"What on Earth Was I Thinking?" How Anticipating Plan's End Places an Intention in Time

Edward S. Hinchman

University of Wisconsin-Milwaukee

[For the the copy-edited and published version of this paper, see Roman Altshuler and Michael J. Sigrist (eds), *Time and the Philosophy of Action* (Routledge, 2016), pp. 87-107.]

How must you think about time in order to form an intention? When you intend to φ at some future time *t*, you must of course think about *t*; perhaps you must in some related way think about the stretch of time between now and *t*. But must you place your endeavor in any broader prospect or retrospect? In what follows I argue that you must: in forming an intention, you commit yourself to a specific prospect of a future retrospect – a retrospect, indeed, on that very prospect. I argue that this broader temporal attitude articulates the species of self-accountability necessary for diachronic practical commitment. In forming an intention you project a future from which you will not ask regretfully, referring back to your follow-through on that intention, "What on earth was I thinking?"

In thus thematizing "plan's end" I build on Michael Bratman's approach to the stability of intention, on which the commitment at the core of intention puts you into a complex species of rapport not only with the later *you* that will execute the plan but with the still later *you* that will look back on that relation of self-influence.[1] Bratman plausibly argues that the key to undertaking a plan that will rationally resist subsequent shifts in your preferences lies in how you anticipate the retrospective stance that your future self will, if rational, take toward your having

brought the plan to completion. You make your practical commitment stable by looking thus forward to how you'll look back. If you anticipate looking back with fair regret on your having executed an intention – that is, on your having done what you intended to do – then you simply cannot form the intention, because you cannot institute the species of rational stability that defines an intention. As I develop the approach, the problem lies not in any disappointment targeting the results of follow-through – we can hypothesize that you wouldn't be disappointed – but in regret targeting the self-relation that you would institute in following through.

What exactly is wrong with intending in a way that you expect to regret? Note, first, that you may aim to avoid regret in individual acts of intending without aiming to live a life altogether free from regret – just as you may aim at truth in individual acts of belief without aiming to live a life altogether free from false belief. We're talking about an aim to avoid regret in the particular case, not in the general. Expecting that you'll regret a particular relation of self-influence posits three temporally distinct perspectives. Because you now expect that your plan's-end self will prove unable to make retrospective sense of the self-trust relation whereby you performed the action – "What on earth was I thinking? How could I have trusted that intention?" – you cannot now make prospective sense of how you could aim to institute such self-trust. When you claim the intrapersonal authority characteristic of intention, you expect that the relation whereby you execute the intention will continue to strike you – all the way out to *plan's end* – as in this way "speaking for" you. I aim to explain the stability of intention in terms of that self-trust relation, with its cognate species of self-accountability.[2]

In focusing specifically on the aim to avoid such regret, which I'll call "trust-regret," my approach differs crucially from Bratman's. Though Bratman earlier emphasized a pragmatic need for diachronic stability in intention, he later came to ground the no-regret condition in a

broader account of *agential authority*.[3] The "problem of agential authority" poses this question: what kind of logical functioning is sufficient for you to count as engaging in full-blown agency?[4] The problem of agential authority thus poses a question of attributability: what does it take for the action to count as straightforwardly attributable to its agent? I aim to provide an alternative to both Bratman's later emphasis on attributability, via agential authority, and his earlier emphasis on the pragmatic need for stability. Beyond a merely pragmatic emphasis, we need to see how the stability of intention rests on a norm of intrapersonal self-constitution. But we cannot explain this species of intrapersonal self-constitution in terms of attributability.

My approach differs from Bratman's not merely because I'll argue that there can be stable intentions without attributable agency – for example, an unwilling but planning addict. My approach more fundamentally differs from his because I doubt one of Bratman's central claims: that there is any "metaphysical imperative" to maintain your identity through time.[5] I'll argue that we must distinguish two aspects of how personal agency unfolds through time. On the one hand, we think that the person who performs an action must remain the same person through the interval of this performance. On the other hand, we think that the action itself must amount to a single performance – that is, a single instance of agential self-governance – through the interval in which it occurs. The first assumption poses a core facet of the issue of diachronic personal identity, but I do not believe that the issue of personal identity, even in this facet, need directly inform the issue of diachronic self-governance. On my approach, the second assumption poses the most fundamental explanandum for a theory of diachronic self-governance. The core question is not "How does the agent's settled identity establish the identity conditions for this instance of agential self-governance?" but "How does the agent's *claim* to a settled identity establish identity conditions for this instance of agential self-governance?" My

approach emphasizes how the claim of attributability informs an aim to avoid the reactive-attitudinal sanction of trust-regret, rather than any aim at attributability as such.

1. Toxin and temptation

Why should we regard intention as sensitive to future regret? I agree with Bratman that appeal to the perspective from which you might look back with regret provides the basis of a particularly compelling resolution of Gregory Kavka's toxin puzzle.[6]

To get Kavka's puzzle case, imagine that an eccentric billionaire with a reliable intention detector reliably promises to give you a million dollars if at midnight tonight you form the intention to drink a certain toxin tomorrow at noon. You must do so, he stipulates, without ignorance, manifest irrationality or such external mechanisms as a toxin-administering machine or a side bet. If you do thus form the intention at midnight, the billionaire will deposit the money in your account tomorrow morning as soon as the bank opens. He doesn't care whether you drink the toxin, which you know will make you quite ill for a day or two but leave you thereafter unharmed. To get the money, you need merely form the intention to drink it. Kavka's puzzle is that forming this intention seems impossible under the circumstances.

Let's assume that Kavka's treatment of his original case is correct: you cannot get the million dollars merely for forming the intention because you foresee that it would be irrational to follow through on that intention. How exactly does a toxin case differ from a run-of-the-mill temptation case, in which you expect to undergo a transient shift in your preferences as the time to act grows near? And why should this difference make such a difference, preventing you from even forming the intention in a toxin case? I agree with Bratman that our explanation should

appeal to a perspective that comes after the time of action, but why should that perspective have such authority?

Had the billionaire offered to reward you for drinking, rather than for forming the intention to drink, it would have been perfectly possible for you to form the intention to drink. In forming and retaining that intention, you'd have had to expect that you'd be tempted to reconsider when presented with the noxious liquid. How might you have countered those temptations? One natural strategy is by reminding yourself that you'd not regret following through on this intention. One salient change introduced by the billionaire's offer to reward you for forming the intention to drink, rather than for drinking, is that this strategy no longer works. You expect that, with the million dollars in your bank account, you *will* regret drinking the toxin. This shift in your expectations – with a reward for drinking you would expect not to regret, but with a reward for merely forming an intention to drink you expect to regret – helps explain why you are unable to form the intention in Kavka's puzzle case.

Bratman more generally distinguishes ordinary "temptation" cases from "toxin" cases modeled on Kavka's in terms of expected regret (1999, 79*ff*):

(Toxin) As you form or retain the intention you anticipate that you will regret following through on it and that you will not regret not following through.

(Temptation) As you form or retain the intention, you anticipate that you will not regret following through on it and that you will regret not following through.

Bratman treats the anticipation as a matter of expectation. In a temptation case, you expect you will eventually regret having given in to the temptation to act contrary to your intention, but in a toxin case you expect you will not regret. Moreover, in a temptation case you expect you will not regret having followed through on your intention, but in a toxin case you expect you will regret. Bratman argues that this "no-regret condition" explains why you can form the intention in the temptation case but not in the toxin case.

The no-regret condition imposes a key distinction between what we may call the "once-future" perspective of action and the "twice-future" perspective from which you might look back on that action with regret. That distinction is especially clear in a toxin case, in which forming the intention would require expecting both that you will regret following through on it and that you will not regret not following through, expectations that in turn prevent you from forming or retaining the intention. Bratman's moral from the toxin puzzle is that in order to form an intention you must expect both that, given what you expect about the circumstances in which you will act, you will not later regret following through on the intention and that you will later regret not following through. This expectation targets the once-future circumstances of your future action *as they will be regarded* from a still later, *twice*-future perspective.

Let me briefly anticipate two worries about the no-regret condition that I'll deal with more fully as my argument proceeds. First, one might worry that the no-regret condition does not formulate a necessary condition on present-directed intentions, since we all commonly perform impulsive actions while thinking "I'm going to regret this." Though in other work I argue that an analogue of the no-regret condition does apply to present-directed intentions (Hinchman 2016c), I lack space to pursue that issue here. For present purposes, I claim only that the no-regret condition serves as a necessary condition on future-directed intentions. Still, one

might worry that we can form future-directed intentions that are akratic, if not exactly impulsive, and that we might express the akratic element in these intentions by thinking "I'm going to regret this." I agree that this is possible, but the question is what it shows. In the next two sections, I'll argue that the possibility cuts against Bratman's approach to stability but not against mine.

2. Why plan's end must be projected, not expected

Why should this difference in what you expect to happen long after the intended action – at what we're following Bratman in calling *plan's end*: the twice-future point beyond which you will not change your attitude toward your once-future action[7] – 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,[8] to how you will view the action from plan's end? Now that we see the need for a no-regret condition, we can see the need for what I'll call a "projection" out to plan's end. The need for such projection arises from a problem that I'll now develop for the simpler view on which the no-regret condition appeals to mere expectations about the future.[9]

Consider any case in which you expect both that your sensibility will change after you've acted on the intention and that after the change your regret will have no bearing on whether you should have acted on the intention. 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 context 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 will 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 agential authority when you form the intention or on the stability of your intention once you form it.[10]

If Bratman's interpretation of the no-regret condition were correct, then the cult case would be a toxin case, wherein you cannot even form the intention because you expect you'll regret following through on the intention at plan's end. But the cult case is clearly not a toxin case. Clearly, it is a temptation case, wherein you may naturally expect a preference shift as the time of action grows near and you begin to appreciate how much easier your life would be if you acted as if you'd already undergone the "reprograming" that you expect you'll undergo after visiting your parents. Expecting you'll come to hate your parents, you may be tempted to act as if you already do, since you may prefer the simplicity of a consistent attitude to the anguished complexity projected by your intention. But, no: you're resolved to resist that preference shift, not because it is merely transient – by hypothesis, you believe it isn't merely transient – but because plan's end lies in a projected future that happens not to coincide with your expected actual future. And we may say the same of the less cartoonish cases that share this structure.

It obviously won't help simply to prohibit applications of the no-regret condition that depend on a shift in sensibility. Sometimes looking ahead to a shift in sensibility is entirely appropriate and not to do so irrational. Think of any case wherein a shift in sensibility marks the stereotypical effects of moving forward in your life. You expect you'll grow more "conservative" as you get older. Or you expect you'll have a different take on your life post-parenthood, no longer taking for granted, as you now do, that suburbs are for sell-outs. You do

not expect that this change of mind will reflect better values, merely values better attuned to your needs as you expect them to become as you move into a later stage of your life. Everyone undergoes such shifts in sensibility over the course of their lives. You might even expect that such a shift of sensibility will emerge as a causal byproduct of some intention that you have formed and now retain, without regarding the shift as determining how the no-regret condition should apply. The irrelevance of such expectations may become especially clear in hindsight. Perhaps you find your sensibility shaped by your decision to have children but don't think your earlier bouts of diachronic agency should have been shaped by expectations of a parenting reality that you were then barely able to imagine. Or perhaps you've always expected that you'd wind up depressed and destitute, just as your parents did, with an abiding sense of the pointlessness of all your efforts to avoid this fate. Why should these expectations of where you'll actually wind up have any normative bearing on your capacity to intend, plan, and resolve well before the expectations are confirmed (or not)? Some of the expectations may mark a horizon at plan's end in your actual life, since you may actually look back with an altered sensibility and regret your earlier follow-through. We need to understand how the no-regret condition can work within a projection that continues the form of sense-making that informs your intention.[11]

3. What is it to project plan's end? The key is accountability, not attributability

In light of these problem cases, plan's end cannot figure quite as Bratman proposed, as the horizon of *expectation* beyond which you will not change your attitude toward the action. But we can nonetheless make use of his idea, by viewing plan's end as a notional perspective internal to the sense-making projection that informs your intention. I'll now argue that the best

explanation of the normative role of plan's end treats the projection as grounded in relations of intrapersonal accountability.

The no-regret condition itself motivates the shift to accountability. What exactly do you think you won't regret when your intention is governed by the no-regret condition? It is clearly too strong to require that you project a future in which you'll not regret the action that you intend to perform, since you know that any action may have unforeseen consequences or turn out to violate some principle in a way you hadn't foreseen.[12] The projected regret must target not the action as such – whether in its nature or in its consequences – but how you perform it. What aspect of how you perform the action? I'll now argue that the question of projected regret targets the self-relation that you manifest in this instance of self-governance, given what you now expect about the circumstances of your action. What aspect of the self-relation? I'll argue that you must project a plan's end from which you will not regret the dimension of your self-relation – specifically, the self-trust relation – that would be manifested in your following through on the intention. To form an intention, you must project a future free from what I'll call *trust-regret*, by which I mean regret targeting specifically that relation of self-influence.[13] It is this appeal to self-trust that will get self-accountability in play.

To see how, let's see how this focus on trust-regret generates a deeper difference from Bratman's approach. As I mentioned at the outset, in recent work Bratman aims to ground the no-regret condition in a broader account of *agential authority*.[14] 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 of what you do in a way that would make the action attributable to you – rather than to, say, a force within you. My account also appeals to attributability, but in a completely different way that does not draw any link between

attributability as such and authority. On my approach, the question of attributability arises from the agent's own point of view and asks not whether onlookers can coherently attribute the action to this person but whether this person coherently attributes the action to herself in the course of holding herself accountable for it. This approach suggests a different link between attributability and agential authority. If your would-be intention projects a future in which you become, as I'll say in shorthand, *self-attributively unsettled* by the thought that it was you who performed the action, then you cannot coherently claim the authority distinctive of that intention, and you thereby fail in your attempt to form or self-consciously retain the intention. To say that you are "self-attributively unsettled" is to say, in fuller formulation, that within the projection informing your intention you come to regret the self-trust relation whereby you followed through on the intention and thereby implicitly claimed "ownership" of both intention and action. When that happens, you cannot – without confusion or manifest irrationality – claim agential authority, because you see that any such claim would, by your present lights, lack authority for your acting self. Within your projection you see that when the time came to follow through on the intention, you'd be irrational if you did not abandon it. You see you'd be irrational to follow through because you see that you'd regret the relation of self-influence from plan's end. You see you might express the regret with retrospective self-chastisement: "What on earth was I thinking?"

This is why you cannot form the intention in Kavka's toxin case: the most fundamental problem is not that you cannot *be* an authority (from your own point of view) but that you cannot coherently *claim* authority (from your own point of view). For Bratman, you cannot coherently claim authority because you do not satisfy the conditions for being an authority (from your own point of view). I do not, by contrast, aim to explain why you cannot coherently claim authority in terms of any deeper explanation of what constitutes your authority as such (from your own

point of view). For present purposes we need not give any deeper explanation of agential authority. We need merely see why you cannot coherently claim such authority when you project that you would be self-attributively unsettled by the self-trust relation that would emerge if you let that claim of authority influence you in the way that it aims to influence you. That's all we need in order to explain the species of agential authority distinctive of an intention.

While I lack space for a full exploration of this difference between my approach and Bratman's, let me offer one reason for separating the core issue of the rational stability of intention from Bratman's conception of the issue posed by agential authority. Assume for the sake of argument that ambivalence of a sufficiently deep or systemic sort can undermine the attributability of an action.[15] Imagine that you are stricken by such ambivalence but also that you engage in what to all appearances looks like planning agency, like the unwilling addict who pursues complex means in order to obtain his next batch of drugs. Given our assumption, the actions that you perform through such planning will not in fact be (fully) attributable to you. But, setting that aside, it seems that we can nonetheless explain how your intentions are rationally stable. How might we vindicate the rationality of your expectation, as you intend, 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. Again, 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. So Bratman cannot allow that your intention could be stable in our ambivalence case. My approach, by contrast, appeals not to attributability but to the *claim* of attributability, a claim that you do make, even in such a case, insofar as you aim to avoid your own trust-regret by aiming to perform only those actions that you will not later find

self-attributively unsettling. It is part of the self-trust dynamic that you do *regard* as attributable to you the action that you would perform by self-trustfully following through on your intention. If you did not so regard the action, you could not find the self-relation manifested by the performance self-attributively unsettling.[16]

We thus explain the projective form of sense-making that informs your intention. Even if you do not expect to regret trusting your intention in your actual future, if you project plan's-end trust regret, you cannot claim the species of agential authority that would inform that intention, and so you cannot – without confusion or manifest irrationality – form the intention. But you can claim that species of authority when the course of action to which it would lead is not attributable to you. It is hard to see why we shouldn't allow that an unwilling addict in the grip of his addiction can nonetheless plan how to purchase his drugs and when or how to use them. Even if these actions are not (fully) attributable to him, his planning manifests his agential point of view and as such can prove rationally stable. Gripped by his addiction, he makes a plan to obtain drugs at his dealer's house, but upon arrival our hero finds himself struggling with what seems like an akratic temptation to abandon the plan.

Can we make real sense of such apparent "akrasia within akrasia"? Imagine two possible cases in which our hero might thus pace nervously on the dealer's front step, reluctant to knock. In one case, our hero struggles with a conflict between his akratic intention to buy drugs and the unabandoned resolution to give up drugs that makes that intention akratic, his addicted planning agency thus in conflict with a "better judgment" now functioning as an apparent source of temptations pulling him away from his plan. If he gives in to this temptation, though without thinking of himself as having abandoned his plan to buy drugs, that would not count as "akrasia within akrasia" but a resolution of the original akrasia into a confused form of ambivalence: he

has put himself back into motivational touch with his better judgment, but he is strangely confused by this newfound enkrateia, believing that the intention to buy drugs continues to "speak for" him.

But that is not the only form that "akrasia within akrasia" might take. In a second case, imagine a different temptation pulling our hero away from his plan to buy drugs – the all too familiar temptation to flee that grips him every time he arrives at his dealer's doorstep. He fears his dealer, always has, and coping with this fear is a familiar obstacle en route to each of his drug purchases. This drug purchase, unlike those others, is akratic, since it is countered by an unabandoned resolution to end his drug habit. But that difference makes no difference to the familiar temptation now gripping him at the dealer's doorstep. If he gives in to this temptation, it wouldn't rescue him from the akratic plan that left him there. Rather, it would make his akratic predicament worse. Giving in to this fear of his dealer would not, on its own, amount to abandonment of his plan to buy drugs. Say he paces back and forth for a few minutes grappling with the fear. That does not amount to grappling with the temptation to buy drugs, given that he intends to end his drug habit. It amounts to grappling with the temptation to give in to fear and thereby fall short in this plan to buy drugs, given that he nonetheless retains the plan. If he does give in, that would amount to akrasia within akrasia – as such, a violation of the rational stability of his intention.

How might we explain rational stability in light of this possibility? The possibility of such cases appears to show that the rational stability of intention cannot be a matter, most fundamentally, of attributability. So how else might we explain it? I have begun to offer an alternative that emphasizes not attributability but the *claim* of attributability: our addicted hero can form an akratic intention to buy drugs because, even though the intention is not actually

attributable to him, he can nonetheless claim that it is. A proponent of Bratman's explanatory approach may reply that we cannot make sense of this appeal to a mere claim of attributability without understanding the role of attributability in constituting the intrapersonal relations at the core of rational stability. Is our hero merely claiming that his case is a normal case, in which the intrapersonal relations do suffice for stability? That's how we might have to interpret my talk of a "claim of attributability" if there were no other way to make sense of it. As I've already suggested, however, we can make sense of it within the economy of the agent's self-accountability relations as they project a future out to plan's end.

Our hero can akratically intend to buy drugs, even though he retains his resolution to end his drug habit, because the intention and the resolution project different futures.[17] When he resolves to end his drug habit, he projects a plan's end from which he will not regret having given up drugs and will regret not having done so. But when he intends to buy drugs, he projects a plan's end from which he will not regret continuing his habit ("just one more time") and will regret not having done so ("since anyone who cares about me won't want to see me in this pain"). Notice the parenthetical rationalizations. We needn't imagine our hero blindly lashing out in his addiction, as if he could score some drugs by having a temper tantrum. Though he has resolved not to do so, he gives in to the temptation to project a plan's end that conflicts with the projection at the core of his resolution. We could say that his addicted projection expresses the addiction rather than the judgment with which he identifies; that formulation captures the intuitive force of our hypothesis that this entire course of action, including its planning, is not straightforwardly attributable to him. Even so, there is planning, and the planning occurs in his psychology. From the perspective of this planning, there is rational stability: from this perspective, it would be irrational of him to give in to his fear-based temptation to flee. The

projection, with its rationalizations, reveals the influence of the very addiction that he has resolved to overcome. But it enables him to formulate and pursue, through intrapersonal rational commitments, a diachronically robust course of action.

In the next section I'll more fully explain the species of intrapersonal accountability that gives shape to the projection at the core of such diachronic rational commitment. My alternative to explaining rational stability as a matter of attributability begins from my emphasis on how a *claim* of attributability lies nested within intrapersonal accountability from plan's end. But before we move on, let me use the dialectic developed thus far to reply to a worry about how I am using the concept of a projection. I claim that your intention rests on a projection into a future that you need not regard as your actual future. But if you do not expect plan's end to lie in your actual future, how could you could distinguish actual correctness from a mere feeling of correctness? As Wittgenstein framed the problem for a different but structurally parallel aspect of rule-following (1953, sect. 202), how could you distinguish being right from merely seeming right? If you expect that you'll go off the rails in the way of the cult case, how could any projection into a future that you expect to be non-actual serve to keep you in line?

This worry presupposes the emphasis on attributability that I'm rejecting. It is true that if we combined Bratman's appeal to attributability with my appeal to projection, we'd have to say that you constitute yourself as a unified subject of your actions through your projection out to plan's end. But the worry reveals why that would not work in many cases: your projection leaves open the possibility that you will fail to have actual future "partners" in the normative endeavor – that is, actual future selves within the projection. By contrast, it is not hard to see how a species of self-accountability could unfold through a projection. After all, we often feel accountable "in the eyes of others" even when there are no others around, or when we know that

no actual others in our community will hold us accountable. The link between accountability to others and accountability to *actual* others is isomorphic to the link between self-accountability and accountability to *expected* actual future selves. In each case, you could not learn how to feel accountable if no one ever held you accountable. (Self-accountability could not be something that you only ever imagine. You learn how to hold yourself accountable in this dimension by coming to feel actual regret!) But in each case, whether interpersonal or intrapersonal, accountability is possible when there is no actual person in position to hold you accountable.

4. Projecting a retrospect: regret as reactive attitude

We need to understand how this species of self-accountability works in detail. What *are* trust and regret such that they should play this role in diachronic agency? I'm arguing for the importance of a species of agent-regret that I'm calling "trust-regret," since what you regret is that you entered into an unwise trust relation.[18] And I'm arguing for the importance of a species of trust that we could describe as "reasonable but not deliberated confidence in someone's anticipation of what you'll regret," where the "someone" might be your own earlier self. How do these species of trust and regret combine to serve as a backward-looking reactive attitude?

Following an approach pioneered independently by Susan Wolf (1990), Jay Wallace (1994), and Gary Watson (1996), we must understand two things in order to see how trust-regret functions as a reactive attitude: how it involves a sanction, and how application of this sanction involves a norm of fairness. The fairness norm applies as follows. Just as you don't hold someone fully responsible when your don't think she had a reasonable opportunity to avoid this sanction, so you don't regret your intention-to-action self-trust relation when you don't think

your earlier selves had a reasonable opportunity to avoid the regret. When I say that you "don't" react or regret in these ways, I mean that we agree that it would be wrong to.

Several different types of case illustrate the intrapersonal fairness norm. First, you don't regret – that is, *trust*-regret[19] – when you don't think your acting self could have foreseen the circumstances in which the self-trust has come to seem foolish. Second, and more complexly, you don't regret when both of these conditions hold: (a) your acting self acted on self-trust – that is, without redeliberation – and (b) you don't think your *intending* self could have foreseen the circumstances that make the self-trust foolish. Third, there is an analogue of the point emphasized by Wolf: you don't regret self-trust informed by *deep* deficiencies of character. This is typically because you can't see them: since the deficiencies inform the perspective from which you deliberate, intend and act, you typically can't recognize yourself in that description. Even if you can recognize the deficiency, your inability to do much of anything to change – which follows from the hypothesis that the deficiencies are "deep," in Wolf's metaphor – makes regret (again, *trust*-regret) the wrong reaction. What you feel is better classified as disappointment.

This role played by fairness is revealed less in the reactive attitude itself than in your efforts to avoid it. Just as there is no constructive deliberative role for the worry that drink or disease will make you regret – apart, of course, from the strategic question of coping with that consequence – so there is no constructive role for the worry that you'll regret unfairly. Though in forming an intention you aim to avoid twice-future regret at the relation that you thereby aim to institute with your once-future self, there is no constructive role for a worry either that you will regret for reasons that you cannot foresee – as opposed to those you can foresee but merely overlook – or that you thereby manifest deep deficiencies in your character, or some other deficiency that it would be unfair for you to regret. Still, such worries can *infect* practical

reflection – precisely because they cannot inform it. When reflection is thus infected, you may expect to regret following through on the intention that you're forming, but this expectation does not prevent you from forming that intention, though you may do so with a feeling of despair. It is part of what makes such self-accountable agency possible – in part by preventing such self-despair from undermining your agency – that the norm you're aiming to meet is thus constrained by fairness. Since you aim to avoid *fair* regret, an inability to respond constructively to all the worries that you might feel about taking responsibility does not impede your ability to act. When such anxieties unsettle you, it is the fairness norm that makes intention possible.

We see here another reason to interpret the no-regret condition in terms of a projection that need not coincide with any expectation. In section 2, we saw how it is possible to expect that you'll regret following through on your intention without being thereby constrained in your ability to form the intention. In those cases, your expectation of regret targets the far side of an expected "shift in your values" (as we might say, in shorthand for this complex range of normative changes). You expect that the regret will be informed by values that you expect you'll adopt for reasons that fail to recommend those values on their merit: in our cartoonish example, that you'll be brainwashed by a cult into hating your family; or, with more realism, that you'll become narrow-mindedly "conservative" with age, or overridingly concerned with the needs of children for whose existence you haven't even begun to plan. But what if you expect not a shift in values but a sheer deficit in your concern with retrospective fairness? Imagine you expect, not the species of "narrow-mindedness" that marks a shift in values, but the species that marks a lack of concern to do justice to any perspective but your own perspective at that very moment. You see both your parents grow embittered by life's frustrations and take out their bitterness on their younger selves, "regretting" nearly everything they did when young, including things that you

can see made perfectly good sense at the time. You may come to expect that this will happen to you; but the expectation does not give you pause as you form and carry out the plans that you expect will thus provoke your embittered self at plan's end. You can see that this is not regret but self-recrimination. The expected self-recrimination may trouble you, but it does not disrupt your efforts to plan in the way that projected regret would do. Recrimination is a charge, a species of psychic aggression, and as such it may be entirely arbitrary. It is not a reactive attitude unless it represents itself as responsive to a norm of fairness. Just as mere hostility may not yet amount to a responsibility-imputing reactive attitude on the order of blame or resentment, so mere self-recrimination may not amount to a responsibility-*self*-imputing attitude on the order of trust-regret.

This reflexivity is crucial to the normativity of the self-relation. Interpersonal hostility may or may not represent itself as responsive to a norm of fairness. Even if it does, and even if it thereby counts as a reactive attitude (transforming mere hostility into, say, contempt), that does not yet mean that it *is* fair. If it is not fair, then it may not succeed in establishing any species or degree of actual responsibility.[20] Intrapersonal regret may likewise fail to prove fair, but such a failure does not in the same way compromise its normative role – that is, the normative role that I'm arguing it has in stabilizing intention. Even if the regret proves unfair, when you project the regret you project it *as* fair. And that – the projection – is what ensure that it plays the normative role. You may hold another person in an attitude of contempt that proves unfair. But if you try to project "unfair regret," what you project is mere self-recrimination that, as such, fails to bear on the stability of your intention. In each case, the interpersonal and the intrapersonal, your attitude is governed by a fairness norm: if unfair, it is wrong or unjustified. What makes the difference is that in the intrapersonal case, unlike the interpersonal, you cannot adopt the attitude

insincerely. The intrapersonal attitude cannot be insincere because it functions prospectively: you take a prospect toward a future retrospect. Within that prospect, the retrospect is projected as fair, in part because a retrospect that you did not project as fair, within that prospect, could not function to stabilize your intention. The normative force of the prospect thus derives from the retrospective attitude that it projects, and the whole point of projecting is to stabilize your intention. Within that prospect, you project the retrospective attitude as fair, simply because if you didn't you couldn't thereby undertake the normative attitude – the intention, plan or resolution – that the prospect informs.

You can hold someone in just a degree of contempt. Can you likewise project just a degree of trust-regret? There are at least two ways to answer this question: (a) yes, and the role of projected trust-regret in stabilizing intention involves the idea of a threshold beyond which the degree of projected trust-regret is too high; or (b) no, since any degree of trust-regret serves, within the projection, to undermine the stability of your intention and thereby your ability to form it. I believe that the correct answer is (b), but nothing in my argument depends on that answer. I believe that the answer is (b) because I believe that there is at least this much truth in Harry Frankfurt's proposal (1988) that agential authority requires freedom from ambivalence: the *claim* of agential authority requires *projecting* an unambivalent retrospect from plan's end. Even if you are ambivalent, as in our example of an unwilling though planning addict, when you invite self-trust you project an image of yourself as wholehearted in the retrospect that you adopt at plan's end. It is in this respect, I believe, that diachronic agency involves an ideal. Just as you project fair trust-regret, a normative idealization, so you project unambivalent trust-regret, a psychological idealization. Both idealizations may fall short of ensuing reality: at plan's end, you may regret your self-trust unfairly, or you may regret to this or that degree. But *within* the

projection you provide yourself with an idealized image of what you are up to in intending, an image whose idealized nature enables you summon the clarity of purpose that you may need to resist temptation.[21] If we prefer answer (a), with its apparently more realistic psychology, we lose this simple account of how clarity of purpose structures the stability of intention. But it is no part of my argument in the present paper that we cannot provide such an alternative account.[22]

5. The sanction of trust-regret

Talk of "fair" or "unfair" regret makes little sense if the regret does not impose a sanction. What sanction might trust-regret impose? The no-regret condition specifies how the sanction of regret functions in prospect: when you expect that your twice-future self will prove unable to make the right sort of sense of how this action could be retrospectively attributable to your once-future self, you cannot now make the right sort of sense of how you could form the intention and thereby prospectively attribute this action to your once-future self. To understand more fully how this works, it helps to develop the analogy between intra- and interpersonal agency suggested by the similarities between intra- and interpersonal trust.

When you accept another's invitation to share an intention, it makes sense to regard you as entering into an interpersonal trust relation premised on that other's expectation that you'll not regret the union. Your co-worker invites you to accompany her on a tour of the premises, but in trusting her (she gently took your arm, and you reciprocated by walking with her) you played right into one of her schemes – what on earth were you thinking? ("A 'tour'?" you ask yourself from plan's end. "How could I have fallen for that scheme?") This species of regret addresses whether the agential union "speaks for" you in the sense at issue in this paper: are its actions

attributable to you as agent (that is, as one of its agents)? Call this an *identification-expressive* species of regret. Since the invitation to share the intention is an invitation to be thus identified with the action, the invitation must represent itself as manifesting the projection that you will thus identify with the action. But then it must represent itself as manifesting the projection that you won't experience fair identification-expressive regret concerning the action. The need for this projection lies in the invitation's claim of an interpersonal analogue of agential authority – not the union between parts of your self that gets a unified you into your behavior, but the union between two agents that gets them both as a unified "we" into an instance of shared agency. Why should you let your co-worker influence you in the way that she proposes? As she takes your arm, she may be explicit: "Let's take a little tour. Trust me. You won't regret it!"[23]

I'm arguing that the same identification-expressive species of regret shapes the intrapersonal case. What you aim to avoid when you intend to perform an action is a species of regret that amounts to the paradoxical – and therefore unsettling – feeling that the action does not speak for you. If it does not speak for you, how could it be attributable to you? But you attributed it to yourself – you staked an implicit claim of attributability – when you entered into the self-trust relation! From this retrospective perspective the sanction registers as *the disorientation you feel when this self-attribution turns out to look wrong* – wrong (in light of the fairness norm) in a way for which it would be fair to hold you accountable. The regret is thus a form of self-accountability: you staked a claim of self-attribution in which, as things now turn out, you cannot recognize yourself. Prospectively, the sanction registers as an inability to presume the intrapersonal authority over your acting self that defines an intention. As we've seen by reflection on toxin cases, you simply cannot form the intention without holding yourself thus accountable. My suggestion, then, is this: just as another cannot coherently claim the

authority inherent in inviting you to share an intention while representing herself as projecting fair regret that you entered into the relation, so you cannot claim the authority inherent in forming an intention while projecting fair regret that you followed through on the intention.

An understanding of this sanction explains why you should care about your own future regret. But why should you care about your expectations of future regret? Since in expecting to regret you give evidence of your untrustworthiness in intending, we can rephrase the question: why care about your own status as trustworthy or untrustworthy in intending? Even if you have other reasons to care, can you derive a reason simply from the thought that your twice-future self will regret your untrustworthiness? Here is my proposal: without projecting a relevantly regret-free future, you cannot find your claim of agential authority intelligible. And without finding that claim intelligible you cannot count as making it.[24]

Forming an intention differs on this point from inviting another to share it. You can make agential sense of a presumption of authority over another when you know that you are not relevantly trustworthy. (Your co-worker knows that her self-presentation is insincere.) But you cannot make agential sense of such a presumption of authority over yourself. You can make causal sense of this presumption, if you expect to succeed in deceiving yourself in relevant ways. But, as the problem of "deviant" causation for causal theories of action makes vivid, mere causal intelligibility does not suffice for intelligibility as an action.[25] We can explain this difference by observing that, as many philosophers have emphasized, you cannot form an intention for the "autonomous benefits" of forming it.[26] When you form an intention to φ at *t*, you must deliberate only from considerations that concern your φing at *t*. This, again, is why you cannot form an intention to drink the toxin – thereby getting the autonomous benefit of a million dollars – in Kavka's puzzle. By contrast, to count as sincerely inviting someone to share your intention

to φ at *t*, while you must present yourself as deliberating from considerations that manifest an appropriate concern for your invitee's φing at *t*, you needn't actually feel that concern at all. If an eccentric billionaire offered to reward you for inviting someone to share a putative intention to drink toxin, you could easily do it. The only rub would lie in your confidence that your invitee wouldn't understand your motive in inviting him to drink and would form his intention to drink entirely on the basis of trust in you. If he thought the case through on his own he would be unable to claim authority over his conduct, but simply trusting your claim of authority he can. Knowing this, you can invite him to claim that authority – thereby bagging your reward.

Note well that the restriction on individual intention doesn't derive from your *owing* anything to your twice-future self. This is not like the relation between you and someone who calls you out on wrongful conduct. Nor is it like the relation between your intending and acting selves when the latter fails to trust the former. It is not, in sum, a justificatory or more broadly forensic relation in which you purport to give reasons. In terms of its content, it is a *finding-intelligible* relation.

I'm arguing that this finding-intelligible relation involves the species of accountability also at issue in forensic relations. We can see that this is a form of accountability by understanding how the unsettled feeling at the core of trust-regret functions like other sanctions in keeping you "in line." You view yourself as accountable to your twice-future self insofar as your aim to avoid this disorientation guides your practical reflection from the onset of deliberation to the formation of your intention. Your practical-reflective stance actually extends farther than that, since it is up to you whether to re-open deliberation in the interval between forming the intention and acting on it. The self-accountability at the core of your self-

intelligibility thus guides you from deliberation to action, as you project a regret-free future all the way out to plan's end.

The sanction reveals how trust and regret are both acknowledgments of risk. On a natural but mistaken view, trust would manifest a vulnerability to specifically personal influences, over and above our vulnerability to the rest of the world, and regret would register all this vulnerability in retrospect as disappointment that things did not go your way. That misses what's most interesting in both attitudes. Agential self-trust is not mere self-reliance, nor regret mere disappointment, because what each attitude most fundamentally registers is the basis of enkratic rational requirements.[27] Self-trust registers this basis prospectively by positing the absence of what regret registers retrospectively. This basis is an intrapersonal relationship that such vulnerability – while very real – does not define. It would not be the relationship that it is without that vulnerability, but that is because the risks define a genus of broader trust relationships to which self-trust relations belong, not because they specifically define self-trust relations.

Trust amounts to a way of *coping constructively* with these risks, by putting the agency of another into the service of your own – where that "other" may be your own earlier self. Typically there is no other way to proceed. Much work on trust starts from the thought that it's risky to act in a way that relies on others, and much work on regret starts from the thought that it's generally risky to act. Yet if you did not know how to trust yourself, thereby risking the sanction of trust-regret, you would not be capable of acting commissively through time.

Bratman, Michael
1999a *Faces of Intention* (Cambridge: Cambridge University Press)

1999b "Cognitivism about Practical Reason," in Bratman 1999a
1999c "Toxin, Temptation, and the Stability of Intention," in Bratman 1999a
2007a *Structures of Agency* (Oxford: Oxford University Press)
2007b "Introduction" in Bratman 2007a
2007c "Temptation Revisited" in Bratman 2007a
2007d "Three Theories of Self-Governance," in Bratman 2007a,
2007e "Two Problems about Human Agency," in Bratman 2007a
2007f "Valuing and the Will," in Bratman 2007a
2014 "Temptation and the Agent's Standpoint," *Inquiry* 57:3

Davidson, Donald
1980 "Freedom to Act," in his *Essays on Actions and Events* (Oxford: Oxford University Press)

Farrell, Daniel
1989 "Intention, Reason, and Action," *American Philosophical Quarterly* 26:4
1993 "Utility-maximizing Intentions and the Theory of Rational Choice," *Philosophical Topics* 21:1

Frankfurt, Harry G.
1988 *The Importance of What We Care About* (Cambridge: Cambridge University Press)

Hinchman, Edward S.
2003 "Trust and Diachronic Agency," *Noûs* 37:1
2009 "Receptivity and the Will," *Noûs* 43:3
2010 "Conspiracy, Commitment, and the Self," *Ethics* 120:3
2013 "Rational Requirements and 'Rational' Akrasia," *Philosophical Studies* 166:3
2015 "Narrative and the Stability of Intention," *European Journal of Philosophy* 23:1
2016a "On the Risks of Resting Assured: An Assurance Theory of Trust," in Paul Faulkner and Tom Simpson (eds), *New Philosophical Perspectives on Trust* (Oxford: Oxford University Press)
2016b "How to Settle on a Shared Intention," unpublished manuscript
2016c "Intention and Time," unpublished manuscript
2016d "What is the Rational Stability of Intention?," unpublished manuscript

Holton, Richard
2009 *Willing, Wanting, Waiting* (Oxford: Oxford University Press)

Kavka, Gregory
1983 "The Toxin Puzzle," *Analysis* 43:1

Velleman, J. David
1989 *Practical Reflection* (Princeton: Princeton University Press)
2000 "Introduction" to his *The Possibility of Practical Reason* (Oxford: Oxford University Press)

Wallace, R. Jay
1994 *Responsibility and the Moral Sentiments* (Cambridge: Harvard University Press)

Watson, Gary
1996 "Two Faces of Responsibility," *Philosophical Topics* 24:2

2005 "Hierarchy and Agential Authority," in John Fischer (ed.), *Free Will: Critical Concepts in Philosophy* (New York: Routledge), volume IV

Williams, Bernard

1981 "Moral Luck," in *Moral Luck* (Cambridge: Cambridge University Press)

Wittgenstein, Ludwig

1953 *Philosophical Investigations* (Oxford: Basil Blackwell)

Wolf, Susan

1990 *Freedom Within Reason* (Oxford: Oxford University Press)

---

[1] Bratman coined the term "plan's end" (1999c, 85*ff*). For some qualifications on how I'm going to use the term, see note 7 below. See note 3 for more on Bratman's approach.

[2] This paper extends the inquiry into how norms of trust structure diachronic agency that I began in Hinchman 2003 and 2009.

[3] For Bratman's earlier approach, see his 2007f, 56. This restates his justification of the no-regret condition in Bratman 1999c, where he first proposed the condition. Bratman 2007c emphasizes the difference between the pragmatic orientation of his own earlier treatment and a new orientation toward agential authority, noting how these orientations would make different use of an appeal to regret. Bratman 2014 appeared too late to discuss here; it revises his position in one respect that I criticize in Hinchman 2016d (see notes 5 and 7 below).

[4] See Bratman 2007e, where he says "necessary and sufficient" (92). Bratman later states that he no longer wants to formulate the problem of agential authority in terms of necessary conditions (2007b, 4-5 and 11).

[5] For "metaphysical imperative," see Gary Watson's interpretation of Bratman (Watson 2005, 94-5), an interpretation which Bratman has subsequently confirmed, characterizing his own 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)" (2007d, 246); Bratman says in a footnote that he is replying to Watson's interpretive claim. Bratman's 2014 formulation may no longer rest on this imperative, though abandoning it raises problems that I discuss in Hinchman 2016d.

[6] For the original case, see Kavka 1983.

[7] My formulation makes somewhat more precise Bratman's characterization of plan's end as "the conclusion of one's plan" (1999c, 86). It makes sense to prefer my more precise formula over his vaguer phrase 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 (2009, 158) presses this latter observation against Bratman.) Bratman's phrase does appear to differ from my formulation insofar as appeal to the "conclusion" of the plan appears to exclude death-bed conversions and the like. While one might find that an attractive implication in some instances (why should death-bed delirium matter?), the exclusion imports a substantial assumption about diachronic agency that could not be vindicated in all instances (why shouldn't conscientious death-bed reflection matter?). Note well that any such differences between my formulation and Bratman's will not make any difference for my argument against Bratman's

understanding of plan's end, since my argument will target not such details but the general idea that plan's end must be expected (rather than projected).

[8] 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 agential authority.

[9] Note that Bratman uses "projects" alongside "anticipates" to refer to the forward-looking attitude that is governed, in part, by expectations about plan's end (e.g. 2007c, 275). That is, he doesn't draw the principled distinction between expectation and projection on which I am insisting.

[10] Why not? It points in the right direction to observe that, unlike a "projection," a mere expectation does not amount to anything worth describing as "telling a story." When you form the present intention, it is plausible to regard you as projecting a plan's-end perspective from which you'll resonate to this "tragic" narrative: "giving my parents their due before I become 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. I hypothesize that a mere expectation fails to stabilize your intention because it lacks the element distinctive of "telling a story" about your future. For more on this appeal to narrative, see Hinchman 2015. But my present argument does not require that particular elaboration.

[11] I treat this form of sense-making at greater length in Hinchman 2016c.

[12] 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" (e.g. 1999c, 88, and 2007f, 56). In any case, he does not consider the possibility of regretting your relation of self-influence as opposed to your action.

[13] By the no-regret condition, you must also anticipate that you would regret it if you failed to realize that relation. But because you typically (albeit not necessarily) expect that you will follow through, this hypothetical anticipation does not play the same role as the categorical one.

[14] For references and discussion, see note 5 above.

[15] I'm not sure this really is so, but I lack space to investigate the issue.

[16] Note this respect in which accountability, while different from attributability, nonetheless presupposes self-attribution. I pursue further the idea that anticipating regret amounts to a species of self-accountability in Hinchman 2010 and 2016a.

[17] Both the resolution not to take drugs in general and the intention to obtain drugs now are intentions. I call the first a "resolution" merely for ease of reference, and to remind us of its status in making the other intention akratic. Having both intentions at once is a contradictory state of mind, and this incoherence poses the traditional problem of akrasia: how could the contradiction (i.e. acting against your own intention or resolution) fail to undermine attributable agency? I cannot treat the traditional problem here (I do so in Hinchman 2009 and 2013), beyond noting that taking seriously the problem of the rational stability of intention – how could it be rational to follow through on an intention that you would abandon if you reconsidered? – presupposes that we can solve the traditional problem of akrasia. If it were simply impossible to act (sc. with attributable agency) against your own intention without thereby counting as having changed your mind, then the "rational stability" of intention would amount to a rational pressure not to change your mind. But that is not the species of stability at issue here, and it is not clear that there actually is any such species of rational stability. The

species of rational stability that gives the content of intention its point provides rational pressure not against changing your mind about what to do but against failing to follow through on your intention *without* having changed your mind about what to do. (This point undercuts Bratman's 2014 attempt to broaden rational stability beyond mere non-reconsideration; see my 2016d.)

[18] "Agent-regret" is Bernard Williams's term (1981, 27*ff*).

[19] Henceforth I'll resume my practice of mostly leaving the qualifier implicit.

[20] Wolf 1990, Wallace 1994, and Watson 1996 have in different ways argued for this proposition.

[21] Here is where my approach calls for something like the elaboration sketched in note 10 above.

[22] Whichever elaboration we prefer, we'll need to explain how the behavior at issue counts as a single and to that extent unified instance of self-governance. I pursue these issues in Hinchman 2016c.

[23] I develop such an account of shared intention in Hinchman 2016b.

[24] I agree here with David Velleman: the constitutive aim of diachronic agency is a species of self-intelligibility (1989, 94-100 and Chapter 4; and 2000). But I disagree with Velleman's interpretation of this insight at two key points: (a) my account is not "cognitivist" (in the respect criticized by Bratman (1999b); and (b) I hold that self-intelligibility matters specifically in the claim of agential authority at the core of an intention.

[25] For "deviant" causation, see Davidson 1980, 63-81. The older issue of action explanation unfolds from a third-personal perspective, whereas our issue unfolds within the first-personal perspective of the agent. Mapping our issue of agential intelligibility onto that older issue, despite this difference, we might view *mere* causal intelligibility – for example, you expect to forget your actual deliberative basis for deciding to drink the toxin, deceiving yourself into believing that the billionaire requires that you drink – as a form of causal deviancy: your decision-making causes your own behavior without thereby amounting to the uncompromised performance of an action because it does not cause it "in the right way," i.e. in a way that engages your deliberative basis for acting.

[26] For some early articulations of this general point, see Farrell 1989 and 1993 (esp. 58-9).

[27] I explain how norms of trust inform enkratic rational requirements in Hinchman 2013.