing self is *worthy* of the governed's trust. Even if Velleman is right that intentions are self-regarding expectations, an autonomous agent governs himself by acting only on those self-regarding expectations that are worthy of his trust. There is no reason to follow through on an intention formed by an untrustworthy earlier self. How could the fact that an untrustworthy earlier self of yours formed an intention give you a reason to fulfill it? It seems no more rational to do the bidding of an untrustworthy self than it is to do the bidding of an untrustworthy advisor.[8]

Velleman presents his view in criticism of David Gauthier's conception of constrained maximization.[9] For Gauthier, the rational bearing of your intention to φ at t on your conduct at t has not a theoretical but a practical rationale: it would derive from the fact that you thereby most effectively pursue your rational aim (we can remain uncommitted as to exactly what that is). But do you not sometimes most effectively pursue your rational aim by forming an intention and then changing your mind? You can sometimes benefit simply by forming an intention: consider how you might benefit by assuring a partner you will follow through on a cooperative agreement. Why not form the intention, secure the benefit, and then see whether it makes sense to follow through? The problem is that if you expect that you'll change your mind, you cannot form the intention in the first place. But this prevents you from getting the benefit.

Gauthier's solution is elegant: he argues that the concept of rationality applies primarily not to actions as such but to the dispositions that motivate them. A disposition is rational, he argues, if it best promotes your rational aim. We must therefore ask: does the sequence {intending to φ at t, φing at t} promote your rational aim better than other sequences that could be motivated by other dispositions (including of course the sequence {not intending to φ, not φing})? If it does, then φing at t is rational, and its status as rational derives from its relation to your intention to φ in the following sense: both are motivated by the disposition by which it is rational to be motivated in the circumstances. It is not that you would have an intention-based reason to φ at t. Rather, your φing at t would be rational as a result of being motivated by the same rational disposition that motivated your formation of the intention to φ at t.[10] Your intention would have rational bearing on your future conduct in the sense that each would be motivated by the disposition that best promotes your rational aim.[11]

Yet how could it be rational for some disposition to move you at t (the time of action) merely because it was rational for that disposition to move you at *t-n* (the time of decision)? Even if it is rational at *t-n* to be moved by a disposition that will move you to φ at t, how does that show that it is rational at t to be moved by that disposition? Imagine that at *t-n*

you faced a range of options concerning what to do at t and decided to φ. Now t has almost arrived and you wonder whether to follow through on that decision. Gauthier contends that the question you should ask yourself is whether the sequence of acts {intending to φ, φing} better promotes your rational aim than any contending sequence. But which sequences count as contenders? Is the sequence {intending to φ, not φing} a contender? Here Gauthier appeals to the idea that only certain sequences can be motivated by coherent dispositions and claims that {intending to φ, not φing} cannot be so motivated.[12] But what of your disposition to reconsider your intentions when you suspect you may thereby do better? Whatever else we may say of that disposition, it hardly seems incoherent.[13]

Velleman and Gauthier of course have replies to my challenges,[14] but since my aim in this paper is not primarily critical, I'll move on without pausing to consider them. I aim instead to use Velleman's and Gauthier's positions to frame an issue on which I'll offer a contribution of my own. Velleman is right that an agent forming an intention aims thereby to give his future self a reason to follow through. And Gauthier is right that the rational stance of an agent wondering whether to follow through on an intention looks back and assesses its formation. We need to preserve both insights and use each to supplement the other. We need to view an intention-forming self as aiming thereby to give its future self a reason to follow through. And we need to view intention-executing selves as having that reason only by exercising the capacity *not* to follow through when that earlier self is untrustworthy in one of the ways characterized by conditions (a) through (d).

Is the basis of your reason to follow through on your intention theoretical or practical then? The dichotomy is specious. Your predicament at the end of section I is pathological because, as I'll explain, you are not appropriately responsive to the mere fact that you have formed an intention. This sounds like a theoretical basis. Where following through on an intention will not promote your rational aim because you satisfy one of (a) through (d), however, the appropriate response is to redeliberate; therefore, if you have a reason to follow through without redeliberating it must be because you satisfy none of those conditions and following through thereby promotes your rational aim. This sounds like a practical basis. As I'll explain, it is constitutive of your having an intention that it gives you a reason to follow through on it unless certain defeating conditions are met. The defeating conditions concern its tendency to promote your rational aim. So the basis of the reason is theoretical by virtue of being in this way practical.

III

Now to my positive account of diachronic agency. Let $S_{F/A}$ be a subject separated into F, the earlier self that forms and then continues to have an

intention, and A, the later self that acts on it. I begin by proposing this analysis of intending:

(I) $S_{F/A}$ intends to φ at t iff F expects and desires that A will at t have a preemptive reason to φ at that time simply through A's memory that F desires that A will at t have a preemptive reason to φ at that time.[15]

The analysis has a quasi-Gricean structure: the first party aims to give the second party a reason simply through the latter's understanding of that aim. That both parties are aspects or elements of a single subject merely shows that the relation is a self-relation.

How could F's expectation be true and F's desire satisfied? How, in other words, could A come to have a reason to φ simply through A's memory that F desires that A have a reason to φ? The answer is straightforward: by trusting F. How could that give A a *reason*? It could give A a reason if F is trustworthy on the matter at hand and there is no good evidence available to A to the contrary. We might conceptualize these as defeating conditions on the presumptive reasonability—and therefore reason-givingness—of self-trust. We might say that self-trust is presumptively reasonable, and therefore reason-giving, unless (i) your intention-forming self is untrustworthy on the matter at hand or (ii) there is good evidence available to you that it is untrustworthy on the matter at hand.[16] The first is an 'externalist' defeating condition on the reasonability of trust; the second, an 'internalist' defeating condition.[17]

I criticized Velleman's account for failing to explain how merely forming an intention could give you a reason to follow through on it. I have now in effect located the basis of this reason in the trustworthiness of the intention-forming self. This marks a parallel with interpersonal cases of trust. In each, (a) reasonable trust presupposes trustworthiness in the trusted and (b) reasonable mistrust presupposes untrustworthiness in the mistrusted, or evidence thereof. Let me offer some remarks in clarification and defense of these theses.[18]

Begin with (a). One complication is that bestowing trust can itself make a person trustworthy: perhaps your trust in someone, in making her trustworthy, will causally determine its own status as reasonable. We may concede that trying to reform this untrustworthy person was reasonable, but in order to attempt that reasonable thing you had to do something unreasonable considered in its own right; namely, trust someone not (yet) trustworthy. Considered in its own right, trust would be unreasonable unless and until it succeeded in making the trusted trustworthy—though it may of course figure as a means toward this reasonable end. Perhaps this is a case of reasonable unreasonability.[19] Yet whatever we say about this kind of case, it is crucially unlike the case of *self*-trust. There is no prospect

of reforming your untrustworthy earlier self by following through on its intention. That self, after all, no longer exists.

In putting forth not positive but defeating conditions for the reasonability of trust—conditions whose *non*-satisfaction is necessary and sufficient for trust to be reasonable—we embrace the everyday assumption that trust can be reasonable even when there is no evidence available that the trusted is worthy of it. We usually do not assess our intention-forming earlier selves for trustworthiness before following through (though sometimes we do), and we often do not assess advisors before following their advice. This in itself is not unreasonable. But if we discover that the earlier self or the advisor was untrustworthy, or that there was evidence of untrustworthiness available, we view the presumption of reasonability as defeated.[20]

Now turn to (b), the claim that reasonable mistrust presupposes untrustworthiness in the mistrusted or evidence thereof. By "X mistrusts Y" I do not mean merely *X does not trust Y*. It is obviously not unreasonable not to trust someone you've never met, even though she is trustworthy and you've no evidence to the contrary. If you've never met her, you've had no opportunity to trust her. Moreover, by "X mistrusts Y" I do not mean merely *X has an opportunity to trust Y but does not*. It is not unreasonable not to trust a stranger whom you've overheard give someone else a betting tip, though she's trustworthy and you've no evidence to the contrary. But it would, I think, be unreasonable not to trust this stranger if she gave *you* the tip in response to your request for advice. In requesting the advice you create a context in which it is appropriate for her to invite you to trust her judgment on the matter. If she is trustworthy on the matter and you've no evidence to the contrary, then refusing her invitation is unreasonable. By "refusing" her invitation I mean refusing to treat her intervention as giving you any reason—even an overridden *pro tanto* reason—to bet as she advises. (You can of course trust someone's advice without following it: trust is not in general the same as compliance.) By "X mistrusts Y," then, I mean *X refuses Y's appropriately tendered invitation to trust*. I claim that it is unreasonable to refuse someone's appropriately tendered invitation to trust unless either (i) the person is untrustworthy or (ii) there is good evidence available to you that the person is untrustworthy.

Why is it unreasonable to mistrust when neither of these conditions is satisfied? We can explain its unreasonability by viewing it as a conversation stopper. Imagine that it's contextually appropriate for your interlocutor to offer you advice on some matter, that she does so, and that you do not believe that you thereby acquire a *pro tanto* reason to comply with it. If you can cite (or at least gesture at) no consideration that would make reasonable this refusal of her invitation to trust, the refusal seems unreasonable. I do not claim, of course, that it would be unreasonable to refuse to cite that consideration. That's an entirely different matter: perhaps citing it would

give her a context in which to vindicate her trustworthiness; perhaps it would simply enrage her. If you deem it best not to cite the consideration and this has the effect of stopping the conversation, then *you* have stopped the conversation. If, on the other hand, there is no such consideration known to you, then your *mistrust* has stopped the conversation—in the way that any unreasonable act would. It isn't merely that your interlocutor won't know what to say next. *You* don't know what to say next, since there is nothing by your own lights that would rationalize your act in a way that would permit the conversation to go forward (without changing the subject). Of course, you retain your power of speech. But you do not retain the ability to make sense of your conduct—even to yourself. This deficit manifests itself most saliently in the fact that you'll be unable to offer a sincere and non-self-deceived defense of your conduct to one in position to take offense at it. But you'll be equally unable to pursue a conversation with yourself on the matter. (Of course, you could attempt to *diagnose* your mistrust, but that would change the subject. The conversation I have in mind would involve criicism and defense of the mistrust itself, not merely inquiry into its causes.) In short, the mistrust is unreasonable because it cannot be reasoned.

I hope this sketch of a framework for comparing the intrapersonal case with the interpersonal suggests an account of the pathology inherent in baselessly refusing to follow through on an intention. (I) models intending on essentially interpersonal speech acts such as telling and advising (as opposed to stating or asserting, which may lack an audience). Baselessly refusing to follow through on an intention is pathological in a way that we can model on the pathology in baselessly refusing an interlocutor's appropriately tendered invitation to trust. In each case, we can view the one performing the act—intending or advising—as aiming to give you a reason simply through your recognition of this very aim.[21]

IV

If neither defeating condition is satisfied, and if you do trust your earlier self now that the time has come to act, then (I) shows how you have a special kind of reason to execute the intention. Your reason to execute the intention isn't on a par with the reasons that weighed in your deliberation. Those reasons guided your intention-formation. What you now need is a reason that will guide your intention-execution. You execute an intention not by redeliberating—not by weighing or reweighing considerations whether to φ—but by letting your conduct be guided by the reason *not* to redeliberate that you act on when you execute an intention. This is what I mean when I say that a reason transmitted through the quasi-Gricean mechanism is "preemptive." It is a reason merely to follow through on an intention you've already formed. It is not a reason specifically to perform that action. If you

redeliberate, you cannot weigh this reason alongside the reasons pro and con performing the action, or use it to reaffirm your intention. It is not a reason that can be weighed in deliberation at all. It functions at a different level of your practical reflective life.

This level lies beneath that of deliberation. Following through on an intention in the spirit of self-trust does not involve deliberation whether to perform the action you intend to perform, whether to follow through on the intention, or whether to redeliberate. What, if you cannot weigh it in deliberation, does this reason *do* for you? It puts your deliberations in rational touch with your actions. Such a reason is not a dim echo of the reasons you weigh in your deliberations. Such a reason operates at a deeper level, that of trust.

Here we find an important disanalogy between intra- and interpersonal cases of trust. I remarked in the previous section that trust is not in general the same as compliance: you can trust a piece of advice by giving it weight in your deliberations without in the end complying with it. Of course, that distinction does not hold in the case of self-trust. Forming an intention to φ is not the same as advising your future self to φ. When you advise yourself to do something you remain volitionally two-minded on the matter. Forming an intention makes you volitionally one-minded[22]—and in a way incompatible with any deliberative weighing of the reason to follow through that the intention gives you. So in the intrapersonal case, unlike the interpersonal, trust does reduce to compliance. Reasons based in trusting advice are not preemptive, but reasons based in trusting an intention are. The rational authority of an intending self is not advisorial but executive.[23]

When you φ purely on the basis of self-trust, you do not act on the basis of deliberation whether to φ: you simply accept your own earlier assessment of what reasons you have to φ. When you've deliberatively assessed what reasons you have, it remains simply to act on that assessment. But you cannot simply act on it unless you accept the assessment non-deliberatively. You cannot engage in diachronic agency without accepting an assessment of what reasons you have on trust. Pathology emerges when you refuse to accept an assessment you've made of what reasons you have without a basis that would make reasonable a *re*assessment of what reasons you have. For if your mistrust lacks a basis in considerations that would point you towards a reassessment of what reasons you have, you'll for the present prove incapable of acting on an assessment of what reasons you have. If diachronic agency just is being governed by an earlier assessment of what reasons you have, you'll for the present prove incapable of it.[24]

We might generalize this account of trust as the basis of diachronic agency. When others advise you, they too assess what reasons you have, and earlier selves are not inherently more trustworthy than concerned and well-informed advisors. So why should you not sometimes act purely on the basis of trust in a trustworthy advisor, if you typically act purely on the

basis of self-trust? Say you don't trust yourself to deliberate some matter but can't refrain from acting. (Perhaps the cultural context is unfamiliar. Or perhaps you're too nauseated to think clearly, or simply inebriated.) Here you'll have to treat your advisor's assessment of what reasons you have as a substitute for your own. Although advice aims to engage the deliberative faculties of the advisee, by your own consent it would bypass those faculties and serve as a source of preemptive reasons. I don't see why following through on this advisor's assessment without deliberation need be irrational.[25] You could form your intention by simply resolving to take this person's advice. The self-trust required in order to execute your intention would in fact be trust in another.

At the basis of diachronic agency lies simple trust. It seems incidental to agency that the one trusted should typically be you yourself. From this perspective, the stubbornness that consists in a baseless refusal to give the advice of a trustworthy advisor appropriate weight—whether deliberative or non-deliberative—looks like an interpersonal species of your pathology at the produce stand.[26]

V

When you intend to φ at t, by (I), you expect that you'll reasonably trust yourself at t and act on the reason you thereby give yourself to follow through on this intention without redeliberating. I've argued that you cannot give yourself this reason unless from your perspective at t you were trustworthy in forming the intention,[27] where that requires (*inter alia*) not basing it on false beliefs. So one implication of (I) in the context of my account of diachronic rationality is that you cannot form an intention unless you believe that you are trustworthy in forming it and will appear so to the self that will execute the intention.

We may have here a diagnosis of the intuition articulated by a number of influential philosophers that you cannot form an intention to φ if you believe that you will not φ.[28] I'm not sure that I share this intuition, thus stated. But a related intuition lies at the core of my approach to intending. It isn't that your intention-forming self must believe that your intention-executing self will succeed in actually φing. Nor is it quite that the earlier self must have the belief that the later self will follow through on the intention—that is, try to φ. Rather, the earlier self must believe that the later self will follow through on the intention *through an exercise of reasonable trust*. And, as I've now argued, the earlier self must believe that it will appear trustworthy on the matter at hand. Here we have the *derived* basis of the earlier self's belief that the later self will follow through on the intention. But this basis suggests more. In viewing itself as viewed by the later self as trustworthy on the question whether to φ, the earlier self views itself as viewed by the later self as authoritative on the question whether to φ.

Because this authority on a practical matter manifests itself in a way that bypasses the later self's deliberative reason—since the earlier self expects that the later self will follow through on the intention without deliberating—it's natural to assume that this practical authority entails a corresponding epistemic authority. That is, it's natural to assume that the intention-forming self will view itself through the eyes of the intention-executing self as authoritative not only on the practical matter whether to φ but on the epistemic matter whether its φing will occur. Since forming the intention involves taking on the mantle of the former authority, one might assume that it involves taking on the mantle of the latter authority. On this assumption, if you were not confident that your φing would occur, you could not form an intention to φ—just as you could not if you were not confident that you would appear trustworthy on the matter. I take no position on whether the assumption is correct. I merely observe that, if it is, its basis lies in the earlier self's attempt to engage the later self's reasonable trust.

(I) speaks not only of expectation but also of desire. According to (I), you cannot intend to φ merely by expecting that when the time comes to act you'll have a preemptive reason to φ. You must also desire that you'll have that reason. Your desire that you'll have this reason is, in effect, simply the desire that you'll follow through on the intention without redeliberating—provided it's trustworthily formed. When you form an intention you acknowledge that you may violate one of the conditions (a) through (d) listed in section I and that if so you would lack an intention-based reason to follow through. When you form an intention you do, however, desire that you'll follow through on it if you satisfy the conditions—if, that is, from the perspective of your future self you are trustworthy.[29]

Throughout this discussion I've meant by a "reason" a *normative* (or *justifying*) reason. I have asked what it would take for you to have a normative reason to follow through on an intention. Your normative reason would be useless, of course, if it were not also motivating. But this is not a problem for my account. When you trustworthily form the intention, you give yourself a reason to follow through on it that is both normative *and* motivating.

How do you acquire the motivating reason? You acquire it by virtue of the fact that trust is inherently motivating. When you trust your intention-forming self, you are motivated to follow through on the intention without redeliberating. That's simply part of what it is to trust. When you form the intention you desire that your intention-executing self will trust you in this way. But, as with your expectation discussed above, that desire is not basic: it follows from your desire that your intention-executing self will trust you if and only if trust in you is reasonable, together with your expectation that trust in you will be reasonable. Since in forming the intention you believe that you will be viewed as relevantly trustworthy, you expect that this condition will be satisfied. But the content of your desire includes the

condition. You desire that your intention-executing self not trust you if trust in you would not be reasonable. You aren't trying to trick your future self; you're inviting that self to trust you. If your future self does trust you, it will be motivated to follow through on the intention simply because it remembers that you invited it to. But, again, when you invite your future self to trust you, you desire that it trust you only if trusting you is reasonable. You desire that it trust you only if it has a reason to trust you. What reason? The reason you expect you'll thereby give it. The motivating reason to follow through derives from an appreciation of the normative reason to follow through.[30]

In forming an intention, in sum, you adopt a complex attitude toward the self of yours that will execute it. You expect and desire that your future self will trust you. But you equally expect and desire that your future self will *not* trust you unless you are indeed worthy of trust. What reconciles these attitudes is an attitude you bear toward yourself—that is, toward your present self. You cannot form an intention unless you view yourself as worthy of the trust that you expect and desire will lead you to follow through on it.

VI

I'll now defend my conception of trust as the basis of diachronic agency against objections and alternative accounts.

Michael Bratman objects that any account on which merely forming an intention creates a reason to execute it will allow agents to bootstrap their way into possession of reasons we do not think they possess.[31] Bratman has in mind cases in which you irrationally form an intention to φ. He is right that you cannot come into possession of a reason to φ simply by forming an intention to φ. On my account, you cannot come into possession of an intention-based reason unless you formed the intention not only rationally but in a way that manifests a more general trustworthiness on the matter at hand. This creates no space for unacceptable bootstrapping. You can no more bootstrap yourself into possession of a reason to execute an intention you've formed than you can bootstrap yourself into trustworthiness. But if you've earned your trust by forming the intention in a way that manifests trustworthiness on the matter, then you do come into possession of a reason to execute the intention. You come into possession of a reason, that is, to treat the matter as settled. As long as you continue to view your intention-forming self as trustworthy, and circumstances emerge as you'd envisioned them when you formed the intention, then you have an intention-based reason to do what you intend to do.

Bratman also objects to the idea that an intention-based reason might be weighed alongside other reasons.[32] This seems to lead to a kind of double-counting: you form an intention on the basis of various reasons, then when

you redeliberate, or merely reaffirm your intention, those reasons not only retain their force but acquire an indirect force through the force of the intention-based reason. The reasons created by the quasi-Gricean mechanism in the context of appropriate trust do not, however, simply transmit the force of other reasons. Rather, they *preempt* those reasons: they preempt reasons pro and con performance of the action in question. How can they do this? This preemptive capacity marks the fact that a reason created by the quasi-Gricean mechanism is a reason merely to follow through on an intention you have already formed. It is best conceived as a reason to treat the matter as settled and in that spirit simply to do what you intend to do.

VII

I've been asking how we can explain the rational bearing of deliberation on future action, of the formation of an intention on its execution. One alternative approach, Strategy A, denies that there's anything in need of explanation here. A proponent of this strategy views the relation in question as follows. You deliberate and form the intention. While you retain the intention, it serves as a structuring principle as you deliberate about means, for example, of fulfilling it. Then when the time comes to act, if events have emerged as you'd envisioned them and you haven't changed your preferences or values, you act. Forming the intention was lighting the fuse; the explosion is now your action. From this angle, there is no question concerning the rational bearing of deliberation on action. If your deliberation causes you to act via the relation just described, it has exercised its rational bearing on your conduct. That's just *what it is* for a deliberation to get a rational grip on conduct. So claims the proponent of Strategy A.

This is the strategy that Bratman takes in *Intention, Plans, and Practical Reason*.[33] On his view an intention does not provide a reason for its own execution, but an intention does provide a *framework* reason "whose role is to help determine the relevance and admissibility of options."[34] These would be options as to how to implement the intention, not whether to implement it. On Bratman's view, forming an intention is a way of committing yourself to act. While he distances himself from one feature of the "lit fuse" metaphor by emphasizing the role intentions play in structuring processes of practical reasoning, he accepts the feature of the metaphor that interests me, which is the idea that the relation between intention-formation and intention-execution need involve no rational dimension beyond this structuring. The manner in which the fuse burns down can be quite structurally complex, and I agree with much of Bratman's characterization of this complexity. But I disagree with his claim that deliberation could get a rational grip on future conduct merely by means of the process he characterizes.

To see the problem with Strategy A, consider a case it describes perfectly. You love eggplant and okra equally. But you decide on the basis of relevant

considerations that it would be best to serve eggplant tonight. You worry, however, that your love of okra will overwhelm you at the produce stand and lead you to change your mind—from your present perspective, irrationally. So just before you leave for the market you force yourself to eat a large meal of ill-prepared okra, hoping thereby to dampen your taste for okra for a few hours at least. And it works: at the produce stand you can hardly manage to look at the stuff.

Here we have an instance of what Jon Elster calls "precommitment."[35] When you precommit yourself you take steps to make it more likely that you'll execute an intention, either by physical means (as when Ulysses has himself tied to the mast to resist the sirens' attractions), by psychological means (as when you stuff yourself with okra to resist its attractions at the market), or by means of external measures (such as a side bet).[36] Precommitment is compatible with everything Bratman says in *Intention, Plans, and Practical Reason* about the role your intention may play in structuring your deliberations on matters of means. Your intention to buy eggplant, even implemented in the way just described, gives you framework reasons to consider only stores that you believe will have eggplant in stock and only those wines that you believe will complement eggplant as you deliberate further the details of your errand. It gives you a framework reason to consider inviting only those guests who will not be displeased with eggplant. This is not, in Bratman's phrase for an over-simple view he rejects, a picture of intention "exerting a ghostly mode of influence on later action."[37] The picture posits forethought, planning, and multiple sub-deliberations conducted in the light of that planning. Even so, however, it is equally not a picture of your deliberation exerting rational bearing on your conduct.

I do not claim that it is irrational to precommit yourself. I agree with Elster that it can be eminently rational. But when it is, what is rational is the act of precommitting yourself, not the bearing of your intention on your subsequent action. Given that you've precommitted yourself, your doing what you intend to do simply follows (if all goes well). Your doing what you intend to do is not here an instance, however, of your deliberation's getting a rational grip on your conduct. You decide to precommit yourself, after all, precisely because you doubt that your deliberation *can* get a rational grip on your conduct in this case. Precommitting yourself is an *alternative* to following through on an intention. You don't trust your future self to φ simply for the reason that you intend to φ. You think you'll act on that intention only if, when the time comes, you are prevented from doing anything else or, as here, from doing specifically what you fear you'll do if you aren't prevented. But you don't follow through on an intention by being prevented from doing other things. At least, "following through" on an intention in this way would be pathological. It turns autonomy into automatism, agency into a mere mechanism for getting intentions executed.

Agents aren't conduits through which a mechanism set in motion by intention-formation produces an action. Agents are not patients of their own intentions. When you rationally follow through on an intention, you remain the agent of that intention by retaining rational governance of it. To be in *rational* governance of it is to govern it by acting on a reason. You've already decided whether to φ. Now the issue is whether to follow through on your intention to φ. You rationally follow through on your intention to φ by acting on a reason, not specifically to φ, but to follow through on this intention you've formed on the basis of an assessment of what reasons you have to φ. You cannot *rationally* bridge the gap between deliberation and action without acting on an intention-based reason.[38]

I should note that Bratman specifically denies that his account models intentions on precommitments. He does not, to be sure, view intentions as functioning by means of the sort of "psychological resistance" that characterizes precommitments.[39] But precommitment employs psychological resistance toward the end of bypassing the need to follow through rationally on an intention. Without intention-based reasons, when you did not employ such resistance you'd merely have to hope that your intention would get executed. Perhaps it usually would. But it would no more count as your rationally following through on your intention than it would if you'd precommitted yourself. Precommitment involves actively setting out to make yourself the patient of your own intention. Merely hoping or even quite vividly anticipating that you'll be the patient of your intention won't get you in on your act in the way that you'd be in on your act as an agent. As an agent you'd be in on your act by rationally governing it. At this stage in the process, you'd rationally govern your act by acting on an intention-based reason to follow through on your intention to act.

VIII

Whereas I view them as practical reasons based on self-trust, alternative Strategy B views intention-based reasons as epistemic reasons deriving from a kind of evidence that your intention gives you. It views the relation between deliberation and action as follows. You deliberate and form the intention to φ. If circumstances then emerge as you envisioned them and you don't change your preferences or values, you have good evidence for believing that if you now redeliberated whether to φ you'd reach the same conclusion as you then reached. This belief gives you a reason to follow through on that intention *without* redeliberating. The fact that you have an intention to φ gives you good reason, *ceteris paribus*, to believe that if you redeliberated you'd form the same intention. But the belief that if you redeliberated you'd form the same intention gives you good reason not to redeliberate. So when instead of redeliberating you simply follow through

on your intention you do so on the basis of an intention-based reason to follow through. This is not, however, a reason based on the relation of self-trust that I am trying to characterize. It is a reason based on the evidence your having the intention gives you that redeliberating won't lead you to change your mind. A past deliberation has rational bearing on present conduct, according to this view, by testifying to the conclusion you would reach in a present deliberation.[40]

There is a problem with Strategy B, a problem that reveals what self-trust must be like for it to play the role I have assigned it. We are asking how an intention-forming self can exercise rational influence over an intention-executing self. Consider instead an interpersonal case of influence. Imagine that my attitude toward my brother has this implication: I treat the fact that he forms an intention to φ in such-and-such circumstances as good evidence that I would form an intention to φ in similar circumstances. Does that give me a reason—even, on the broadest construal, *ceteris paribus*—to follow through on an intention so formed? Hardly. I may believe that my deliberative tendencies resemble my brother's yet lament the fact. That I treat his forming an intention to φ as good evidence that I would form an intention to φ in similar circumstances doesn't show that I trust my brother. So described, it shows at most that I believe myself similar to him. To believe yourself similar to someone is not to trust that person. By parallel reasoning, to believe your present self similar to your earlier intention-forming self is not to trust that earlier self. So described, you have no more reason to follow through on your intention than I have reason to imitate my brother. Strategy B is meant to provide a reason for you to follow through on your intention. All it in fact provides is a reason to believe that if you redeliberated you'd affirm that intention. It leaves unanswered the question why you should follow through on it.

Self-trust is not the same as self-constancy. Strategy B appeals to a belief in self-constancy where what we need is an attitude of self-trust. Even on the assumption that the belief is true, and that you would redeliberatively affirm your intention, your affirmation would not give you a reason to follow through on it. The question is not whether you would reaffirm your intention but whether you should do so, just as the question is not whether you will follow through on the intention but whether you should do so. Your forming the intention gives you evidence that you *should* reaffirm it only if you are warranted in believing that the self that formed it is trustworthy. But Strategy B leaves the question of your trustworthiness unaddressed.

IX

The quasi-Gricean analysis (I) formulates a kind of 'linking principle.' It links the conditions for rationally forming an intention with the conditions for rationally following through on it. But (I) is weaker than the linking

principle influentially defended by Bratman. According to Bratman's principle, when on the basis of deliberation you form an intention to φ at t if conditions C hold, you must not expect that if C hold at t then you should, upon reconsideration, rationally abandon your intention in favor of an intention to perform some alternative act.[41] This is too strong. When you deliberate whether to φ instead of Ψ you needn't view yourself as rationally constrained by whether your future acting self would rationally decide to φ instead of Ψ in a last-gasp reconsideration.

Bratman's linking principle fits the advice-model I considered in Section IV. But, as I remarked, it is a mistake to model the intrapersonal rationality of intending on the interpersonal rationality of advising, as if intending to φ had the force of advising your future self to φ. If that were the correct model of intending then in intending, as in (sincere) advising, you'd be irrational if you expected that your addressee could not upon reflection rationally do as you say. But intending requires less than advising here. Intending is fully compatible with expecting that your future self could not *upon reconsideration* rationally do as you intend. What is incompatible with intending is, more generally, expecting that you do not thereby give your future self a rational basis for doing as you intend. According to my account of intention-based reasons, you can give your future self a rational basis for following through on an intention simply by forming it—so long as you are trustworthy in forming the intention and there is no good evidence that you are not. Though weaker than Bratman's, the linking principle is still robust. When you form an intention you cannot expect that your future acting self will have good grounds to view your present self as untrustworthy.

Bratman formulates his linking principle in the course of discussing Gregory Kavka's Toxin Puzzle.[42] It will help situate my contribution if I specify how the view I've defended applies to that much-discussed case. Imagine that an eccentric billionaire with a reliable intention detector reliably promises to give you a million dollars if at midnight tonight he detects that you intend to drink a certain toxin tomorrow at noon. You must form this intention, he stipulates, without ignorance, manifest irrationality or such external mechanisms as a side bet. If you do thus form the intention at midnight, the billionaire will deposit the money in your account tomorrow morning as soon as the bank opens. He doesn't care whether you drink the toxin, which you know will make you quite ill for a day or two but leave you thereafter unharmed. To get the money, you need merely form the intention to drink it.

Now I agree with Bratman that if you redeliberate the matter at 11 a.m. tomorrow, you'll be irrational if you do not change your mind. And I agree that, foreseeing this, you cannot form the intention to drink without ignorance, manifest irrationality or external measures. So your prospects of getting the million look pretty dim. But that's only, I think, because the case is set up in such a way as to make it *obvious* to anyone in your deliberative predicament that redeliberating will be in order as the hour to

drink draws nigh. Your options are described, after all, in a way designed specifically to draw your attention to this fact. In an ordinary case your attention will be directed more to the attractions of your present options, or perhaps of the corresponding plans taken as wholes, and less to the attractions of reconsidering.

Imagine then a somewhat different toxin case in which it isn't so clear in advance that you'll be inclined to redeliberate before the time comes to drink. Imagine that you expect to be very busy or otherwise distracted as the hour draws nigh and therefore inclined to follow through on intentions formed earlier—as long as your intention-forming self appears trustworthy. (That, after all, is usually the point of making your mind up in advance.) As you form the intention to drink, you believe that were you to redeliberate at 11 a.m. you'd be irrational if you did not change your mind—in light of new options now open to you. But you reasonably do not expect that you'll conduct that deliberation. You reasonably expect that you'll be too busy or otherwise distracted to redeliberate in light of any new options that you'll come to have in the interval between forming and executing your intention. Under these new terms, I don't see why you could not both form the intention to drink the toxin and without irrationality follow through on it.

Following through would be irrational, of course, if your intention-forming self were untrustworthy. Is it? You formed the intention to drink despite knowing that doing so would make it possible for you to pursue the different outcome—getting the payoff without drinking—that you all along prefer to all others. This outcome was not, of course, an option for you as you formed the intention; but forming the intention immediately made it an option. Is your intention-forming self untrustworthy in not giving this fact due weight in deliberation? Well, how *could* you have given it due weight in deliberation? It wasn't an option then! Upon deliberation you chose the option that you then most preferred and continue to most prefer among those open to you at that time. Like Gauthier, I think it's important that you continue to believe that you took the right option among those then actually open to you.

But this point of agreement rests on a deeper disagreement. Gauthier holds that you should follow through on an intention to φ at t if at t you prefer the sequence {intending to φ at t, φing at t} to the sequences corresponding to actions you might have chosen instead. Since you do have this preference in the toxin puzzle (where $\varphi =$ drink the toxin), he argues that it is rational to follow through on your intention and drink.[43] Bratman, by contrast, holds that you should follow through on an intention to φ at t if at t you continue to prefer to φ—unless, he now adds, you would regret that preference later, at what he calls "plan's end." Since you would regret drinking the toxin as soon as the pointlessness of doing so occurred to you, Bratman argues that it is irrational to drink the toxin.[44] Each therefore holds that the rationality of following through on an intention

to φ at t can be explained in terms of a principle for choosing at t to φ—or, as it were, for 'seconding' at t an earlier choice to φ at t. Bratman's linking principle is therefore a point of agreement between Bratman and Gauthier.[45] In rejecting it as too strong, I disagree with both of them.

My approach to diachronic agency—to what Bratman calls the "stability of intention"—is utterly different. I am defending neither a principle of choice nor a principle for 'seconding' a choice. I therefore in fact have no stake in adjudicating Kavka's Toxin Puzzle, if doing so means explaining how choosing to drink may or may not be 'seconded' in a last-gasp deliberation. In forming an intention to drink, as with any intention, you aim to provide your future acting self with a reason to follow through without further deliberation, and without 'seconding' an earlier deliberation. Yet in Kavka's original case we imagine you faced with a deliberative quandary as the time comes to drink. Imagining your future self thus would indeed prevent you from forming an intention to drink. But that is only because it would prevent you from ascribing to yourself the executive authority that you presume in forming an intention. You cannot view yourself as trustworthy from the perspective of your future self if you view that future self as likely to redeliberate the matter in light of options it has gained in the meantime. In viewing your future self as likely to redeliberate, you do not view your present self as having resolved the question of what to do. But if you do not view yourself as having resolved that question, you do not view yourself as having formed an intention. You can, however, resolve the question of what to do while nonetheless believing that *if* you redeliberated you would change your mind. You need merely believe that you will not redeliberate. 'Making up your mind' in the way characteristic of intention *requires* that you view your future self as rationally *entitled* to follow through on the intention without redeliberating; you need merely believe that it will act on this entitlement. Bratman's linking principle links earlier and later deliberative perspectives. What an account of diachronic agency needs, however, is a principle linking your deliberation with a later perspective that is *non*-deliberative.[46]

For similar reasons, I do not claim that it is deliberatively irrational to change your mind in my earlier example and buy okra instead of eggplant. I do not claim that it is deliberatively irrational to spend your afternoon at the produce stand repeatedly redeliberating. I claim only that these manifest a pathology of agency because mistrusting yourself in this way is more broadly unreasonable. The species of irrationality in play is non-deliberative. That deliberative rationality is compatible with such pathology is what reveals how agency depends on non-deliberative rationality.

X

Self-trust provides the mechanism, I've argued, by which we are capable of such a thing as guiding our conduct by an earlier deliberation. We are

capable of diachronic agency even when we mistrust ourselves. But we are not thus capable in the absence altogether of trust, since when we mistrust ourselves our capacity for diachronic agency—for making diachronic sense in what we do—will depend on our trust in others' advice.[47]

I have not argued that we therefore have a reason to trust. I do not believe that we do have a reason to trust simply because failure to trust precludes diachronic agency. It is not reasonable to trust those who are not worthy of our trust, whether ourselves or others. If no one is trustworthy, no one has a reason to follow through either on his or her intentions or on others' advice. If no one is trustworthy, diachronic agency—the life of practical reason—is simply impossible.[48]

What makes diachronic agency possible is what makes practical commitment possible: the fact that we regard ourselves as—*and are*[49]—trustworthy executives over our own future conduct, and sometimes, at least *in extremis*, over one another.

Notes

* I wrote the first draft of this paper during a pleasant term visiting at Bowdoin College; thanks to former colleagues there (especially Scott Sehon) for brainstorming with me. I presented later drafts to a seminar at Michigan (thanks to Krista Lawlor, Laura Schroeter and Nishi Shah for giving me grief) and to an audience at the Eastern APA (thanks to Michael Robins for commenting). For additional comments, I'm grateful to Michael Bratman, Steve Darwall, David Hills, Andrea Westlund, and two anonymous referees. Special thanks to Jim Joyce and David Velleman, both for comments and for getting me engaged on the issue.

[1] Assume that there's little cost in redeliberating. (I'll consider how it might help to appeal to deliberation costs in section I.)

[2] I focus exclusively on cases in which there is a temporal gap between formation and execution of the intention. Of course, sometimes there is not. There is sometimes no distinction to be drawn between an earlier intention-forming self and a later intention-executing self—as, for example, when you choose a path while fleeing a predator. I do not have such cases in mind. (For an account of the more general synchronic role of self-trust in judgment, see my "Judging as Inviting Self-Trust," in preparation.)

[3] One might say instead that they doubt their ability to determine whether this apparent memory really is a memory. Though the pathology is ripe for philosophical examination, I do not pursue it here—nor does my argument depend on it on any analogy between the epistemic case and the practical.

[4] We face a more general problem if we bridge the gap between deliberation and action, between intention-formation and intention-execution, with another deliberation, or with the formation of another intention. If you get from deliberation to action by another deliberation, the question remains how the latter deliberation succeeded in rationally guiding your conduct. But that is the same question we asked regarding the former deliberation. Appealing to a second deliberation merely deflects the question without answering it. This is the practical-inferential analogue of the point Lewis Carroll makes about theoretical inference in "What the Tortoise Said to Achilles," *Mind* 4 (1895). On the analogy, see Simon Blackburn, "Practical Tortoise Raising," *Mind* 104 (1995), and Peter Railton, "On the Hypothetical and Non-Hypothetical in Reasoning about Belief and

Action,'' in Garrett Cullity and Berys Gaut (eds), *Ethics and Practical Reason* (Oxford: Oxford University Press, 1997), 76–8.

[5] One might, I suppose, simply insist that it is positively rational to stand there and redeliberate as long as it takes to make the itch go away—so long as you've nothing better to do, and even if the only thing capable of making the itch go away is having something better to do. It is harder to imagine how, so gripped, you could view yourself as rational. You don't, after all, view the redeliberations as likely to change your mind. You expect merely that they will reduce the unpleasantness of self-doubt—briefly. Assessing the costs and benefits may lead you repeatedly to redeliberate (since you've nothing better to do), but you thereby become in an important respect unintelligible to yourself. Perhaps this is why the stance strikes me as intuitively pathological. (This anticipates the account I'll offer in section III.)

[6] J. David Velleman, *Practical Reflection* (Princeton: Princeton University Press, 1989), Chs. 3 and 4.

[7] Velleman, ''Introduction,'' in his *The Possibility of Practical Reason* (Oxford: Oxford University Press, 2000), p. 26.

[8] Elizabeth Anderson joins Velleman in arguing that autonomy requires treating any intention that you've formed as giving you a reason to follow through on it:

> We can frame coherent intentions only under a conception of ourselves as autonomous agents. We act autonomously only if we act on reasons that we regard as warranting our actions. But agency is exercised over time. Autonomy thus requires that decisions we make now be able to control our future actions. But we cannot act autonomously at that future time unless the sort of control our prior selves exercised is compatible with our acting for good reasons at that time. This is possible only if we regard the commitments we made then as giving us reasons to fulfill them now. (''Reasons, Attitudes, and Values: Replies to Sturgeon and Piper,'' *Ethics* 106 (1996), p. 542)

I agree with these observations up to the last sentence. A prior self that controls your actions without being worthy of your trust does not, I would say, control them in a way compatible with your acting for good reasons. It is possible for you to act for good reasons only if you regard the commitments you made then *that are worthy of your present trust* as giving you reasons to fulfill them now. The italicized phrase marks an important divergence from Anderson's and Velleman's attempt to derive intention-based reasons from the nature of autonomous agency. Both Anderson and Velleman hold that because the link between deliberation and action must be reason-giving it must also be autonomy-preserving. (See also Velleman's ''Deciding How to Decide,'' in Garrett Cullity and Berys Gaut (eds), *Ethics and Practical Reason* (Oxford: Oxford University Press, 1997), pp. 42–50.) My view of the role that you may need others to play in your practical-reflective life when you do not trust your own deliberations leads me to suspect that trust can provide the crucial link between deliberation and action without preserving autonomy. I don't see why it should undermine your rationality if you trust *another's* deliberation when you act. Here, perhaps, we have rational agency without autonomy. Although I am going to agree with Anderson and Velleman that there are intention-based reasons, I do not argue for their existence from considerations of autonomy. In my argument from trust, I model the intrapersonal on the interpersonal.

[9] See ''Deciding How to Decide.''

[10] Here I dispute John Broome's reading of Gauthier's position (''Are Intentions Reasons? And How Should We Cope with Incommensurable Values?'' in Christopher W. Morris and Arthur Ripstein (eds.), *Practical Rationality and Preference* (Cambridge: Cambridge University Press, 2001), pp. 100–102). Broome views Gauthier as maintaining that the intention *is* a reason. But Gauthier holds merely that possessing the intention is

a necessary condition for possessing the reason. What generates the reason is not the individual intention but the general principle just stated in text. (On Gauthier's appeal to generality, see Michael Thompson, "Two Forms of Practical Generality," in Morris and Ripstein (eds), *Practical Rationality and Preference*, pp. 131–133.)

[11] I focus on Gauthier's formulation of his position in "Rationality and the Rational Aim," in Jonathan Dancy (ed.), *Reading Parfit* (Oxford: Blackwell, 1997). While its composition dates from the late 1980s, this paper formulates especially clearly a feature of Gauthier's view that survives his subsequent reformulations, a feature that he presents by contrast with Derek Parfit's view as follows: "Parfit's view implies that any theory of rationality tells me that my reasons for acting are considerations about what would be conducive to the aim it gives me. And I hold that any theory of rationality tells me that my reasons for acting are considerations determined *by my being disposed* in the ways that are most conducive to the aim it gives me" (*ibid.*, p. 40; emphasis added).

[12] "Rationality and the Rational Aim," p. 33: "but no single coherent disposition can motivate both members of this set."

[13] Holly Smith poses the challenge as directed at what she calls the "causal efficacy thesis" (to which she views Gauthier as committed), the thesis that forming an intention to do A will cause the performance of A ("Deriving Morality from Rationality," in Peter Vallentyne (ed.), *Contractarianism and Rational Choice* (Cambridge: Cambridge University Press, 1991, pp. 235–237). As she notes, "it is implausible to suppose our commitments always compel our future acts—especially in the kind of case in question, where considerations of utility press the agent to change her mind when the time comes. Indeed, if the causal efficacy thesis were true, it is hard to see how Prisoner's Dilemmas could have been the deep historical problem for social cooperation that they have been" (p. 236). Since I'm not sure that Gauthier is committed to the causal efficacy thesis, I prefer to pose the challenge to his thesis that the disposition in question is not "coherent."

[14] My debate with Velleman will turn in part on whether he can explain in terms of what he calls "constancy" the intrapsychic phenomena that I treat in terms of self-trust. Constancy, as he defines it, is an inclination consisting in "the desire for grounds on which to make plans," which he glosses as "the desire for evidence of having precisely this inclination" (*Practical Reflection*, p. 227). Velleman considers a case in which an agent loses confidence in herself by coming into possession of evidence that she will not stick to her plans: evidence that she will not follow through on her plans begins to shake her faith in herself as a planner. Her lack of constancy—her manifest lack of desire to give herself evidence that she will follow through on her intentions—might if it persists prevent her from forming intentions in the first place. This is, I concede, a failure of self-trust. But it is nonetheless a different failure from the failure of self-trust that I posed in my challenge to Velleman. My challenge derives from the possibility, not that you lack grounds for trusting your future self to execute your intention, but that you lack grounds for trusting your former self to have formed an intention that you now have reason to execute. The problem isn't the untrustworthiness exemplified by fickleness but the untrustworthiness exemplified by foolishness. (I say more about the relation between self-trust and self-constancy in section VIII.)

For Gauthier's reply, see the papers cited in note 43 below.

[15] I'll explain what I mean by a "preemptive" reason in section IV. Read the "at that time" in (I) to exclude such failures of self-knowledge as John Perry discusses in "The Essential Indexical," *Noûs* 13 (1979). It is, in Hector-Neri Castaneda's term, a quasi-indicator (see his "Indicators and Quasi-Indicators," *American Philosophical Quarterly* 4 (1967), esp. pp. 93–96).

[16] I'll henceforth omit the qualification "on the matter at hand"—but it should always be understood when I speak of trustworthiness. Your intention-forming self may of course be trustworthy about some matters (say, how to find food) but not about others (say, how to find love).

[17] I add the 'externalist' defeating condition because we often find it natural to say that an act that seems reasonable to an agent 'turns out' to be unreasonable after all, or that an act that seems unreasonable 'turns out' to be reasonable after all, in light of truths that the agent was not in a position to know. Thus we might naturally say that trust without satisfaction of the second defeating condition turns out to be unreasonable through satisfaction of the first, or that mistrust without satisfaction of the second defeating condition turns out to be reasonable through satisfaction of the first. But nothing of importance in this paper hangs on the inclusion of this externalist element. One could avoid the externalism by dropping the first condition, or by interpreting it as shorthand for the second. 'X is *worthy* of Y's trust' says that Y is (*pro tanto*) entitled to trust X. We can leave it open that the entitlement should be given an 'internalist' reading.

[18] In what follows I have been helped by some perceptive comments from an anonymous referee.

[19] Perhaps it is in this respect similar to Derek Parfit's cases of ''rational irrationality'' (see *Reasons and Persons* (Oxford: Oxford University Press, 1984), pp. 9–13; and ''Bombs and Coconuts, or Rational Irrationality,'' in Morris and Ripstein (eds.), *Practical Rationality and Preference*).

[20] Again, I do not insist on the externalism. (See note 17.)

[21] I treat the advisorial case more fully in my ''Advising as Inviting to Trust,'' in preparation.

[22] Being volitionally one-minded, as I conceive it, is compatible with being desideratively two-minded—with retaining ambivalent desires on the matter. In that case you do not, in Harry Frankfurt's sense, 'identify' yourself with one of the desires (see, for example, Frankfurt's ''The Faintest Passion,'' in his *Necessity, Volition, and Love* (Cambridge: Cambridge University Press, 1999)). Having ambivalent desires is not the same as *being* ambivalent.

[23] This distinction raises a possible complication. One could view the issue as a question of whom to trust as you arrive at the produce stand: your earlier intention-forming self or your current self that itches to redeliberate. But the self that itches to redeliberate does not aim at executive authority. I don't actually think it helpful to view the temptation to redeliberate as issued by a 'self,' since there is no temporal gap, but if we did we'd have to view that self as aiming at advisorial authority—if we viewed it as aiming at rational authority at all, rather than at some bruter sort of influence. The temptation would be like an advisor offering you the advice to redeliberate. But advisors, unlike executives, aim to give reasons that engage their interlocutor's deliberative faculty: the advisee can weigh this *pro tanto* reason against competing reasons in deciding what to do. So we'd have to imagine a deliberation on the question of whether to redeliberate. That needn't be incompatible with the aim of your intention-forming self: it aims at your following through on its intention without redeliberating the first-order matter of what to do. We're now imagining you conducting a deliberation on the second-order matter of whom to trust concerning what to do.

[24] For different versions of the view that a capacity to treat something as a reason is the key to agency—without special reference to the diachronic case—see Bratman, ''Identification, Decision, and Treating as a Reason,'' in his *Faces of Intention* (Cambridge: Cambridge University Press, 1999); Christine M. Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996), pp. 94, 97–8; T. M. Scanlon, *What We Owe to Each Other* (Cambridge: Harvard University Press, 1999), Ch. I; and Velleman, ''What Happens When Someone Acts?,'' *Mind* 101 (1992), and the Introduction to *The Possibility of Practical Reason*. Bratman suggests that Harry Frankfurt may embrace a version of this view when he speaks of ''identification'' with a desire (*ibid.*, notes 33, 38 and 40). For an argument against the thesis that this can be the *substantive* (as opposed to merely formal) aim of agency, see Velleman, ''The Possibility of Practical Reason,'' *Ethics* 106 (1996).

[25] Of course, you'll deliberate whether to perform the act described as 'following through on this advice without deliberating.' And performing that act may be deliberatively rational. The question is whether the advice could give you a reason in the way that an intention gives you a reason. I don't see why not. Following advice in this way is mediated by your capacity for reasonable trust and is therefore a manifestation of your agency—even though the content of your will gets determined by the will of another. It isn't like resolving to let yourself be buffeted about by the next strong breeze. You could decide to do that, but then your behavior from moment to moment specifically wouldn't count as manifesting your agency. Trusting an advisor needn't in that way amount to an abdication of agency.

[26] I consider such interpersonal pathologies more fully in ''Advising as Inviting to Trust.''

[27] By ''From Y's perspective X is trustworthy in forming the intention'' (and similar formulations) I mean: Relative to the question of Y's follow-through, X satisfies neither of the defeating conditions on trustworthiness in forming the intention. I don't mean to imply that Y has *assessed* X as trustworthy.

[28] The question whether intending to φ entails believing that you will succeed in φing has been much debated. For an affirmative answer, see H. P. Grice, ''Intention and Uncertainty,'' *Proceedings of the British Academy* 57 (1971); Gilbert Harman, *Change in View* (Cambridge: MIT Press, 1986), Ch. 8; Velleman, *Practical Reflection*, Ch. 4. (See Velleman's note 8 on pp. 113–114 for a list of further adherents.) For a negative answer, see Donald Davidson, ''Intending,'' in his *Essays on Actions and Events* (Oxford: Oxford University Press, 1980); Michael E. Bratman, *Intention, Plans, and Practical Reason* (Cambridge: Harvard University Press, 1987), Ch. 3.4.

[29] For present purposes I assume that the list in section I sketches an adequate account of trustworthiness. If it doesn't, let ''conditions'' in this sentence and the last refer to whatever conditions figure in an adequate account.

[30] The word ''appreciation'' here may carry misleading intellectualist overtones. But recall that I characterized the role of the earlier self's trustworthiness in creating the reason in terms of defeating conditions. The later self needn't form any belief about the earlier self's trustworthiness. Its trust need merely be informed by a capacity to respond appropriately to evidence of untrustworthiness should any be available.

[31] Bratman, *Intention, Plans, and Practical Reason*, pp. 23–27, 86–87. More recently, see John Broome's ''Are Intentions Reasons? And How Should We Cope with Incommensurable Values?'' *op. cit*. Broome argues that the postulation of intention-based reasons rests on a conflation of reasons with normative requirements. (This interesting paper came to my attention too late for me to treat it here. In my ''Regret and the Unity of Agency,'' in preparation, I raise an objection to Broome's crucial appeal to the *repudiation* of an intention. In a nutshell, I don't think he has appreciated this intrapersonal relation's potential for pathology—and thus for non-deliberative irrationality.)

[32] Bratman treats this as a subproblem of the general bootstrapping problem.

[33] In more recent work he takes a somewhat different strategy, which I consider briefly in section IX. When I speak of ''Bratman's view'' in the present section I may thus be referring to a thesis which Bratman no longer endorses.

[34] Bratman, *Intention, Plans, and Practical Reason*, p. 109.

[35] Jon Elster, *Ulysses and the Sirens* (Cambridge: Cambridge University Press, 1984), Ch. 2, and *Ulysses Unbound* (Cambridge: Cambridge University Press, 2000), Ch. I.

[36] By an ''external measure'' I mean here any measure designed to increase the costs or decrease the benefits to you of not following through on your intention. For example, you might bet someone that you will follow through (which will increase the costs to you of not following through), or you might do something now that has the consequence when the time comes to act that you've nothing to lose—perhaps, nothing more to lose—by following through (which will decrease the benefits to you of not following through).

[37] Bratman, *Intention, Plans, and Practical Reason*, p. 108.

[38] I am not arguing that rationally following through on an intention is incompatible with the intention's *causing* you to follow through. To say that the relation between the formation and the execution of an intention cannot be *merely* causal is not to say that it cannot be causal. (Holly Smith suggests that some may find an incompatibility here (''Deriving Morality from Rationality,'' p. 237, n. 13). I agree with her, and with Gauthier, that there is none.)

Let me re-emphasize that, despite the central role that it gives intention-based reasons, my account is not *rationalistic*. In claiming that these reasons derive from trust, I claim specifically that you don't have to *think* your way into possession of them. As I emphasized in sections I and IV, these reasons differ from reasons that you can weigh in deliberation in precisely that respect.

[39] Bratman, *Intention, Plans, and Practical Reason*, p. 12.

[40] This strategy simplifies a strategy sketched by James M. Joyce in *The Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press, 1999), p. 60. Joyce's own less simple version of it obviates the objection I am about to raise by adding conditions that amount to the stipulation that you view the results of your prior deliberation as giving you good evidence, not about what you would decide if you redeliberated, but about the present interests that would rationally inform a redeliberation. You can meet this stipulation only if you view your intention-forming self as trustworthy.

[41] See ''Toxin, Temptation, and the Stability of Intention,'' in *Faces of Intention*, p. 64. Bratman's formulation does not contain the clause 'upon reconsideration,' but it is clear from his applications of the principle that this is how he understands it. (I've left off a second condition in the antecedent of this conditional—''and you expect that if C you'll retain rational control of your action''—which Bratman includes in order to allow cases of rational irrationality not at issue here.)

[42] Gregory S. Kavka, ''The Toxin Puzzle,'' *Analysis* 43 (1983).

[43] For Gauthier's treatment of the Toxin Puzzle, see ''Assure and Threaten,'' *Ethics* 104 (1994); ''Commitment and Choice: An Essay on the Rationality of Plans,'' in F. Farina, F. Hahn, and S. Vannucci (eds), *Ethics, Rationality, and Economic Behaviour* (Oxford: Oxford University Press, 1996); ''Rethinking the Toxin Puzzle,'' in Jules L. Coleman and Christopher W. Morris (eds), *Rational Commitment and Social Justice* (Cambridge: Cambridge University Press, 1998); and ''Intention and Deliberation,'' in Peter Danielson (ed.), *Modeling Rationality, Morality, and Evolution* (Oxford: Oxford University Press, 1998).

[44] See ''Toxin, Temptation, and the Stability of Intention'' for the additional element. This is a modification of the view he defended in *Intention, Plans, and Practical Reason*. I discuss Bratman's ''no-regret condition'' at length in ''Regret and the Unity of Agency.''

[45] As Bratman acknowledges: see ''Following Through with One's Plans: Reply to David Gauthier'' in Danielson (ed.), *Modeling Rationality, Morality, and Evolution*.

[46] We can explain Bratman's and Gauthier's convergence on their linking principle by noting that the principle serves as the contrapositive of what Michael Thompson calls a ''transfer principle'' (''Two Forms of Practical Generality,'' *op. cit*., pp. 129*ff*), to wit: *a rational disposition or intention makes the actions manifesting it rational*. Or rather, by noting that it thereby serves as an interpretation of this transfer principle, ensuring that its concluding ''rational'' means *deliberatively rational*. In rejecting the linking principle, I also reject the transfer principle, at least thus interpreted. Where Thompson worries about how to put a properly practical—as opposed to purely psychological—conception of disposition or intention to work in such a principle, I worry that no such principle can codify the non-deliberative species of rationality under discussion in this paper. If we're sensitive to the trust-negotiating dynamics I've made a stab at depicting here, I see no reason why the explanantia in practical philosophy cannot in the end be 'purely' psychological. (The present inquiry is of course barely a brushstroke on that canvas.)

[47] Advisorial trustworthiness seems to require both more and less than the species of trustworthiness characterized in section I above. It requires less insofar as it does not require that condition (c) be met: you can receive good advice from advisors who do not share your values. (The phenomenon of rational conversion requires this; see my ''Conversions, Reasons, and Causes,'' in preparation.) It requires more insofar as, in place of condition (c), it requires a broader advisorial competence. (In prudential advice, this competence takes the form of appropriate care for the advisee. In moral advice, it takes the form of competence as a moral judge. For the beginnings of an account, see my ''Advising as Inviting to Trust.'')

[48] There is room here for a sceptical thesis. While I am not myself a sceptic, it is unclear to me whether or how scepticism about diachronic agency might be refuted—as opposed to contextually defused.

[49] Note again that I have not ruled out an 'internalist' reading of ''trustworthy.'' (See one last time note 17.) On that reading, the point would be that you don't give one who trusts you evidence of your trustworthiness merely by virtue of either party's *believing* that you're trustworthy. Whether or not your status as trustworthy can be exhausted by the evidence you make available to those who trust you (the internalism debate in this context), the condition on the possibility of agency at issue cannot be met by mere belief that the corresponding entitlement obtains. Even if not 'external,' the norm governing diachronic agency is thus objective.