European Journal of Philosophy

DOI: 10.1111/ejop.12100

Narrative and the Stability of Intention

Edward S. Hinchman

Abstract: This paper addresses a problem concerning the rational stability of intention. When you form an intention to φ at some future time *t*, you thereby make it subjectively rational for you to follow through and φ at *t*, even if—hypothetically—you would abandon the intention were you to redeliberate at *t*. It is hard to understand how this is possible. Shouldn't the perspective of your acting self be what determines what is then subjectively rational for you? I aim to solve this problem by highlighting a role for narrative in intention. I'll argue that committing yourself to a course of action by intending to pursue it crucially involves the expectation that your acting self will be 'swept along' by its participation in a distinctively narrative form of self-understanding. I'll motivate my approach by criticizing Richard Holton's and Michael Bratman's recent treatments of the stability of intention, though my account also borrows from Bratman's work. I'll likewise criticize and borrow from David Velleman's work on narrative and self-intelligibility. When the pieces fall into place, we'll see how intending is akin to telling your future self a kind of story. My thesis is not that you address your acting self but that your acting self figures as a 'character' in the 'story' that you address to a still later self. Unlike other appeals to narrative in agency, mine will explain how as narrator you address a specifically intrapersonal audience.

Anyone who has spent time near a toddler at play must suspect that storytelling figures somewhere in how we learn to perform self-governed actions. Small children are constantly telling stories as they plan and carry out their play, addressing not only others but themselves. One often feels one is merely overhearing (an impression confirmed when the story continues uninterrupted while one leaves and returns). One has the impression that the child is learning how to form and follow through on intentions and more generally on felt practical commitments (witness the kicking and screaming when one intervenes) and that the stories help organize the diachronic dimension of the enterprise. Must this dimension of how we learn diachronic self-governance figure in an explanation of its very nature? An impressive number of philosophers, psychologists, and cognitive scientists argue that it must.[1] But there are three fundamental puzzles in the hypothesis.

As you watch the toddler narrate what he's doing—'I'm building a fort to keep safe from the animals', he says as he piles pillows on one end of the sofa, stuffed hippo and booby on the other—he may strike you as curiously disconnected from his actions. He seems to be doing two things at once: piling up some stuff, and interpreting his actions in terms of a fanciful story.[2] What is the relation

between the action and the interpretation? There is no puzzle in viewing the child as simultaneously acting and offering a fanciful narrative of what he's doing.[3] But intentional agency unfolds prospectively. ('Now how will we be safe?' the toddler warns as you interrupt his project.) A puzzle lies in the hypothesis that the story could figure in the intentional basis of the action. How could narrative go deeper than onlooking commentary and somehow figure in an intention's or action's very nature? Call this the puzzle of prospectivity.

The second puzzle arises for an attempt to resolve the first. As the child wants it to be true of him that he is building that fort, perhaps agents are more generally driven to make true the stories they tell of themselves. Doing that from moment to moment would involve satisfying the story's description of its protagonist at each of these moments. But now we get a new puzzle. The story's moment-to-moment description of its protagonist is merely a series of descriptions. If we say that an action is informed by the agent's desire or drive to satisfy a description figuring at a given moment in the story, the narrative element drops out. The pillow-stacking would manifest a desire to be fort-builder, but not in a way that makes any *essential* use of the story the child is telling.[4] How could the narrative figure *as such* in the intention or action? Call this the puzzle of diachronic integrity.

Each of these puzzles has been discussed in the literature,[5] but a third arguably more fundamental puzzle has not. The literature on narrative and agency makes frequent appeal to the telling of stories, but a story is typically told not into the ether but to an audience. If our toddler is practicing narrative self-governance with his antics on the sofa, to whom is he addressing his story? If to himself, how could he at once serve as both teller and audience? We don't otherwise typically address stories to ourselves. When we read silently we are the audience and not the teller. And when we rehearse a story to ourselves—'Here's how the car got that dent, honey . . .'—we are typically addressing someone other than ourselves in imagination. Whom exactly do agents address when they tell the stories that somehow inform their intentions or actions? What exactly is the connection between how they address those stories and their capacity to intend or act? Call this the puzzle of the addressee.

The puzzle of the addressee provides not merely a puzzle but a clue. This paper's thesis is that intention specifically rests on a capacity to address one's future self and that the mode of address cannot do its work with mere act descriptions: what you thereby tell your prospective self must have the diachronic integrity of a narrative. We'll thereby see the puzzles resolved while coming to appreciate how deeply narrative figures in the learned capacity for self-governance. But why think that narrative really does inform the nature of agency? We need a reason to believe that we are not misled by the appearances from which we began. We can derive a basis for my thesis from outside those appearances, I'll argue, by confronting a problem that goes to the core of diachronic agency: how could an intention have rational stability?

I'll discuss the rational stability of intention at length in section 1. Here is a brief statement of the problem it poses. Observe that forming an intention to φ

at some future time *t* generates this rational commitment: you thereby make it 'subjectively' rational (that is, rational in light of your present attitudes) for you to follow through and φ at *t*, *even if*—hypothetically—you would abandon the intention were you to redeliberate at or just before *t*. Indeed, part of the point of forming an intention is that your earlier perspective—the perspective from which you deliberated whether to φ at *t*—thereby overrides the rational authority of your later perspective as the time of action approaches. Of course, if you actually redeliberate at or just before the time of action and decide not to φ at *t*, that counts as changing your mind and your earlier perspective loses the rational authority it had. But it seems odd that you should have to deliberate from it for the later perspective—*your* perspective, at or just before the time of action—to have rational authority for you. Even if you don't actually redeliberate what to do at *t*, shouldn't the perspective of your *acting* self at *t* be what determines what is then subjectively rational for you?

My approach to the problem of rational stability will highlight a crucial role for narrative in intention. I'll argue that your claim to rational authority when you intend posits an intrapersonal relation whose normative properties we can best understand by comparing it to the rhetorical relation in which a storyteller stands to her audience. The idea is not that you address your acting self. On the account I'll develop, your acting self figures as a 'character' in the story you address to a still later self. Unlike other accounts of agency that appeal to narrative self-understanding, mine will emphasize how the 'story' in question addresses its intrapersonal audience. I'll argue that committing yourself to φing at *t* by forming an intention to φ at *t* crucially involves the expectation that your self at *t* will be 'swept along' by its participation in an appropriately addressed structure of narrative influence. My thesis is that we theorists need to conceive the intrapersonal relation at the core of intention in these terms if we are to understand how the intentions of creatures like us rationally resist reconsideration—that is, are stable—in the way we'll consider. In the long concluding section I'll frame my thesis more abstractly, asking what it says about the nature of rationality and whether it need hold of rational beings less invested in narrative than we are.

1. What is the Rational Stability of Intention?

Let me clarify how the rational stability of intention poses a problem. I'll make seven points.

(1) The rational commitment at the core of intention is defeasible. Again, if you actually redeliberate and change your mind whether to φ at *t*, the commitment at the core of the now abandoned intention vanishes. The problem is not that you're in rational error if you redeliberate but that you're *not* in rational error when you follow through on an intention that you would have abandoned if you had redeliberated. The problem arises from reflection on the hypothetical case.

(2) You need not address the commitment to anyone but yourself, and the way you address it to yourself cannot be understood as an intrapersonal analogue of a promise. The problem is not that you will *owe* it to yourself to act but that committing yourself now can override your own later perspective by making it then subjectively rational for you to act.

(3) The problem thus amounts to the challenge of explaining a species of intrapersonal influence. How it can be subjectively rational to put your later acting self under the influence of your earlier intending self? Again, by 'subjectively rational' here I mean rational from the perspective of your acting self. At first glance, it is easy to understand how it can be rational from the perspective of your intending self to put your acting self under its influence. You may expect that your acting self would not have time to deliberate or a disposition to conclude as you now see is best. But why should such a project of self-influence rationally bind you when the time comes to act? One might think it sufficient to observe that you remain the same person through these transitions, but that observation actually makes the self-influence harder to understand. Insofar as you treat your future acting self as rationally subject to your influence, you appear to treat it as if it were not a self of your own, since you do not treat its deliberative perspective as speaking for you. Again, that attitude at first glance seems easy to understand looking forward, since you expect that this future perspective may be corrupted by a transient preference reversal. But why should your acting self, from *its* point of view, let itself undergo this influence?

(4) The fact that this is self-influence—that is, *intra*personal influence—makes that question equally pressing from your perspective as you form the intention. Ease at first glance gives way to difficulty on reflection. Forming an intention is not like binding your future self to a mast in the way of Ulysses, in order to sail your action safely past temptation.[6] When you form an intention, you expect not merely to override the perspective of your acting self but to engage it in a species of influence that you will then, at the time of action, regard as appropriately rational. But it is hard to see how you could have this expectation of rational self-influence, given that you may also expect that your acting self will be out of deliberative accord with your influence. How can you expect that the influence will look appropriately rational while also expecting (as it may be) that the influence will look substantially wrongheaded? You may expect that your dessert-craving self will view your intention to diet as 'unfortunate' and perhaps 'unfair' as you face down the dessert tray. But as long as you don't let the disgruntled feelings lead you to reconsider whether to diet, your intention to diet will—you expect—nonetheless continue to strike you as rational. What is this nondeliberative species of rationality? And how can it be rational to expect that you will, at the time of action, achieve it?

(5) For reasons that I'll elaborate in section 5, I lack space to do more than gesture at a full explanation of how such intrapersonal influence could count

as subjectively rational from your perspective as you undergo it. What needs explaining first is the *claim* to rational authority that you make when you form an intention. In order to make this claim you must expect that your acting self will (or at least could) follow through not only without redeliberating but without thereby ceasing to maintain the species of subjective rationality at which intending, by its nature, aims. Again, intending in the face of expected temptation is not merely a matter of tying yourself to the mast. In attempting to understand how you could presume to exercise such intrapersonal influence, we must first understand what form of responsiveness to the influence is compatible with its claim to rational authority. As we've seen, there is reason to think the responsiveness cannot take the form of a deliberative disposition. But what is the alternative?

(6) A point about the connection between stability and rationality. Again, it is through being 'stable' in the present sense that an intention rationally resists being reconsidered or abandoned in the interval between formation and follow-through.[7] By 'rationally resists' I do not mean that there need be any psychological force making it difficult to redeliberate. The point is merely that a 'stable' intention preserves this rational status till you follow through on it or, by actually redeliberating, abandon it: even if you ought to redeliberate, and even if you would change your mind if you redeliberated (or would have changed it if you had redeliberated), it is nonetheless the case that from your own point of view then—that is, 'subjectively'—you ought to follow through on the intention. If that seems odd—well, again, that's the problem. How could your earlier deliberation control what's subjectively rational for you at *t* even when it would be (or would have been) overridden by a deliberation performed at or just before *t*?

(7) Elaborating points 4 and 5 in light of point 6, what's most immediately noteworthy is not so much the fact that an intention *is* stable as your need to expect that it *will* be as you form it. When you form an intention now to act at some time in the future, that's usually because you don't expect you'll be able to deliberate the matter at that future time, because you expect that the cost of waiting till then to deliberate will be too high, or because you expect that your deliberative perspective at that future time will be infected with a preference reversal that you do not now believe accords with what you'll all along have reason to do. Since forming an intention is compatible with expecting that you'll be thus tempted to abandon it before the time comes to act, it seems you must be able to expect that your intention will survive such a shift in preference. Or perhaps that's too strong; perhaps intending is compatible with not expecting that you'll succeed in resisting the temptation.[8] At the very least, you must be able to expect that if you do give in to temptation, that would constitute a departure from what is then subjectively rational for you. (Note how this formulation directly poses the problem that we're addressing: how could that be a departure from what is then subjectively rational for you?) There are thus two points of rational assessment: you must *rationally* expect that your intention will *rationally*

survive a shift in preference. As I've said, I lack space for a full treatment of that second 'rationally' in this paper, though I'll sketch part of an account in section 5. (For ease of formulation, I'll henceforth usually assume that you expect that you will follow through, not merely that you will be such that you subjectively ought to follow through.) Our focus will therefore be on your perspective as you form the intention. But our question about that perspective addresses its capacity to engage your perspective as you act on the intention.

Here's an overview of the argument that we'll pursue. Though the problem figures in the intrapersonal dimension of traditional questions about morality's relation to self-interest,[9] I'll motivate my approach with criticisms of more recent approaches that do not highlight any application to morality or broader norms. I'll begin, in section 2, with a critical discussion of Michael Bratman's and Richard Holton's recent treatments of the stability of intention, though the account I'll develop also borrows from Bratman's work. In section 3, I'll criticize and borrow from David Velleman's work on narrative and self-intelligibility. In section 4, I'll borrow from the psychologist Daniel Stern's work on the proto-narrative emotional relationship that he calls 'affect attunement'. In section 5, I'll clarify the dimension of subjective rationality in play and address broader questions of methodology. When the pieces fall into place, we'll see how the intrapersonal attitude at the core of intention aims at rational stability insofar as you thereby address your future self within a narrative structure articulating a species of self-relation modeled on the formative trust relations you once had with caregivers.

Here's a preview of the dialectic. We'll see how Holton and Bratman, while understanding the general problem perfectly well, nonetheless misconstrue the challenge you face as you form an intention. Holton's purely pragmatic approach overlooks important aspects of the intrapersonal relation in play and therefore requires too little. Bratman's metaphysical approach demands a kind of metaphysical authority over your follow-through and therefore requires too much. The claim of rational authority at the core of intention is best understood neither as a way of getting something you want nor as a way of making yourself metaphysically coherent. The key to understanding how we exert this species of rational influence over our future selves lies in tracing that influence to a species of intrapersonal relation that is neither pragmatic nor metaphysical but rhetorical.

By contrast with Holton's and Bratman's, my approach will emphasize the need for case-by-case self-intelligibility in intending: the need to understand what you're up to, and how you might plausibly succeed, each time you undertake the rational commitment at the core of intention.[10] When you form an intention, you project a sense-making perspective after—perhaps long after—the intended action is performed. Foreseeing that temptation may lead you astray, your intending self expects to move your acting self—at any stage of the unfolding teleological structure—by your ongoing sense that the story you're telling is believable. You foresee that when your tempted self considers this story,

it will feel more like scoffing than like complying. But that is not, you expect, how it will respond to the story. You expect it to look not only back but forward. You expect it to resonate not to the felt absurdity of your stage-directions at that moment but to the broader attractions of the play, attractions it feels by an empathetic projection onward to the horizon marked by plan's end.

2. Contra Bratman and Holton

Both Bratman and Holton defend their views via Gregory Kavka's Toxin Puzzle, so let's begin with that.[11] In Kavka's puzzle case, you have very good reason to form an intention though no reason whatsoever, it seems, to follow through on it. Imagine that you have good reason to expect that an eccentric billionaire will pay you a million dollars for forming, at midnight tonight, an intention to drink a certain toxin tomorrow at noon. The toxin will severely nauseate you for a time but leave you otherwise unharmed. If you had to drink to get the million dollars, you'd certainly do so, but the trick is that the billionaire does not require that you actually drink: all you need do is form, tonight, the intention to drink, and he will deposit the money in your bank account tomorrow morning when the bank opens. So at noon tomorrow you will have had three hours to reflect on the fact that you have no forward-looking reason to drink the toxin—that, if you redeliberated, you'd be crazy not to change your mind. Knowing this, you cannot form the intention. Though you could get the million dollars if the billionaire were rewarding your action of drinking, you cannot get it if he rewards only your formation of the intention to drink.

Consider first Bratman's treatment of the puzzle. Bratman argues that what distinguishes ordinary 'temptation' cases from 'Toxin' cases modeled on Kavka's is that you expect that you would regret abandoning your intention in the former but not in the latter.[12] You cannot form the intention to drink the toxin in Kavka's puzzle-case, despite expecting that you'd get a million dollars for forming that intention, because you also have to expect that, money in hand, you'd reasonably abandon the intention before the time came to drink. This change of mind would not merely manifest a temptation that your intention could if 'stronger' overcome. In a true temptation case, you expect you'd later regret having given in to the temptation to act contrary to your intention, but in the Toxin case you expect you would not. Moreover, in a temptation case you expect you would not regret having followed through on your intention, but in a Toxin case you expect you would. Bratman argues that this 'no-regret condition' explains why you can form the intention in the temptation case but not in the Toxin case.

That explanation seems plausible as far as it goes. Still, why should your regret matter to you in this way? How does this difference in what you expect to come later—at what Bratman calls *plan's end*: the point beyond which you will not change your attitude toward the action[13]—register in your ability to form the intention? Why should it be a necessary condition on forming an intention that you give any regard, even dispositionally,[14] to how you will view your present

action from plan's end? That is the specific question to which my appeal to narrative will provide an answer. Though I agree with Bratman's approach in broad outline, I think he goes wrong in some important details—mistakes that my emphasis on narrative can help correct. The main problem lies in how he formulates the no-regret condition.

Begin by observing that it is manifestly too strong to *require* that you expect not to regret an action that you intend to perform, since you know that any action may have unforeseen consequences. The expected regret must target not the action as such but your status as rational in performing it, given what you now expect will be its nature and consequences.[15] In light of this observation, Bratman's approach generates the proposal that we view this ascription of rationality to your acting self in terms of how you expect it will strike your self looking back from plan's end. We should therefore make the no-regret condition more specific: you must expect that your plan's-end self will not regret specifically the dimension of your self-relation manifested when you follow through on the intention.[16] To form an intention, you must expect that you'll not regret that relation of self-influence.[17]

The problem revealed by Toxin cases is that you cannot make the right sort of sense of what you're doing when you attempt to form an intention that you expect you'll later 'feel like an idiot' for having executed. That is a problem not of disappointment targeting the results of follow-through—by hypothesis, you wouldn't be disappointed—but of regret targeting the self-relation you would institute in following through. The breakdown posits three temporally distinct perspectives. Because you now expect that your plan's-end self will prove unable to make retrospective sense of how this action could be attributable to your acting self—'I can't believe I φed! How could I have trusted that intention?'—you cannot now make prospective sense of how you could, in forming the intention, attribute this action to your acting self. Your present perspective is linked in this way to your twice-future perspective because, as we'll see, the executive authority over your once-future acting self that you claim when you form an intention rests on an expectation not about how that acting self will view you but about how your twice-future self will view the executive self-relation. When you claim the authority, you expect that the relation whereby you execute it will continue to strike you—out to plan's end—as 'speaking for' you. You project that this retrospect will not be, as we may put it in shorthand, *self-attributively unsettling*.

Let that be step one of my objection: Bratman is not as explicit about the distinction between regret and disappointment as he should be.[18] Step two shows that we need to understand the role of regret partly in terms of narrative. Indeed, we're already in position to note a respect in which narrative will be crucial to the no-regret condition. When you intend, I build on Bratman in observing, you expect that your plan's-end self won't regret the aspect of your follow-through that we've just discussed. As we've seen, Bratman understands plan's end as the point beyond which you will not change your attitude toward the action.[19] But the observation, so understood, generates an obvious problem: you can intend

to φ at *t* despite expecting to regret your having φed at *t* as long as you don't expect that this regret will be fair.[20]

We can use a cartoon case to make the problem vivid. Say you're planning to join a cult that will, among other things, 'reprogram' you into hating your parents. This isn't the reason you're joining the cult, but you can foresee that it will have this effect. Still, you do now love your parents and want to spend some time with them to make vivid your love for them, expecting that you'll soon thereafter come to hate them and that the timing of your changed attitude will help them to see that the change is not their fault. So you form an intention to spend a week with them in which you make your love for them as clear as you can, expecting that you'll soon thereafter adopt a stable attitude of regretting having followed through on that intention. The fact that you expect that you will have this regret at plan's end—as Bratman defines it—obviously has no bearing either on your ability to claim executive authority when you form the intention or on the stability of your intention once you form it. But that's because your expectations of permanent regret do not target the narrative arc of the story that you're telling as you form the intention. The audience for *that* story is not your cult-addled future self but a self that stands beyond, or at any rate outside, that future. When you form the intention, you're thinking of plan's end as a perspective that can resonate to the 'poignancy' of this 'tragic' narrative: 'giving my parents their due before I take steps that will render me unable to appreciate what I owe them'. That story can count as well-told only to a future self of yours that *is* able to appreciate what you owe your parents—by hypothesis, not the self that you expect that you will become and forever remain.

Plan's end cannot, therefore, simply be the point beyond which you will not change your attitude toward the action. So the idea of 'plan's end' cannot figure quite as Bratman proposed. But we can nonetheless make use of his idea, by viewing plan's end as a notional perspective internal to the narrative you posit when you intend. We'll see that it is not a perspective within the story but the perspective—often not discrete but ongoing—from which the story is to be appreciated. This will enable us to understand why the story must be not merely thought but *addressed*. How you're committing yourself depends in part on how you conceive of the self or selves to whom you're addressing your story.[21] I'll explain the stability of intention by explaining the nature of this projection. As we'll see, the projection explains how your intention can manifest a rational expectation that you will resist shifts in preference that would otherwise—that is, without the intention—rationally require that you redeliberate. (I'll say more about this plan's-end perspective at the end of section 3.)

This refinement of Bratman's no-regret condition enables us to see where Holton's approach goes wrong. Holton presents his argument as a defense of 'resolute choice'.[22] But unlike David Gauthier's well-known defense of that doctrine, Holton's does not yield the implausible result that you should persist in your intention to drink the toxin even if you redeliberate between 9 a.m. and noon tomorrow.[23] Holton argues only that you should not reopen deliberation, if you can count on the rationality of your disposition to follow through, not that

you should deliberatively reaffirm the intention tomorrow morning. He thus agrees with Bratman that you cannot form the intention in a one-off Toxin Case, though he argues that you could if Toxin Cases were common. If Toxin Cases were common—Holton asks us to '[s]uppose that, for his own mysterious ends, a perverse god arranged things so that the necessities of life were distributed to those who intended to endure subsequent (and by then pointless) suffering'[24]—it would be rational for you to inculcate in yourself and in others whom you want to enrich (say, your children) a disposition not to reconsider one's intention in such cases. But if it is rational for you to have a disposition not to reconsider your intention in such a case, Holton argues, then you can rationally expect that you'll follow through on your intention in such a case—even if it would not be rational for you to deliberatively reaffirm that intention in the hours before you follow through on it.

Though this is an improvement on Gauthier's position, there is nonetheless a problem for Holton: how can you regard as rational the disposition to follow through on such an intention *when you can see so clearly that you will regret doing so*? As we have seen, when you form an intention you must expect not to regret following through on it *even when* you expect not to be disappointed by the results of following through. And we are not imagining that you are too dumb or distracted tomorrow morning to notice that the payoff has been securely deposited in your bank account—and thus that nothing is to be gained by making yourself sick.[25] Making the Toxin puzzle a common occurrence does not change this feature of Kavka's one-off version. In each case, you expect you won't feel disappointed by the results: sure, you'll be sick, but you'll reasonably believe that that was a way of getting a million dollars that you'll be very happy to have secured. In each case, you may well not be disappointed with the results while nonetheless 'feeling like an idiot', as you'll naturally put it, for having followed through and made yourself sick when there was no need to. Even if you regard your possession of the disposition to follow through as a good thing because you regularly confront such cases, you have to concede that you could have costlessly—indeed, with a vivid benefit—reconsidered *in this instance*, thereby keeping the money and avoiding illness. That will, you can now see, be the subjectively rational thing to have done: come morning realize that you already have the money, let that realization reopen deliberation, and decide that there's now no reason to drink. 'Feeling like an idiot' for following through marks a kind of *regret* at following through. Since you can see that you will regret following through, you cannot form the intention to drink—even though you reasonably expect you would not be disappointed with the results of drinking. But, again, that is how we might also describe a one-off Toxin case. Making Toxin cases common has not helped.

On my alternative account, the moral of Kavka's Toxin Puzzle is that you cannot tell yourself the 'story' that you'll drink the toxin unless you expect that you'll continue to find it believable—not necessarily that you *will* follow through but that it will make sense to.[26] This recasts in terms of narrative the observation that you cannot form an intention unless you expect that it will have rational

bearing on your conduct when the time comes to act, such that you may follow through on the intention without redeliberating. Again, that expectation codifies the claim of executive authority at the heart of intention. We'll return to that claim, and to the question how the present account compares with Bratman's, in section 5. I am focusing on the stability of intention, and my thesis concerns how a claim of authority must shape the agent's expectations in forming the intention. We're beginning to see that the claim of authority is fundamentally rhetorical, with an expectation or projection of influence revealingly like that of an effective storyteller over his or her audience.

3. How Your Intention Makes Narrative Sense of What You're Doing

Before we consider this mode of address more fully, let's consider how it projects a sense-making perspective. How does a story make sense of things? David Velleman argues that a story makes specifically emotional sense of things, by 'enabl[ing] its audience to assimilate events, not to familiar patterns of *how things happen*, but rather to familiar patterns of *how things feel*'.[27] Of a well-told story's ending, he writes:

> [T]he emotion that resolves a narrative sequence tends to subsume the emotions that preceded it. . . . Hence the conclusory emotion in a narrative sequence embodies not just how the audience feels about the ending; it embodies how the audience feels, at the ending, about the whole story. Having passed through the emotional ups and downs of the story, as one event succeeded another, the audience comes to rest in a stable attitude about the series of events in its entirety.[28]

I'll accept this account of narrative understanding and proceed by explaining how the puzzle of diachronic integrity described in the introduction above leads me to dissent from the use to which Velleman puts the account in his broader theory of agency.[29] I'll develop my approach further in section 4, where I'll specify a role for a regret-like attitude modeled on a species of alienation typical of 'unbelievable' storytelling. I am not putting forth an analytic thesis about the nature of narrative, or of narrative understanding, so we need not worry whether Velleman's account admits of peripheral counterexamples. Our interest lies in the species of understanding that he describes and in the claim that we are familiar with it not only through our emotions but through our experience of narrative.

We'll work from this core example. As Velleman notes, one of the best ways to begin to break a smoking habit is to imagine yourself, falsely, as a nonsmoker—say, at the party you'll attend tonight where you know it will be hard to avoid cigarettes.[30] This gives you a motive to resist temptations to smoke, Velleman argues, grounded in your more fundamental desire to make sense of what you're doing: if you're a nonsmoker, the intelligible thing is not to smoke. That appeal to fiction is not yet quite an appeal to narrative, since it makes no

essential use of the idea of a *story*. The narrative proper figures only in the background, explaining the nature of your psychological investment in the fictional self-description—like an actor onstage, you get caught up in the game of make-believe. Velleman hypothesizes that like all of us you are independently motivated to do what would allow you to understand yourself in light of your self-conception,[31] and an appeal to narrative *could* explain how your self-conception includes a self-description—for example, that you are a nonsmoker.

We must say 'could' here because Velleman does not actually offer this explanation. Simply sidestepping our puzzle of diachronic integrity, he does not explain what it is to be invested in the fiction.[32] We must concede that making sense of yourself tonight *given* your imaginative project would not itself be an instance of grasping how things feel as opposed to how they happen. There you stand at the party thinking your nonsmokerish thoughts, wearing an anti-smoking T-shirt, and nodding sympathetically to your interlocutor's rant about the filthy smokers with whom he works. Against that social and psychological background it would be causally unintelligible if you suddenly lit up a cigarette—that could only be a joke, but you are not inclined to humor about this matter. So *this* self-understanding is causal-psychological. Still, what about the fictional self-understanding—the taking-yourself-to-be-a-nonsmoker—figuring in the background? You know you *are* a smoker. You're quietly rocking from foot to foot craving a cigarette! You know that if you let yourself think about it, you'd find this puritan's rant against his coworkers appalling. And so on. In what respect, then, do you understand yourself as a nonsmoker?

This aspect of your self-conception appears to involve the very sort of narrative understanding that Velleman theorizes as an emotional grasp of how things feel. His appeal to a 'fictional' self-understanding requires confronting our puzzle of diachronic integrity head on. Despite Velleman's silence on the question, then, we can use his theory of narrative understanding to explain the background that enables you to make causal-psychological sense of yourself as not smoking despite knowing that you smoke. There is no reason to think that causal-psychological sense-making must always presuppose narrative sense-making. But in the everyday example that we're considering, we'll naturally hypothesize that your ability to make causal sense of yourself as not smoking depends on your ability to make narrative sense of yourself as a nonsmoker. To make narrative sense of yourself as a nonsmoker, in this case, you have to resonate emotionally to the narrative arc of a story that would cast you in that role. This is a nontrivial requirement. It may be that aspects of your psychology prevent you from resonating to such a story. Your craving for cigarettes may, for example, grow severe enough that it undermines your imaginative project. Velleman draws an analogy with improvisational method acting, and we might compare your predicament with that of a method actor who has difficulty depicting a starving peasant after he—the actor, not the character—has ingested too large a mid-day meal. The problem isn't that he cannot make causal sense of the burps and belly-based lethargy given that he's a starving peasant. The problem is that he cannot make emotional sense of a story casting him as a

starving peasant given the burps and belly-based lethargy. That's why a method actor takes steps to 'prepare' for a role: in part to make it believable to the actor himself that he inhabits the role. This looks like Velleman's species of narrative sense-making rather than causal-psychological sense-making. Causal-psychological sense-making in this context—doing what it would make sense for a starving peasant to do—presupposes a capacity for such narrative sense-making. Again, I don't claim that it must always do so, merely that when causal-psychological sense-making does presuppose narrative sense-making in this way we get an explanatory appeal to narrative that, unlike Velleman's explicit account, resolves the puzzle of diachronic integrity.

There is another instructive puzzle, however, in this application of Velleman's theory of narrative understanding. If acting involves resonating to a narrative arc, we should be puzzled how that can happen when you already know the ending. A Broadway actor might perform the same role every evening and twice on Sundays for years: how is he supposed to resonate directly to this narrative arc? The arc on its own must before long leave him cold.[33] Drama is unlike music in this respect. Your ability to hear the resolution of tension in a passage of music does not depend on curiosity how it will resolve: whatever exactly this involves, you can hear musical tension and resolution more directly. You can be absorbed in a piece of music played over and over many times a day for many days. But if you watch a film or read a story over and over like that—and of course people do—it will probably prevent you from remaining simply absorbed in the plot. Perhaps you like the novelistic or filmic technique, or the actors or the sweep of the prose. Or perhaps you enjoy remembering how it felt to be caught up in that narrative arc the first time you encountered the work, and your imagination breathes life into the work through that captivating memory. To give his lines life, the Broadway actor must resonate to something rather like that: if not to the perspective of an audience ever renewing itself, then to the memory of his own first encounter with the work. This lesson draws on aspects of the second and third puzzles from which we began. The moral is that grasping the diachronic integrity of a narrative implicitly invokes the perspective of an addressee—that is, of one whom we can imagine as emotionally invested in the story's narrative arc.

I propose that we generalize the observation and embrace this hypothesis: that the difference between merely following stage directions and bringing the play to life lies in a capacity to imagine how your performance would resonate with someone curious how it will end. I agree with Velleman that such narrative understanding is primarily emotional: the way to imagine such curiosity is to imagine how it would feel.[34] You don't bring a play to life in performance by imagining merely *that there is* an audience curious how it will end. You do so by imagining yourself *feeling* that curiosity. Imagining yourself feeling curious how the play will end is imagining yourself resonating to the play's narrative arc in Velleman's sense. Since you have to grasp how things feel in this dimension in order to imagine feeling them, bringing the play to life in performance requires (inter alia) narrative understanding in Velleman's sense—narrative understanding specifically from the perspective of your audience.

Returning from the explicitly dramatic case to the everyday, note that your self at plan's end defines such a perspective—as you look onward from your perspective in forming the intention. I'll elaborate this intrapersonal case in section 4, drawing on resources introduced there, but let me first fend off an objection. One might wonder how there could be any analogy between the interpersonal and the intrapersonal case, given that your plan's-end self already knows how the story will end. That would overlook a basic fact about these relata: each is simply, at a different time, *you*. Your plan's-end self is simply you in position to make retrospective sense of things that you are now proposing to do. Of course, you are not now proposing to do these things in a spirit of curiosity how things will turn out. You are in the business of doing things, not merely of observing while things get done. But your capacity to do things by following through on an intention depends on your capacity to project a perspective from which a key portion of what you're bringing about will be observed: the portion that consists of the self-relation that you instantiate when you follow through. 'Does this self-relation speak for me?' becomes 'Will the narrative arc that it traces make narrative sense to me as I look back on it from plan's end?'[35] Narrative sense-making satisfies curiosity, but Velleman is right that the role played by curiosity in narrative sense-making is primarily emotional.

Let me draw this moral more directly. The story you're telling as you intend and follow-through is a story of self-influence, and the question is how you expect—or fictionally project[36]—you'll come to feel about that. It is a question about your emotions at plan's end, but it is also a question about your emotions as you project this perspective. Getting caught up in the story as you enact it manifests an emotional understanding of its arc, and the question is how you'll later feel about the arc thus understood. A question of your emotional self-understanding arises at three distinct points. Your retrospective self-understanding at plan's end targets your self-understanding at the time of action, and the entire structure plays its role in constituting the stability of intention through your prospective self-understanding as you intend—where all three instances of self-understanding are emotional in the way that Velleman describes. The relation linking these self-understandings is not, as we'll now see, a mimetic relation but the internalization of an important way in which we employ emotions nonmimetically in our understandings of others.

4. Narrative Sense-Making as Emotional Influence

What second-personal relation could be internalized to function thus intrapersonally?[37] I propose we model this intrapersonal relation on the type of interpersonal relation that the developmental psychologist Daniel Stern has theorized under the label 'affect attunement'.[38]

Though we've been discussing a possible schism between intending self and plan's-end self, we need an account of intention on which the transition between these perspectives is normally seamless. When you follow through on an

intention to φ at *t*, you normally act at *t* without further thought, and you don't normally feel any pressure to consider the matter further. It's only when things don't go as you expected when you formed the intention that you *need* to conceptualize the sequence as an intrapersonal relation. One beautiful thing about affect attunement is that when it works it is seamless in just this respect. It can be observed only when it breaks down.

Here is what Stern calls the 'characteristic episode':

> Baby A, a nine-month-old crawler, crawls away from Mom and over to a new toy. While on his stomach he grabs the toy and begins to play with it. His play is animated, as judged by his movements, breathing, and vocalizations. Mother then approaches him from behind and puts her hand on his bottom and gives it an animated jiggle side to side. The speed and intensity of her jiggle appear to match well these aspects of the infant's behavior, qualifying this as an attunement. The infant's response to her attunement is nothing. He simply continues to play without dropping a stitch. Her jiggle left no overt trace, as if she had never acted.[39]

But if the jiggle leaves no trace, how do we detect its influence? Simply by making its influence less effective:

> To create the first perturbation, the mother was instructed to do exactly the same as always, except to purposely 'misjudge' her baby's level of joyful animation, to pretend that the baby is somewhat less excited than he appears to be, and to jiggle accordingly. When the mother did jiggle somewhat more slowly and less intensely than she judged would make a good match, the baby quickly stopped playing and looked around at her, as if to say, 'What's going on?' This perturbation was repeated twice with the same results. The second perturbation was in the opposite direction. The mother was to pretend that her baby was at a higher level of joyful animation and jiggle accordingly. The results were the same: the baby stopped and looked around. The mother was then asked to jiggle appropriately as she originally did, and again the infant did not respond.[40]

When the mother does not intervene at all, by contrast, the baby likewise tends soon to stop and look back—from anxiety, as it seems, that she simply is not there. The nonresponse characteristic of attunement seems therefore to be playing an important role in the baby's ongoing activity. The activity itself functions as the response, with the mother now experienced as sharing its affective contour.

Stern argues that these interpersonal influences—influences that unless impeded leave no outward trace—are what enable the developing infant to come to represent an interpersonal world, thereby enabling the growing child to feel connected to others while no longer physically in their presence. My aim is less ambitious. I need merely borrow two ideas: first, that there could be a form of

inter- or intra-personal influence that registers behaviorally only when it breaks down; second, that this form of influence involves what Stern calls the 'matching' of 'vitality affects' that occurs in affect attunement.[41] As Stern views them, vitality affects are emotions displaying the 'arc' that Velleman argues lies at the core of narrative explanation: the experience of a vitality affect is distinguished by its tension and resolution.[42] The possibility of affect attunement can thus help us see how a robust interpersonal influence might be seamless in the way that the influence of present over future self is seamless in intentional agency. It also helps us see how these relations might be mediated by emotions with a narrative arc.

On Stern's model, when the caregiver's vitality affect 'matches' the infant's vitality affect—matching, say, an eyebrow movement to the baby's vocalization or a drawn-out 'Hellooo!' to the baby's squirms—it communicates an influence that transforms something that the infant would be doing on its own into something that they're doing together. The matching may though needn't be cross-modal. What is crucial is that it *not be merely imitative*—for the commonsensical reason that, while imitation is one response to what a partner is doing, it usually amounts to a refusal to act with the partner. Real attunement has three aspects or stages:

> First, the parent must be able to read the infant's feeling state from the infant's overt behavior. Second, the parent must perform some behavior that is not a strict imitation but nonetheless corresponds in some way to the infant's overt behavior. Third, the infant must be able to read this corresponding parental response as having to do with the infant's own original feeling experience and not just imitating the infant's behavior.[43]

The 'correspondence' at stage two is the matching. The claims that the caregiver's behavior is not imitative and that the infant does not experience it as imitative—the claims at stages two and three—entail that this three-stage process aims at exerting an *influence* on the infant, not merely at 'mirroring'. Having through affect attunement achieved a sense of joint activity, the pair can articulate their shared project in new ways, taking turns exerting an influence and being influenced. The caregiver can thereby guide the infant toward new experiences—but only by showing a receptivity, through attunement, to the infant's influence in turn.

So how would this work in a one-person case? Affect attunement provides a mechanism for teaching an infant to experience her expressed emotions as aiming to engage an audience that would 'complete' them by providing a 'matching' response. The lesson consists in the fact that that is what the adult caregiver *does* provide, naturally treating the infant's grunting, wriggling, or cooing as if it had that aim. A form of communication based in emotion thus falls in place, with each side capable of playing the role of initiator or of respondent. The infant comes to experience her expressions as completing her partner's as much as her partner's complete her own and thereby learns a kind of interpretive self-care—not, or course, how to meet her own demands but how to make

emotional sense of their behavioral expression. As the growing child gradually learns to replicate this structure within herself, I hypothesize, she'll keep looking for an audience for her vitality affects—for her 'stories'—but be willing to make do with an audience in her own projected future. This purely projected audience plays the role that the caregiver played in more basic engagements; it is crucially not there simply to replicate or mimic the content of the attitude that addresses it. The story told and the story understood—attitude and response—are not merely the same thing repeated. As in the interpersonal case, the response completes the story by providing emotional closure to its narrative arc considered as a whole.

Here we reach the core of my account. As the stage actor aims to make sense of each action in light of the narrative arc that will resonate with his audience at play's end, so the child learns to make her intentions resonate to how she will experience a narrative arc at plan's end. If she achieves this aim, nothing happens; no actual response is called for, since unlike the audience of a play she has all along inhabited that intentional structure. We see an actual response only when the intention fails in its narrative aim, breaking the spell that made it stable. That failure is like initiating an affect attunement that doesn't materialize. The stability of intention does not require actual attunement with your self at plan's end. But the requirement that you expect this self to attune imposes a significant constraint on your ability to form intentions, and can provide sui generis motivation to follow through on an intention once formed. The motivation consists simply in your anticipation of the expected attunement.

In the one-person case on which we're focusing, 'attunement' takes the form of your not feeling 'self-attributively unsettled'—as I put it in section 2—by the memory of your having followed through on the intention.[44] Precisely how does this resemble the interpersonal attunement that Stern describes? We can assimilate our kind of case to his by making three observations. First, note that an affect attunement between child and caregiver is not restricted to the moment of actual interaction between them. The importance of such attunements on both sides derives from how they continue in memory, giving the attuned-with partner an ongoing sense of connection with the attuning partner. (The attunement will typically be reciprocated but needn't be, as in the 'characteristic episode' above.) Second, note one key respect in which the connection is ongoing: the attuned-with partner will consciously represent the relationship only when the attunement breaks down. (This is the above-noted 'seamlessness'.) Third, note the role played by identity in the relation: it is just *this* person—in basic cases, a parent or other important caregiver—to whom one is ongoingly connected. Should the infant in the characteristic episode glance back and see a stranger instead of his mother, even a perfect 'match' in purely affective attunement would be for naught.

In the one-person case, the attuning partner can engage the attuned-with partner only in memory, since these are simply the same person at different times. But that need not impede the key dimension of attunement, which is the ongoing sense of connection. To clarify the parallel, let's model a one-person

case—crudely—on Stern's characteristic episode. The role of attuning mother is played by the intending self, and the role of attuned-with infant is played by the self at plan's end. The intending self is responsive to how it imagines the plan's-end self will respond to its influence over the acting self. Recall that plan's end is the point—within the projection that informs the intention—at which the disposition to remember this influence will fade. So the question for the intending self is how that memory will figure in the ongoing life the plan's-end self is leading. Will that future self—that is, you—respond to the memory disruptively, feeling self-dissociated by the observation that *that* is how you yourself anticipated your future, expecting the action to continue to speak for you? Or will your response, by contrast, be—nothing? That *nothing* is simply you getting on with your life, remembering the intention and action as how the story of your past once spoke for you and still ongoingly does. On the latter disjunct, you're not unsettled by the memory of how you once projected the perspective of this very memory. That projected you feels like—*you*.

5. Intending as Inviting Self-Trust

As we've begun to see, the core of my account takes a distinctive view of the rationality of letting your present deliberative perspective be overridden by an earlier perspective whose verdict you would not reaffirm if you now redeliberated. But the account needs elaboration. It is unclear how it *could* be rational for you to follow through on an intention that you would not redeliberatively affirm. Though, as I've said, I cannot provide a full explanation here, we need an answer to this question: under what conditions is it rational to follow through on an intention that you would not thus reaffirm? We need to have at least some understanding of those conditions in order to have any understanding of this paper's explanandum: what it is for the expectation of rational follow-through itself to be rational. The rationality of the expectation reduces to the rationality of the belief that those conditions will be met. Even if I cannot give a full explanation of those conditions here, we must understand them well enough to understand how a belief that they have been met could be rational.[45]

The elaboration that we'll now pursue provides an abstract characterization of the *kind* of account given in the previous two sections. As we'll see, the rhetorical self-relations that I've emphasized realize the more abstract pattern of a self-trust relation: they amount to the specific way in which we realize self-relations that adjudicate a dynamic of self-trust.[46] I could have pursued the more abstract issue first, and only afterwards filled in the details. But it is difficult to appreciate the self-trust dynamic in the abstract, especially since there are other questions of self-trust in play (concerning the relation of judgment to intention, doxastic self-trust, etc.), and it is easy to get lost in all that complexity. My emphasis on emotional development and rhetorical engagement marks an aspiration to ground my account in an empirical understanding of what human beings are

actually like. My emphasis on specific self-relations—on your attitude toward your future selves, and specifically toward their responsiveness to that attitude—fundamentally differs from Holton's emphasis on the strength of your disposition to follow through in the *kind* of case you're in. And, as we'll see presently, it fundamentally differs from Bratman's emphasis on coherence relations between your intention and *more general* policy-like commitments that unify all these attitudes into a single self-governing point of view. But now that I've presented a detailed picture of these self-relations, it is time to step back and say more about the kind of solution it is. That will more directly engage the question of rationality at the core of intention stability, commensurating my account more clearly with Holton's and Bratman's. And it will help pose broader questions, which we'll lack space to pursue, about how other beings—nonhumans capable of diachronic agency, or a subset of our fellow humans less capable of empathetic emotions—might solve the problem of intention stability in a different way than we characteristically do.[47]

We can approach the more abstract orientation of my account by contrasting it with a deeper dimension in Bratman's account of his no-regret condition. Though Bratman at first emphasized a pragmatic need for diachronic stability in intention, he more recently grounds the no-regret condition in a broader account of *agential authority*.[48] An account of agential authority would explain a specific dimension of your authority over your action, when you have it: it would explain what makes it the case that you are in charge (rather than, say, a force within you).[49] The problem is that agents appear capable of forming and acting from stable intentions even when racked with Lockean-coherence-undermining ambivalence. We can imagine an unwilling addict—a Frankfurtian wanton—with stable intentions, as long as we imagine him sufficiently invested in his addiction that he projects a plan's-end self who would not regret his addiction-addled plans. This will probably require self-deception, but self-deception is not necessarily at odds with the stability of intention. In fact, as we've seen in section 3, the stability of intention sometimes requires an interesting species of self-deception.[50] We appear not to be engaging the question of attributability that Bratman's account treats as paramount.[51]

Bratman's emphasis on agential authority—on being in charge of what you do in a way that would make the action attributable to you rather than to forces within you—builds too much into the stability of intention and overlooks cases in which ambivalence or some other agential incoherence may prevent you from living up to the claim of authority that you make when you commit yourself in this way. Generalizing the counterexample in the previous paragraph, it seems that we all sometimes form intentions or resolutions that are at odds with policies in which we are at the same time deeply invested. After all, an intention or resolution can express anger or cowardice with which you do not identify yourself, even if you do not suffer from addictions that tie you in attributability-undermining knots of ambivalence. A passionately self-destructive resolution is still a resolution, as is an ambivalent one. A stable intention or resolution, or the expectation thereof, need not speak for 'who you are' as an agent in any

interesting sense. It expresses attitudes that you have that may be at odds with other attitudes that you have, even right then and there. Even if there is no such conflict, your attitudes need not remain consistent, nor need you expect them to remain consistent, stretching out to plan's end.[52]

How, then, can my account vindicate the rationality of expecting that you will be rational in letting your intention override your deliberative perspective as you act? Here is where my approach diverges most sharply from Bratman's. As we've just seen, Bratman explains stability in terms of attributability: it is rational to let the force of your intention override your deliberative perspective as you act just when doing so satisfies the conditions sufficient for making that action attributable to you. My account, by contrast, does not appeal to attributability in its explanation, not merely because there can be stable intentions without attributable agency (e.g., an unwilling but planning addict) but because I doubt there is any 'metaphysical imperative' to maintain your identity through time, or to do only what is attributable to you.[53] It can be rational to let your intention override your deliberative perspective when you act, on my account, because it can be rational to let your dispositions to act be informed by your sense that the 'part' you're 'playing' is 'believable' in the way I've been articulating. My account of how rationality figures in stability thus builds on my account of the role of believability in stability.

How does the believability of a narrative help make it rational to let yourself be influenced by it in this way? Let me first emphasize that my appeal to the narrative's believability is not an appeal to its being *worthy* of belief. I am not appealing to an extrinsic value but to a value that is fundamentally internal to your ongoing project of attempting to influence your later selves in appropriate ways and then to let yourself be thus influenced only when the influence is indeed appropriate—where the criterion of appropriateness lies in your ongoing sense of what will make retrospective narrative sense from the perspective that your intention posits at plan's end. What is believable is just what you expect you will find believable at plan's end. This ongoing project plays the role played by Lockean planning coherence in Bratman's account. It yields a species of genuine executive authority, but it is not the sort of authority that would allow us to draw a sharp distinction between your presence qua agent in your behavior and other influences. As I've suggested, I am skeptical whether we need that distinction to ground the practices of attributability and accountability that give content to our understanding of human action. In any case, I am not attempting to give such an account here. Even if there is an ultimate need for such an account, I do not believe that it need play a role in an account of the stability of intention.[54]

I said that I'm appealing to a value that is 'fundamentally' internal to this project of self-influence. To say that it is fundamentally internal is not, however, to say that it is entirely internal. Recall what we observed in section 2 about the role of fairness in the backward looking self-relation: what you expect is that it would not be fair to look back from plan's end with regret—that is, as we're filling in the details, that it would not be an appropriate response to find the

narrative unbelievable. When you form the intention you project a narrative internal to which is a perspective to which the narrative is addressed. Your expectation that following through on the intention will be rational includes an expectation that the addressee posited by the narrative will continue to strike you as providing an appropriately 'fair' response. The perspective of this addressee is modeled on the perspective that you expect to occupy at plan's end; I've therefore followed Bratman in calling it your 'plan's-end' perspective. But the expectation informing your intention addresses how the narrative, as so addressed, will continue to resonate with you, not necessarily how you expect any future self of yours to respond. That's my rhetorical formulation of the presupposition that Bratman formulates in metaphysical terms as the expectation that the network of self-governing plans and policies informing your intention should be sustained to plan's end. What we must say, amending Bratman's appeal to plan's end, is that your project of self-governance anticipates not how you will actually respond to the self-influence but how it will make sense to respond within the narrative structure that codifies how you think it would be fair for your future self to respond to the self-influence.[55]

What if, when the time comes to act, you cease to project an addressee responding as you projected it would respond when you formed the intention? Again, as the Toxin Puzzle reveals, that is precisely what you must *expect* will *not* happen if you are to form the intention. Still, what would it show if—falsifying your expectation—it did happen? It might show merely that you've forgotten what you intended or that you've changed your mind in response to information that you did not foresee as you formed the intention. But one other thing it might show is that you do not trust your intention-forming self: that the rhetorical stance toward your future selves that you adopted when you formed the intention no longer strikes you as trustworthy. It follows that your intention includes an expectation not merely that you will continue to be moved by the narrative that it projects but that being thus moved will amount to undergoing an influence that will prove to be worthy of your trust. At the core of the rhetorical self-relation that you project lies a relation of substantive self-trust.

The rationality of the expectation informing your intention must be grounded in the rationality of the self-trust relations that it projects. Again, it isn't a problem that the projection may prove wrong; an expectation may be rational while proving incorrect. The rationality of the projected self-trust relation itself in turn lies in its projection of a plan's-end perspective from which it will emerge as retrospectively rational, where to say that it is retrospectively rational is to say that it is not self-attributively unsettling in the way we've considered. The question is not whether you will self-attribute the self-trust relation but whether that self-attribution will resonate with you in right way. The way you must expect it will resonate is the way that a well-told story about you resonates with you when you are its audience: the self-attribution ('I did that') must not be in affective tension with your narrative identification with its protagonist ('that's a part that I continue to see myself playing'). The test of rationality is thus in two respects primarily emotional: what it is to play this part lies in the emotional

intelligibility of a narrative, and what it takes for the narrative to resonate with your plan's-end self in the way you expect lies in its emotional compatibility with your self-attribution of the self-trust relation whereby you performed the action.

My approach thus disagrees with Bratman's over how to explain a dimension of 'subjective' rationality. On my trust-based approach, we can say that subjective rationality is rationality from the perspective of the subject's attitudes without assuming that those attitudes amount to a coherent perspective. What matters for rationality, we can say, is not coherence but the intrapersonal relations among the subject's judging, intending and acting selves.[56] Each transition—from judging, all things considered, that you ought to φ at *t*, to intending to φ at *t*, to following through on that intention and φing at *t*—marks an intrapersonal relation of self-trust: you intend to φ at *t* because you trust your judgment that you ought to φ at *t*, and you follow through on the intention without redeliberating because you trust the earlier self of yours that formed the intention. At each transition, you might go akratic through a failure to trust yourself: you fail to intend to φ at *t* without abandoning your all-things-considered judgment that you ought to φ at *t*, or you fail to follow through and φ at *t* without reconsidering your intention to φ at *t*. Avoiding akrasia isn't something you do through mere force of will. As we've seen in our discussion of the Toxin Puzzle, when you form the intention to φ at *t* you don't merely expect that you *will* φ at *t* but that it would be rational to do so. Indeed, you expect that you will not φ at *t* if at *t* if you do not trust the self of yours that formed the intention—that is, your current self as you form and maintain the expectation. Because of the nondeliberative nature of following through on an intention, you do not expect to have to convince your acting self of your retrospective trustworthiness in having formed the intention. But you do expect that if there is evidence of your untrustworthiness you will refrain from following through. After all, you don't want brutely to force your future self to φ at *t* but to guide it to φ at *t* through the force of your trustworthiness in forming the intention.

These intrapersonal relations can manifest subjective rationality because they are trust-relations, and because the attitude of trust manifests a rational responsiveness. When you act on the basis of your trust in someone—when, say, the trusted takes you by the arm and you 'follow her lead'—her influence on you ensures that you are not acting only for reasons that emerge from your own deliberative perspective, but also that you are not merely letting yourself be influenced by the trusted. You are letting yourself be influenced by her, but in a way that is governed by a counterfactual sensitivity to evidence of untrustworthiness in the trusted. It's a sensitivity to evidence of untrustworthiness, rather than to evidence of trustworthiness, because you don't need to assess someone for trustworthiness in order to trust her; you merely need to be sensitive to evidence of untrustworthiness, should any emerge. And the sensitivity must be counterfactual: had any emerged, you would not have trusted. The influence does not reduce to brute reliance, which can be strategic (in all

sorts of ways) and therefore needn't amount to trust. But the influence also needn't involve active monitoring for trustworthiness, since the need for trust often arises when relevant evidence of trustworthiness is not available. The counterfactual sensitivity to evidence of untrustworthiness at the core of trust ensures that the undeliberated nature of the influence that you undergo in trusting does not on its own make the trust irrational.

To say that at the core of this rational influence lies a counterfactual sensitivity to evidence of untrustworthiness does not, of course, explain how the sensitivity works. Sensitivity to precisely what? To untrustworthiness, but in precisely which respects? And sensitivity precisely how? If there is evidence—or enough evidence—of untrustworthiness, the subject will cease to trust, but what precisely does this sensitivity involve? The account of the stability of intention that I've given attempts to explain how the counterfactual sensitivity governs the rational transition from intending to φ at *t* to following through on that intention and φing at *t*. (I have not addressed the relation between judging that you ought to φ at *t* and intending to φ at *t*.[57]) I've argued that the sensitivity figures in your sense of the 'believability' of the narrative structure that I've argued lies at the core of your intention. You are sensitive to the trustworthiness—that is, to the believability—of the story that you project for your future selves as you form the intention and the sensitivity takes the form of your disposition to reconsider the intention if you don't find this narrative structure believable. As I've explained, the respect in which you will or will not find it relevantly believable lies in your ability to project a plan's-end self that will not be self-attributively unsettled by the action. Transposed into the idiom of self-trust, the question is whether you expect that your plan's-end self will regret the trust-relation.

Such a trust-relation emerges from the invitation to trust that I have explained as a piece of rhetoric. We need the concept of narrative to account for the rationality in these self-trust relations because the relations begin from what amount to an invitation to trust, and because the invitation posits a species of influence that is not brute reliance but depends on the counterfactual sensitivity that I've described: it depends on the invitation's being believable as a narrative addressed to an audience at plan's end. As I've just explained, these self-relations are fundamental to the stability of intention because they manifest your capacity to undertake the rational transition from intending to φ at *t* to following through on that intention and φing at *t*. Again, the key to understanding this transition is to see that it is no less rational for being undeliberated. Instead of redeliberating whether to φ at *t*, you simply follow through on your intention. Though you are not weighing or reweighing reasons for or against φing at *t*, the transition is rational insofar as your follow-through manifests the counterfactual sensitivity that I've explained as your ongoing sense that the story you're telling is believable.

The claim of executive authority over your conduct that defines an intention is thus at bottom this piece of rhetoric: you're telling yourself what amounts to a story about how to act, a story that you expect will move you to act because,

within the story, you will continue to find it satisfying—that is, not self-attributively unsettling—when your twice-future self looks back from plan's end. Your orientation is prospective, but your rhetoric addresses the perspective that will emerge as you ask, retrospectively, how to feel about this self-influence at plan's end. To form your intention you must project an audience that will not be attributively unsettled by the self-relations—by the fact that the protagonist of this story is *you, thus invested*. That is part of what it *is* to be invested in the story: to project such an audience. It is your continuing investment in the story, your continuing projection of this audience, that enables you—that is, rationally enables you—to resist temptation. And it is your expectation of such investment, strikingly absent in Toxin cases, that rationally enables you to move from judging that you ought to φ at *t* to intending to φ at *t*.

The investment may seem like a trick you play on yourself. But in that trick lies the artistry of intrapersonal practical commitment. It is perhaps this artistry that my two-year-old is practicing with pillows on the sofa while I write this, letting his stories (one minute he's building a fort, the next fording a rapids) sweep him through projects whose point lies only in the pleasure of not letting himself be distracted from following through. (Oh, the protests when I interrupt!) For grown-ups, practical commitment comes racked with ambivalence, self-uncertainty, and anguished confusion. In these respects our intentions are anything but stable. It takes the ruthlessness of a skilled artisan to pull the recalcitrant pieces into something recognizable as a course of action. The rational stability lies only—but crucially—in the story.[58]

Edward S. Hinchman
Department of Philosophy
University of Wisconsin-Milwaukee
USA
hinchman@uwm.edu

NOTES

[1] In addition to David Velleman's work, which I'll discuss in section 3, here is a small sample: Alasdair MacIntyre, 'The Intelligibility of Action', in J. Margolis, M. Krausz, and R. M. Burian (eds), *Rationality, Relativism, and the Human Sciences* (Dordrecht: Martinus Nijhoff, 1986); Jerome Bruner, *Acts of Meaning* (Cambridge: Harvard University Press, 1990); Daniel Dennett, 'The Self as a Center of Narrative Gravity', in F. S. Kessel, P. M. Cole, and D. L. Johnson (eds), *Self and Consciousness: Multiple Perspectives* (Hillsdale: Erlbaum Associates, 1992), 103–115; and, more recently, Jeanette Kennett and Steve Matthews, 'Normative Agency', in K. Atkins and C. Mackenzie (eds), *Practical Identity and Narrative Agency* (New York: Routledge, 2008), 212–31; Genevieve Lloyd, 'Shaping a Life: Narrative, Time and Necessity', ibid., 255–268; Kim Atkins, *Narrative Identity and Moral Identity: A Practical Perspective* (New York: Routledge, 2008); and Peter Goldie, 'Narrative Thinking, Emotion, and Planning', *The Journal of Aesthetics and Art Criticism* 67 (2009), 97–106.

[2] Not all of a toddler's action-guiding stories are so fanciful, of course. But a small child seems most self-governing when using imagination in this way. More literal self-narrations tend merely to repeat self-accountings the child has heard others give.

[3] Compare how one might take pleasure in the image of oneself breaking a marathon tape as one merely concludes the final stumble of a middle-aged jog. That interpretation puts a happy face on an action that one would have performed in any case.

[4] The story might explain why he has that desire, but it would not otherwise figure in the core account of his intention or action. After all, we could give the same core account were his desire to build a pillow-fort motivated by a different story, or were it not motivated by a story at all.

[5] Karen Jones engages a version of the prospectivity puzzle in 'How to Change the Past', in Atkins and Mackenzie, op. cit., 269–88; and John Christman presses a version of the diachronic integrity puzzle in 'Narrative Unity as a Condition of Personhood', *Metaphilosophy* 35 (2004), 695–713.

[6] In Jon Elster's term, intention is not the same as 'pre-commitment'. For the concept of precommitment, see Elster, *Ulysses Unbound* (Cambridge: Cambridge University Press, 2000), Chapter I. For a full argument for this claim, see my 'Trust and Diachronic Agency', *Noûs* 37: 1 (2003), section VII. The authors whose work on the stability of intention I'll engage also reject the idea that intention is mere pre-commitment.

[7] Holton defines the stability of intention as an upward shift in the threshold of how relevant information must be in order to reopen deliberation: 'some information that would have been relevant in forming an intention will not be sufficient to provoke rational reconsideration once an intention has been formed' (*Willing, Wanting, Waiting* [Oxford: Oxford University Press, 2009], 2–3). The most pressing sort of information would concern how your preferences have changed since you formed the intention, but some other information may likewise fall short of the shifted threshold. Your intention to picnic this afternoon may thus lead you to ignore not only a sudden bout of picnic ennui but some inconclusive hints of a storm on the horizon, both of which would have registered in your now-concluded deliberation whether to picnic. For simplicity, I'll focus on preference shifts, especially on those that present themselves as 'temptations' to reconsider, though we shouldn't forget that a stable intention will resist reconsideration on other grounds as well.

[8] We'll return to this issue. See note 15 below.

[9] See, for example, David Gauthier's historical discussion—of Plato, Hobbes, and Hume—in *Morals by Agreement* (Oxford: Oxford University Press, 1986), Chapter X.

[10] As we'll see, my thesis is not at all that the stories you tell yourself as you intend should somehow add up to a single grand narrative that would make sense of your life as a whole. This is one—though not the only—reason why my approach does not present an easy target for polemics along the lines of Galen Strawson, 'Against Narrativity', *Ratio* XVII (2004), 428–52, and Bernard Williams, 'Life as Narrative', *The European Journal of Philosophy* 17 (2009), 305–14. Another reason is that my approach does not speak of narrative 'unity'; as I'll emphasize in section 5, I am not using *as explanans* any notion of coherence.

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

[12] 'Toxin, Temptation, and the Stability of Intention', reprinted in his *Faces of Intention* (Cambridge: Cambridge University Press, 1999), 79*ff*.

[13] Bratman coined the term in 'Toxin, Temptation, and the Stability of Intention', 86*ff*. My formulation makes somewhat more precise his characterization of plan's end as 'the conclusion of one's plan' (ibid.), though it is possible that Bratman meant to exclude death

bed conversions and the like—which my formulation (to be modified presently in any case) would allow. It makes sense to prefer my formulation because it obviates the conceptually vexing question of when a plan is ever really 'concluded', given that one can always return to it and reason from it to new plans. (You rediscover the stamp collection that you abandoned at the age of twelve and resume the hobby—so the plan wasn't concluded after all!) Moreover, some intended actions (e.g., maintaining your health) simply do not actually have an envisaged 'conclusion'. (Holton presses this latter observation against Bratman (*Willing, Wanting, Waiting*, 158).)

[14] As we'll see, the thesis is not that you're actively thinking about your plan's-end self. The thesis is that you're making an assumption about this self: the assumption that (as we'll see) grounds your presumption of executive authority.

[15] This complements a further observation: to form an intention, you must expect, not necessarily that you *will* follow through on the intention, but that it would be *then rational*—that is, rational from the perspective of your acting self—to follow through (given, again, what you now expect will be the action's nature and consequences). We can perhaps resolve the longstanding debate over whether you can form the intention to φ when you believe you will not φ by focusing on this aspect of the intention-forming self's presumption of authority. When you form an intention to φ, you don't have to believe you will φ, or even that you will try to φ. You need merely believe that you thereby make it nondeliberatively rational for your acting self to φ—i.e., that it *makes sense* to, in the respect we're going to articulate—insofar as that self simply follows through on the intention. For an argument on this point, see my 'Trust and Diachronic Agency', section V.

[16] Bratman usually refers to the agent's 'follow through', which is consistent with my preferred interpretation of the no-regret condition, but he sometimes explicitly speaks of regretting the 'action' (see e.g., 'Valuing and the Will', in his *Structures of Agency* (Oxford: Oxford University Press, 2007), 56, and '*Toxin, Temptation, and the Stability of Intention*', 88). In any case, he does not consider the possibility of regretting your relation of self-influence *as opposed* to you action, which is the distinction I'm stressing here.

[17] By the no-regret condition, you must also expect that you would regret it if you failed to realize that relation. But because you typically (though not necessarily; see note 15 above) expect that you will follow through, this hypothetical expectation does not play the same role as the categorical one. (See note 44 below.)

[18] In 'Temptation Revisited' (in *Structures of Agency*), Bratman emphasizes the difference between the pragmatic orientation of his own earlier treatment (which Holton echoes) and a new orientation toward agential authority, noting how these orientations would make different use of an appeal to regret. (We'll discuss how Bratman's view has evolved in section 5.) But he does not appear to find it noteworthy how the pragmatic appeal to regret obscures the distinction between regret and disappointment. In any case, he does not use the shift to agential authority to clarify that distinction. (A terminological issue: in 'Temptation Revisited', Bratman contrasts 'the strategy of agential authority' with 'the strategy of intention stability', thereby revealing that he is using 'the stability of intention' to refer to a specific solution to our problem rather than to the problem itself. In my usage 'the stability of intention' always refers to the problem, never to a solution.)

[19] See again note 13.

[20] We might argue that such a fairness requirement is internal to regret in the way that it is internal to all reactive attitudes. For the idea that a fairness requirement is internal to reactive attitudes, see Gary Watson, 'Two Faces of Responsibility', 'Two Faces of Responsibility', *Philosophical Topics* 24 (1996), 227–48; reprinted in his *Agency and Answerability* (Oxford: Oxford University Press, 2004).

[21] Of course, the addressee in this instance is typically not a single point of view but a series of points of view unified by the fact that they do not disagree in how they feel about the action in question. My present point is that you may posit these points of view not as actual but as fictional. I'll explain the nature of the fiction in sections 3 and 4.

[22] *Willing, Wanting, Waiting*, 141.

[23] That is, it does not generate this odd result of Gauthier's view: that you should deliberatively spurn the tactic—forming the intention but not following through on it—that you all along think best when it becomes available to you. That tactic was not available to you last night but now, money in the bank, it is. If, *per impossibile*, you formed the intention last night, surely the thing to do now that you have the payoff is to abandon it. For Gauthier's treatment of the Toxin Puzzle, see e.g., 'Rethinking the Toxin Puzzle', in Jules L. Coleman and Christopher W. Morris (eds), *Rational Commitment and Social Justice* (Cambridge: Cambridge University Press, 1998). The term 'resolute choice' is Edward McClennen's: see his *Rationality and Dynamic Choice* (Cambridge: Cambridge University Press, 1990), 12–13. Holton frames his account as 'broadly consistent with McClennen's' (*Willing, Wanting, Waiting*, 141, n. 5).

[24] *Willing, Wanting, Waiting*, 164.

[25] We might pause to consider Holton's two most natural lines of reply to our question. More broadly, he can deny that the prospect of regret need figure in the rationality of the forward-looking stance you adopt when you form an intention. More narrowly, he can argue that when Toxin cases are common in the way he's imagining, you will not regret acting on the disposition to follow through on your intention and—unnecessarily—make yourself sick. The narrow reply would amount to disputing that interjected 'unnecessarily'. Holton can argue that following through on the intention is necessary to the efficacy of your general disposition to follow through in such cases, since doing so (at least, in most cases) is how you inculcate and maintain such a disposition. He is arguing that the efficacy of the disposition—or, at least, a rational belief in its efficacy—is a necessary condition on your being able to form the intention, and thereby to get the payoff, in the first place. So actually following through (that is, without reconsidering your intention) is not something you ought rationally to regret, Holton can argue, since doing so is a necessary condition on your getting the payoff. (Holton makes the broad reply explicitly to Bratman's claim that the prospect of regret plays a role in stability. And the narrow reply seems implicit in his presentation of the imagined scenario in which Toxin cases are common: he is clearly asking us to imagine that agents do not regret their follow-through in such cases.) The problem is that each reply, the explicit and the implicit, fails to address the species of regret at stake in this dimension of agency, which must not be mistaken for mere disappointment with the results of your agency.

[26] See again note 15.

[27] *How We Get Along* (Cambridge: Cambridge University Press, 2009), 200. A very similar passage occurs in Velleman's 'Narrative Explanation', *The Philosophical Review* 112 (2003), 19.

[28] Ibid. (from both sources).

[29] Velleman addresses what he calls the 'problem of commitment'—the question *whether to follow through* on your practical commitment in 'The Centered Self', in *Self to Self*, 270*ff*. That is not quite the issue that we're considering. The issue we're considering is better framed by the question *how you can count as committing yourself* in the first place—that is, what it takes to count as making a choice or forming an intention. For a discussion of Velleman's resolution of his 'problem of commitment' in terms of what he calls 'constancy', see my 'Trust and Diachronic Agency', note 14 and section VIII. I offer

a more general critical treatment of Velleman's approach to commitment in 'Conspiracy, Commitment, and the Self', *Ethics* 120: 3 (2010). There is no reason to think Velleman would approve of the use to which I'm putting his account of narrative understanding.

[30] 'Motivation By Ideal', *Philosophical Explorations* 5 (2002), cited as reprinted in *Self to Self*, 324. He credits the example to Jim Joyce.

[31] Velleman argues elsewhere that the drive toward such self-understanding actually constitutes us as rational agents (for earlier formulations, see the Introduction to *The Possibility of Practical Reason* [Oxford: Oxford University Press, 2000]; for his current view, see chapter 1 of *How We Get Along*). We need not engage that argument here. (I attempt to engage it in 'Conspiracy, Commitment, and the Self'.)

[32] Velleman more recently describes this as a period of 'inauthentic pretense' (*How We Get Along*, 160). The concept of inauthenticity does not figure in his earlier treatment (which he footnotes) in 'Motivation by Ideal', though he there claims that actions that involve such fictional self-understandings are in some important respects irrational (326–28). We might wonder about this claim. Velleman claims that the smoker in his example is 'insensitive to some of the reasons that actually applied to him' (328), but it's hard to see why that must be so. The smoker needn't count as *insensitive* to a reason grounded in, say, the pleasure he takes in the taste of cigarettes when he lets his desire for them be overwhelmed either (in deliberation) by his understanding of the harms of smoking or (in follow-through) by the partly fiction-involving apparatus involved in practical commitment. This isn't insensitivity to a reason but precisely a way of assessing its force.

[33] Don't be distracted by Velleman's focus on improvisational acting; Broadway improvisers would not need to introduce novelty just to keep themselves interested in their parts!

[34] My proposal does not commit me to rejecting Velleman's account of narrative understanding in favor of a cognitivist account like Noël Carroll's, which appeals to curiosity how the causal-explanatory questions generated by the plot of the narrative will be answered ('Narrative Closure', *Philosophical Studies* 135 [2007], 1–15). I am not claiming that narrative understanding must involve such curiosity; I am agreeing with Velleman that the understanding is primarily emotional. In fact, the puzzle we're considering presupposes that narrative understanding is emotional, since it presupposes that the psychological investment required to keep a narrative 'alive'—to constitute it as an object of narrative understanding—is possible even when one experiences no actual curiosity whatsoever.

[35] Like Velleman, I insist that this is a question of understanding, not of assessment. The question is not 'Will my plan's-end self approve of what I've done?' but 'Will my plan's-end self be able to make (narrative) sense of it?' My explanation of this distinction will draw on the ideas developed in section 4.

[36] This qualification (codifying a conclusion drawn in section 2) should of course always be understood to apply. It would be cumbersome to keep repeating it.

[37] It is clear that the story you tell when you intend should be voiced in the second person ('say you're a nonsmoker at this party where everyone is smoking . . .'), not in the third ('once upon a time a nonsmoker went to a party where everyone was smoking . . .'). Though the third-person version aims in part at your empathetic identification with the story's protagonist, the second-person version enacts that aim in its very mode of address.

[38] I do not assume that Stern's theory of these interpersonal relations is correct. I assume merely that it possesses enough first-glance plausibility to provide a useful model for the intrapersonal relations we're considering. I am not building on Stern's theory but borrowing elements of its structure.

[39] Daniel N. Stern, 'Affect Attunement', in J. D. Call, E. Galenson, and R. L. Tyson (eds), *Frontiers in Infant Psychology* (New York: Basic Books, 1984), 8–9. This discussion was expanded in Stern's subsequent book, *The Interpersonal World of the Infant* (New York: Basic Books, 1985), 150–51, and Chapter 7, passim. For an updated treatment, see Stern's *The Present Moment in Psychotherapy and Everyday Life* (New York: Norton, 2004), Chapter 5: The Intersubjective Matrix.

[40] Stern, 'Affect Attunement', 9

[41] See Stern, *The Interpersonal World*, 53–61; *The Present Moment*, 36–37 and 64–70.

[42] Stern explicitly argues that they have a narrative structure, though he emphasizes that the narrativity as such lies only in how we represent them (see 'The Representation of Relational Patterns: Some Developmental Considerations', in Arnold J. Sameroff and Robert N. Emde (eds), *Relationship Disturbances in Early Childhood* (New York: Basic Books, 1989), esp. 66–69). There is no conflict with Velleman's view, however, since Velleman is not claiming that the emotions in question *are* narratives.

[43] Stern, *The Interpersonal World of the Infant*, 139.

[44] Paralleling the two-sidedness of what we have taken onboard from Bratman's appeal to regret, attunement would also involve feeling self-attributively unsettled by a memory of having failed to follow through on the intention. But again (see note 17 above), this side of your plan's-end perspective does not play the same constraining role in forming the intention simply because in forming an intention you typically expect that you will follow through. Acknowledging the hypothetical regret might nonetheless play a role in motivating you to follow through.

[45] For a much fuller account of the conditions in question, see my 'Intention and Time', in preparation. I keep emphasizing the incompleteness of the present account because my inquiry in that other paper reveals a dimension of complexity in the temporality of intrapersonal practical commitment to which I cannot begin to do justice here.

[46] My argument here converges with the broader project pursued in several other papers, including 'Trust and Diachronic Agency'; 'Conspiracy, Commitment, and the Self'; 'Receptivity and the Will', *Noûs* 43: 3 (2009); and 'Rational Requirements and 'Rational' Akrasia', *Philosophical Studies* 166: 3 (2013); and 'Intention and Time'. Each of these papers addresses the role of self-trust in an aspect of practical agency, but none appeals to narrative.

[47] I thus agree with Bratman that a good explanation of stability need not produce a necessary condition for it: see the change of mind cited parenthetically in note 49 below.

[48] For Bratman's earlier approach, see 'Valuing and the Will', in *Structures of Agency*, 56. This restates his justification of the no-regret condition in 'Toxin, Temptation, and the Stability of Intention', where he first proposed the condition. Again, this is the approach that Holton develops, criticizing Bratman for having abandoned it (*Willing, Waiting, Wanting*, 156–60).

[49] Bratman has argued that there are two fundamental problems in the philosophy of action. As we've seen, one is what Bratman calls 'the problem of agential authority': what kind of psychological functioning is necessary and sufficient for you to count as engaging in full-blown agency? The other is what he calls 'the problem of subjective normative authority': what is it for a desire or other pro-attitude to have normative authority for you? (See 'Two Problems about Human Agency', in *Structures of Agency*, 91, 92. But see also his 'Introduction' to *Structures of Agency*, 4–5 and 11, where he clarifies that he no longer wants to formulate the problem of agential authority in terms of necessary conditions.) Gary Watson has claimed that Bratman posits a 'metaphysical imperative' to

maintain one's identity (see 'Hierarchy and Agential Authority', in John Fischer (ed.), *Free Will: Critical Concepts in Philosophy* (New York: Routledge, 2005), volume IV, 94–95), and Bratman has subsequently confirmed this reading, characterizing his view of agential authority as 'a claim about the metaphysics of agency, not a normative ideal of integrity or the like (though we may, of course, also value some such ideal)' ('Three Theories of Self-Governance', in *Structures of Agency*, 246; Bratman says in a footnote that he is replying to Watson's claim).

[50] In light of the observations in note 32 above, we may wonder if the self-deception must even involve irrationality.

[51] We might put it this way: the questions of self-*governance* in play as you face challenges to the stability of intention seem different from the questions of *self*-governance in play as we consider whether you are responsible for or otherwise relevantly present in your behavior. For my view of the latter—of what I call 'the claim of attributability'—see 'Intention and Time'.

[52] We might say that whereas Holton's approach overlooks the complex intrapersonal relations in play, Bratman's quasi-moralizes them. Holton makes it too easy to form an intention, or easy in the wrong way, whereas Bratman makes it too difficult, or difficult in the wrong way.

[53] For the Bratman's embrace of such a 'metaphysical imperative', see again note 49 above.

[54] Again, you can form and follow through on intentions that manifest ambivalence. Even if that doesn't 'get you into your act' in the right way (which, again, I doubt), the problem does not seem to concern your capacity to form and follow through on the intention.

[55] I say more about this appeal to fairness in 'Conspiracy, Commitment, and the Self'.

[56] As I explain in 'Rational Requirements and 'Rational' Akrasia', this has implications for a recent debate over the nature of rational requirements. (For starters on that debate, see John Broome, 'Normative Requirements', *Ratio* 12 (December 1999), 398–419; Niko Kolodny, 'Why Be Rational', *Mind* 114 (July 2005), 509–63; Broome, 'Wide or Narrow Scope?' *Mind* 116 (April 2007), 359–70; Kolodny, 'State or Process Requirements?' *Mind* 116 (April 2007), 371–85. Broome brings the larger issue to bear on the stability of intention in 'Are Intentions Reasons? And How Should We Cope with Incommensurable Values?' in C. W. Morris and A. Ripstein (eds), *Practical Rationality and Preference: Essays for David Gauthier* (Cambridge: Cambridge University Press, 2001), 98–120.)

[57] I give an account of that relation in 'Receptivity and the Will', section VII.

[58] I wrote a proto-version of this paper for a workshop on Narrative and the Construction of the Self organized by Adrienne Martin and Matthew Smith at the University of Pennsylvania in March 2008. Thanks to Susan Sauvé Meyer for commentary and to workshop participants (especially Michael Bratman) for discussion. That proto-manuscript contained five or six distinct papers, as I began to see when I set out a year later to produce something I might actually publish. I wrote the next draft, which turned out to be merely two papers, in a different workshop: during the down time of a family leave. If you need to get specific about the link between narrative and agency, it helps to spend part of each day watching an infant eagerly watch his two-year-old brother develop a repertoire of actions, few of which make so much as a passing difference to the world but each of which comes with its story (some spoken, some hummed). And the workshop continues, since that infant now tells his own stories.