Conspiracy, Commitment, and the Self*

Edward S. Hinchman

Practical commitment is Janus-faced, looking outward toward the expectations it creates and inward toward the basis of these expectations in the agent's will. Sometimes when you say that you have committed yourself, you mean that you've undertaken the commitment: that you've made that commitment 'to yourself'. Other times you mean that you've—sincerely or insincerely—made the commitment to others: that others may now rely on you to follow through. What is the relation between a commitment that you credibly—however insincerely—make to others and a commitment that you make 'to yourself'? *Promising*, *assuring*, and *pledging* are examples of the former sort of commitment; *choosing*, *intending*, and *resolving* are examples of the latter. How are the two types of commitment related?

We naturally assume that making a commitment 'to yourself' is just committing yourself, sans indirect object, and one reason for the assumption is that it's not obvious how you could do so insincerely. It *is* obvious that you can insincerely present yourself as committed to others, even credibly so (provided you can pull off the deception). But noting that difference presupposes an answer to the question we're considering. What makes it possible to present yourself insincerely as committed is that whether you are genuinely committed—that is, 'to yourself'—can come apart from whether you're committed 'to others'. But how is the separation possible? How, in general, is giving others 'your word' like and unlike having a 'word'—that is, a commitment—to give in the first place? I'll pursue an approach to answering this question that treats the interpersonal and intrapersonal cases in parallel as far as we can. My hypothesis is that we can learn something important about both by asking where they diverge.

* I wrote the first draft of this article while holding a fellowship at the Center for 21st Century Studies at the University of Wisconsin–Milwaukee. Thanks to the other fellows at the center that year for discussion and especially to Julius Sensat for his full comments on that draft. Thanks to several referees and editors at *Ethics* for helpful commentary on further drafts and to Bill Bristow and Andrea Westlund for discussion.

Ethics 120 (April 2010): 526–556
 0014-1704/2010/12003-0002$10.00

Are Kantians right, I'll more specifically ask, that a reason to refrain from insincerity emerges from the very nature of practical commitment? I structure my discussion by criticizing alternative accounts recently offered by two moral philosophers, David Velleman (Sec. I) and Christine Korsgaard (Sec. II), who approach the issues in a broadly Kantian spirit. Velleman aims to show that we cannot separate the interpersonal question (how to present your commitments to others) from the intrapersonal question (how to commit yourself) because you cannot commit yourself unless you can find yourself intelligible from your interlocutor's point of view. I agree that we cannot separate these questions, but I reject Velleman's account of why that is so. The core of his argument is valid but not sound. If thinking instrumentally about a commitment could not generate a stable self-understanding, as Velleman argues, then we would all have reason to make a prior decision to think not instrumentally but 'expressively', where to think expressively is to think of how you present yourself as expressing how you actually are.[1] But there is no good reason to believe the antecedent of that conditional. Velleman's account goes wrong in identifying the perspective from which you must make yourself intelligible with the perspective of the interlocutor to whom you are trying to make it credible that you are committing yourself. This is not the only available external perspective from which to make sense of your conduct, since there is also the perspective of a would-be co-conspirator in deception.

I'll argue that the availability of this conspiratorial perspective is what allows you to make sense of yourself as deceiving your interlocutor. It can provide such a basis, I'll argue, by providing the basis of a self-understanding as conspiring—whether your co-conspirators are real or merely imaginary. I'll then argue that the availability of such a third perspective—not yours, not that of the 'target' of your practical thinking—is what enables you to commit yourself 'to yourself'. Here I offer a basis for rejecting Korsgaard's account of practical commitment, on which forming an intention requires identifying yourself with a principle that you expect will continue to guide you when the time comes to act.[2] You needn't identify yourself with a principle, I'll argue, because

1. For this formulation of Velleman's argument, see J. David Velleman, "The Centered Self," in his *Self to Self: Selected Essays* (Cambridge: Cambridge University Press, 2006), 253–83. "The Centered Self" provides a pithy formulation of a core strand in Velleman's recent book, *How We Get Along* (Cambridge: Cambridge University Press, 2009)—as he makes clear, for example, on 17 and in his treatment of the 'blood feud' example on 46–48.

2. See Christine Korsgaard, *Self-Constitution: Agency, Identity, and Integrity* (Oxford: Oxford University Press, 2009), "Self-Constitution in the Ethics of Plato and Kant," in her *The Constitution of Agency: Essays on Practical Reasons and Moral Psychology* (Oxford: Oxford University Press, 2008), 100–126, *The Sources of Normativity* (Cambridge: Cambridge Uni-

(putting the point in Kantian terms) the 'unity' at which you 'aim' when you commit yourself to act is a unity not with your acting self but with a later perspective, a relation that needn't be mediated by any principle shared with your acting self. This 'twice-future' perspective—neither your present intending self nor your (once-) future acting self but a third perspective that looks back on how the first two are related—plays the intrapersonal role played in interpersonal commitment by potential co-conspirators.

My talk of this 'twice-future' perspective echoes a 'No-Regret' condition coined by Michael Bratman to explain the stability of intention.[3] But I'll raise a counterexample to Bratman's account (Sec. III), a case that reveals how practical commitment is structured by a complex species of self-trust. My alternative account of commitment reorients the No-Regret condition toward this question of self-trust (Sec. IV), proposes a new conception of practical intelligibility (Sec. V), and uses that conception to explain precisely where Kantian conceptions go wrong (Sec. VI). The core of the proposal distinguishes a genuinely interpersonal intelligibility constraint on the *deliberative* authority that you claim when you form a practical judgment from a purely intrapersonal intelligibility constraint on the *executive* authority that you claim when you commit yourself to that judgment by making a choice or forming an intention.[4] Deliberative authority is subject to a forensic constraint, since it looks 'outward' toward interpersonal contexts of justificatory challenge. Executive authority is subject not to a forensic but to a metaphysical constraint, since by contrast it looks 'inward' toward the intrapersonal task of self-constitution. The Janus-faced nature of practical commitment

---

versity Press, 1996), "Morality and the Logic of Caring: A Comment on Harry Frankfurt," in Harry Frankfurt, *Taking Ourselves Seriously and Getting It Right*, ed. Debra Satz (Stanford, CA: Stanford University Press, 2006), 55–76, and "Personal Identity and the Unity of Agency: A Kantian Reply to Parfit," *Philosophy & Public Affairs* 18 (1989): 101–32.

3. Michael Bratman, "Toxin, Temptation, and the Stability of Intention," in his *Faces of Intention* (Cambridge: Cambridge University Press, 1999), 58–90. See also Michael Bratman, "Valuing and the Will" and "Temptation Revisited," in his *Structures of Agency* (Oxford: Oxford University Press, 2007), 47–67 and 257–82; and Edward Hinchman, "Narrative and the Stability of Intention" (unpublished manuscript, University of Wisconsin–Milwaukee, 2010).

4. As I'll use the terms, 'choice' and 'intention' mark practical commitments. In this usage, an intention is always a diachronic commitment, with an expected temporal gap between the formation of the intention and the action, which thereby counts as 'following through' on the intention. A choice, by contrast, can be simultaneous with the chosen action. (Some philosophers use 'intention-in-action' or similar terms to mark the synchronic case, but I suspect that that term marks a third notion. What Velleman calls an 'immediate intention', I'll call simply a choice.) The distinction won't matter till later sections, when it will become useful to pretend that practical commitment is always diachronic. See n. 38 below.

reflects the different orientations of these two species of practical authority.

I

Consider first how Velleman develops his argument. When you act purely instrumentally in a context with the payoff structure of a prisoner's dilemma, you seek to deceive your "interlocutor"—as we'll continue to call the other player—about how you will act.[5] You say, "Okay, I'll cooperate," hoping to induce him to cooperate. But you perform this speech act with an intention to defect, since defecting is instrumentally preferable to cooperating, no matter what your interlocutor does. In performing the deceptive speech act, you aim at the best outcome for you, which is the outcome in which your interlocutor cooperates and you defect. Can you imagine this strategy from your interlocutor's perspective? Well, you can imagine him deceived into believing that you intend to cooperate. But that is to imagine him deceived about your strategy; it isn't to imagine your strategy. To imagine your strategy from his perspective, you have to imagine that he sees through it. But wait: if you imagine him detecting your strategy, you must imagine yourself abandoning the strategy—since, after all, you're imagining the strategy pointless. Since you're imagining how what you're really doing will appear from your interlocutor's perspective, you must now imagine your interlocutor detecting your abandonment of the original deceptive strategy. But wait again: if you imagine him detecting that you've abandoned the deceptive strategy, you imagine that the strategy should not be abandoned—since now it might work. And so it goes, on and on, round and round this loop. When you imagine yourself implementing this strategy, you imagine a context in which you should abandon it. But when you imagine yourself abandoning the strategy, you imagine a context in which you should not abandon it. From your interlocutor's perspective, in sum, the instrumental thinking that gives rise to your strategy is inherently unstable.[6]

5. In the classic prisoner's dilemma, of course, the players cannot communicate with each other. But, as everyone notes, a new game in which players are allowed to say to each other "I promise I'll cooperate" is simply a new version of their earlier predicament, with defection now defined as breaking the promise.

6. Here is how Velleman puts the point ("Centered Self," 269): "As soon as I begin to think instrumentally in this case, I enter a dizzying spiral of anticipating that my instrumental calculations have been anticipated, that their validity has thus been compromised, that their being so compromised has also been anticipated, with the result that they gain new validity, which has of course has been anticipated, and so on. Hence the best instrumental understanding that I can achieve of what I am doing, if I offer to cooperate in these circumstances, is that I am taking a shot at being trusted, a shot whose prospects of success are obscured by endless complications."

Velleman derives a strong conclusion from this argument: that as an agent you face a prior choice between two metastrategies, the instrumental strategy that leads to this unstable self-understanding and an expressive strategy that does not. The expressive strategy embodies a commitment to represent yourself as having the first-order strategy that you in fact have. It is a commitment to represent your strategies honestly. Since your agency is compromised by an unstable self-understanding, if you are in the business of agency you have a reason to choose the expressive metastrategy and commit yourself to honest self-representation. Velleman does not, of course, claim that every instance of dishonesty generates the degree of hermeneutic instability generated by an instrumental strategy in a cooperative dilemma. Moreover, he allows that just as self-understanding comes in degrees, so must agency itself. His conclusion is not that you cannot act dishonestly but that its greater intelligibility gives you a reason to choose the expressive alternative.[7] This concession will not matter. The problem is not that Velleman has failed to rule out dishonesty but that he hasn't presented the slightest reason to disprefer it.

We may accept the major premise of his argument: the claim that agency is compromised when it is not informed by a stable understanding.[8] It is intuitively plausible enough that you cannot govern yourself if you do not know what you're doing—where to say that you don't

7. As he puts it (ibid., 270): "I do not claim to have shown that the rational pressure in favor of sincerity always prevails. In particular, there are extreme losses that it makes sense to take a shot at avoiding, and extreme gains that it makes sense to take a shot at obtaining, no matter how wild or how blind a shot. But there are many gains and losses that it makes more sense to ignore, given the more intelligible alternative of speaking our minds. And what's more intelligible is, on my view of practical reason, the more rational course to take."

8. Velleman argues for his view of interpersonal commitment by first arguing for a view of agency and practical reason. (His "Centered Self" offers a condensed version of this argument; earlier versions can be found in the introduction to J. David Velleman, *The Possibility of Practical Reason* [Oxford: Oxford University Press, 2000], 1–31, and, most fully, in *Practical Reflection* [Princeton, NJ: Princeton University Press, 1989].) In the present discussion, we may ignore this larger argument because it is not required by the specific argument that we'll consider. What Velleman offers is an account of agency that, if true, would explain why an agent cannot act from an unstable self-understanding. But it might, of course, be true that an agent cannot act from an unstable self-understanding, though Velleman's explanation of why this is so is false. Since far more philosophers would be willing to grant this thesis than would be willing to grant Velleman's explanation of it, it seems best to start there and not worry why the thesis is true. (We'll consider a different argument for the thesis in Sec. V.) We'll assume that you can't make a choice or form an intention to act unless that choice or intention is informed by a self-understanding that is both coherent enough and stable enough. Those two 'enough's' leave open for further inquiry how to define or explain the bearing of the relevant notions of coherence and stability.

know what you're doing is to say that there is no description available to you under which your proposed course of action fully makes sense to you.[9] The core of Velleman's argument is, in effect, that a description can fully make sense to you only if it can make sense to anyone who understands it as you do. His minor premise is thus that an instrumental strategy cannot fully make sense to you in a cooperative dilemma because it cannot make sense to just anyone who understands it as you do—including, in particular, your interlocutor in the dilemma. This we should not concede.

Velleman's argument wrongly equates objectivity with universality. We may agree that you cannot understand what you are doing unless there is a description available to someone other than you under which what you're doing can make sense to that person. That ensures that your self-understanding is publicly available, thereby enabling it to figure in the order of justification (the 'space of reasons') on which—arguably, at least—any thought content must have appropriate bearing. It does not follow, of course, that the description must make sense to just anyone. In particular, it need not make sense to your interlocutor in the cooperative dilemma.

In fact, this difference in external perspectives seems to figure directly in your instrumental strategy in the dilemma. What figures in the strategy is not merely the distinction that Velleman discusses, between your narrowly subjective take on what you're doing and a broader objective take. The distinction that you need is not that distinction but the distinction between deceivers and the deceived. The deceiving perspective needn't be narrowly yours, since you can share it with a co-conspirator. What you need in pursuing an instrumental strategy is the distinction between those who are parties to the deception and those who are not. The former category will of course include yourself, but it typically may also include others. If your deception cannot include others in a given case, that will be because of details about the case and not merely because it is a case of deception. Both of these perspectives, that of the deceivers and that of the deceived, are objective perspectives in at least one important sense of the term. Though neither is a universal perspective—since neither is available to those who adopt the other perspective, at least so long as they're adopting it—it is also true that neither is a subjective perspective, unavailable to others in principle. Since you can make sense of what you're up to to a potential co-conspirator when you pursue an instrumental strategy in the prisoner's

9. Note that this is a point about self-governance, not about actions generally. Though it doesn't matter to the present argument, this leaves open the possibility of performing actions without governing yourself. See nn. 32 and 37 below.

dilemma, why should it matter that what you're up to cannot make sense to your interlocutor (lest the strategy be compromised)?

Why, in general, must an objective perspective universalize? Step back and consider how your strategy will look to a potential co-conspirator—that is, to a possible person whom you can imagine in the role. Once again, to imagine how the strategy looks from this perspective, you have to imagine that the co-conspirator sees through it. That was what created the problem when you imagined how it would look from your interlocutor's perspective: when you imagine your interlocutor seeing through your strategy, you imagine a strategy that is no longer justified and must be abandoned. But that problem does not arise here. The fact that a potential co-conspirator—someone whom, by hypothesis, you are not trying to deceive—sees through your strategy does not by itself give you a reason to abandon it. To a potential co-conspirator, it will look as if you're pursuing a strategy that depends on its not being detected by those whom it targets. This is a perfectly intelligible thing to try to do; the fact that it is perfectly intelligible is revealed by the intelligibility—in principle—of recruiting others to your instrumentalist cause. You encounter no difficulty understanding yourself in these terms—again, in principle—because you can imagine this recruitment effort.[10]

II

Why must an objective perspective universalize? Since Velleman's argument simply assumes it must, we don't yet have an answer. Let's now consider the answer given by another Kantian, Christine Korsgaard. This brings us to the other issue of practical commitment: commitment not to others but 'to yourself'.[11] Once again, the problem will be that the Kantian argument overlooks how you can make yourself objectively intelligible by making yourself intelligible to a third party—in this case, a twice-future self of yours whom you expect to arrive after your once-future self has followed through on your commitment.

10. Let me emphasize that practical self-intelligibility does not, of course, require the actual availability of co-conspirators but merely a capacity to imagine them. How this imaginative capacity is like and unlike a capacity to universalize the maxim of your action is an interesting question.

11. What Velleman calls the 'problem of commitment' (as opposed to the 'problem of credibility', which we've been treating) presents a third issue, codified by the question of whether to follow through on your commitment (Velleman, "Centered Self," 270ff.). That is not the issue that we're about to consider. The issue we're about to consider is codified by the question of 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" (in both "Centered Self" and *Practical Reflection*, chap. 8), see Edward Hinchman, "Trust and Diachronic Agency," *Noûs* 37 (2003): 25–51, 47 n. 14, and sec. 8.

The problem goes to the heart of Korsgaard's long-standing project in normative theory. In her 2002 Locke Lectures, published in 2009 as *Self-Constitution*, Korsgaard revisits an issue left over from her 1992 Tanner Lectures, published in 1996 as *The Sources of Normativity*.[12] In that earlier work, she tried to derive the Kantian thesis that whenever you act you manifest a commitment to the value of rational reflection, and thus of Kantian 'humanity', from the premise that you cannot act without identifying yourself with a general principle that justifies your action. When you identify yourself with a general principle, she argued, part of what constitutes your identity is your identification with the species of rational reflection that enables you to identify with principles. In the latter identification, you express your commitment not only to the value of your own humanity but to the value of humanity as such. You are thereby committed to treating others only in ways that do not violate their humanity—all merely because, like all of us, you cannot avoid the need to act.[13]

This argument presupposes that you cannot act from what G. A. Cohen, in his reply to Korsgaard, called "singular edicts" or with what Korsgaard in more recent work calls a "particularistic will."[14] But why can't you? Particularistic willing would be willing for this one case only, with no regard for other cases. In her more recent lectures, Korsgaard provides a fuller argument for her premise that you cannot act with a particularistic will. This argument has two steps:[15]

Step One: When you act, you must regard yourself, rather than (*a*) another person or (*b*) some mere incentive *in* you, as the cause of your action.

Step Two: You cannot draw the distinction between you and a mere incentive *in* you when you "wholly identify with the incentive of your action." But wholly identifying with the incentive of your action—without regard for other occasions—is what particularistic willing would come to. So acting with a particularistic will is impossible.

In considering that argument, we'll naturally wonder why particularistic willing prevents you from viewing yourself as the cause of your action. How does it follow that the incentive or inclination cannot be you? Korsgaard doesn't let this question arise, because in her statement of

12. See also Korsgaard, "Self-Constitution in the Ethics of Plato and Kant."

13. For the core of this argument, see Korsgaard, *Sources of Normativity*, 120–23.

14. For "singular edicts," see G. A. Cohen, "Reason, Humanity, and the Moral Law," in Korsgaard, *Sources of Normativity*, 176.

15. Korsgaard, *Self-Constitution*, 75–76; see also Korsgaard, "Self-Constitution in the Ethics of Plato and Kant," 123–24.

step two she merely assumes that if the incentive is to be you, you can't regard it merely as an incentive in you. But the question simply is why you can't regard an incentive *in* you *as* you! Does Korsgaard give any argument why you can't?

The argument seems implicit in her remark that in particularistic willing the agent would become "not one person, but a series, a *mere heap*, of unrelated impulses."[16] Since 'series' seems to qualify 'heap', rather than vice versa, it looks as if the problem is that the particularistic willer will not be able to achieve diachronic unity: acting on a series of impulses is not, after all, like following through on an intention.[17] But Korsgaard seems to be assuming that the only way to achieve this diachronic unity is for the unity to be represented in the content of your will: as a principle—let's call it a 'principle of choice'—shared between intending and acting selves. Since the content of a particularistic will fails—let's grant—to represent such unity, no particularistic will can count as unified. But then why can't the unity be achieved in some other way?[18]

Korsgaard's argument relies on the claim that there is no other account that can explain the distinction articulated in step one, between being moved by an inclination and governing that inclination. She assumes that the only way to explain this distinction is to view self-governance as identifying with a principle of choice. We'll see that this assumption is false by developing an alternative account in terms of which we can explain the distinction, an alternative suggested by the form of the problem that arose for Velleman. We may accept Korsgaard's

16. Korsgaard, *Self-Constitution*, 76.

17. This reading is confirmed by Korsgaard's formulation of the argument in "Morality and the Logic of Caring," 64–65. (In the corresponding passage of "Self-Constitution in the Ethics of Plato and Kant," Korsgaard says "conglomeration" instead of "heap" [124].) Much earlier, Korsgaard argued that momentary persons are not really persons, since being a person requires being capable of practical commitment: see Korsgaard, "Personal Identity and the Unity of Agency."

18. Since the issue of diachronic agential unity rests on an issue of rational redeliberation, Michael Bratman asks the rhetorical question like this: "Why couldn't I reach a decision about the future and trust that I would make a reasonable judgment about reconsideration if the issue arose, without endorsing some nontrivial general principle that says when to reconsider and when not to?" ("Review of Korsgaard's *The Sources of Normativity*," *Philosophy and Phenomenological Research* 58 [1998]: 699–709, 708–9). We'll consider Bratman's positive view in Sec. III; but in the meantime, observe that a more fundamental role for trust in this context is backward looking. When you form an intention, you do trust your future self to reconsider only when it would be reasonable to do so. But when you act without reconsidering—that is, when you simply 'follow through' on the intention—you trust your intending self. The attitude of trust informing your intention thus targets the trust you would manifest in executing the intention: you trust your future self to act on the basis of trust (i.e., to follow through without reconsidering) when but only when such backward-looking trust would be reasonable. See Sec. IV.

thesis that a successful account of agency must provide a basis for this distinction, since what we mean by 'agency' and 'acting' here depends on a contrast with the case in which one 'acts' only because one is acted upon. Even so, we may deny that only her account can provide this basis. The present account will not appeal to identification with a principle of choice but will instead emphasize the need to make yourself intelligible to the intrapersonal analogue of a co-conspirator.

Interpersonally, agency is perspectivally articulated in such a way as typically to make a co-conspirator's perspective on your action synchronic with your own. But the role played by the co-conspirator in the problem for Velleman does not involve his actually acting with you. The co-conspirator's role is to provide an external perspective that serves as a check on your ability to make the right sort of sense of what you're doing. The intrapersonal analogue of these perspectival relations, by contrast, unfolds diachronically, which simply precludes any analogue of acting 'with' yourself. If we pursue the analogy, the diachronic articulation of these intrapersonal relations means that a later perspective of your own must play the sense-making role played by co-conspirators in the problem for Velleman. As we'll see, to find what you're doing intelligible in the way that serves as a necessary condition on the action's being attributable to you, you must expect to make sense of what you're doing both to a potential other person in the role of co-conspirator and to your own future self in an intrapersonal analogue of that role. We'll consider the two-sidedness of the requirement in Section V. It is this intrapersonal requirement—the requirement that you expect to find yourself intelligible from your own later perspective—that plays the role that the intelligibility of principle plays in Korsgaard's account.

III

I'll motivate this alternative approach to intrapersonal practical commitment by addressing a case in which such commitment is impossible: Gregory Kavka's well-known Toxin Puzzle.[19] I'll focus the issue of intrapersonal practical commitment by developing a counterexample to diagnose a problem for Michael Bratman's otherwise plausible treatment of that case.[20]

Compare two scenarios. First, imagine the sort of case described by Kavka. Imagine that an eccentric billionaire with a reliable intention

19. Gregory S. Kavka, "The Toxin Puzzle," *Analysis* 43 (1983): 33–36.

20. Since it's a case of diachronic commitment, that will focus the discussion on intention—that is, on 'future-directed' intention, not on intention 'in action'. We'll approach the question of commitment by asking how the earlier self of yours that forms an intention can presume to exert executive authority over the later self that acts on it. (On how to generalize the account to 'intention in action' or mere synchronic choice, see n. 38 below.)

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 will 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.[21]

Now compare that case with the following variant. Imagine everything as above, but further imagine that you expect you'll perfectly reasonably not reconsider the matter between midnight and noon. Imagine that you're due to give an important lecture tomorrow afternoon, and you reasonably expect that redeliberating whether to drink will distract you from crucial preparations and will thereby incur a greater cost to you than any incurred by drinking the toxin. Imagine that a serious objection to your argument has just occurred to you, and you're determined to focus all your mental powers on it just as soon as you finish this deliberation whether to drink. (We may stipulate that you will not feel the effects of the toxin till later that night.) If this scenario seems psychologically unrealistic to less anxious readers, imagine along parallel lines a second eccentric billionaire offering you a second million dollars just if you retain your intention to drink—not redeliberating—till the time comes to act. (As we'll see presently, the parallel is revealing.)

If we imagine the case like this, is it intelligible to imagine you forming the intention to drink? Here's a line of reasoning that seems to show it is intelligible. Granted, you expect that if you reconsidered the matter in the morning, you would be irrational if you did not change your mind. Still, you also expect that you'll reasonably not reconsider the matter. The reasonability of not reconsidering does not rest on an

21. Kavka created the puzzle as an objection to David Gauthier's doctrine of constrained maximization, and Gauthier did rise to the challenge. (For Gauthier's reply, see "Assure and Threaten," *Ethics* 104 [1994]: 690–721, "Commitment and Choice: An Essay on the Rationality of Plans," in *Ethics, Rationality, and Economic Behaviour*, ed. F. Farina, F. Hahn, and S. Vannucci [Oxford: Oxford University Press, 1996], 217–43, "Rethinking the Toxin Puzzle," in *Rational Commitment and Social Justice*, ed. Jules L. Coleman and Christopher W. Morris [Cambridge: Cambridge University Press, 1998], 47–58, and "Intention and Deliberation," in *Modeling Rationality, Morality, and Evolution*, ed. Peter Danielson [Oxford: Oxford University Press, 1998], 40–53.) Nonetheless, one might safely call Kavka's conclusion the received view of the case in the literature. For a fuller defense of it, see Bratman's "Toxin, Temptation, and the Stability of Intention."

external mechanism: the difference does not hang on the costs of drinking but of reconsidering. Maybe our billionaire will likewise refuse to reward intentions formed through expectation of heavy redeliberation costs. The point is that that would be a further stipulation. Without it, it seems you could form the intention to drink.

Let's follow this reasoning a bit further before diagnosing its mistake. You could not form the intention to drink in the first scenario, one might think, but you could in the second. What would make the difference? Bratman explains why you cannot form the intention to drink in the first scenario by appeal to a dual-claused No-Regret condition.[22] When you form an intention, Bratman argues, you must expect that the intention will be stable, where stability requires the satisfaction of two conditions:

(*a*) If you follow through on the intention, you won't regret it.
(*b*) If you don't follow through on the intention, you will regret it.

In the first scenario, you expect neither condition to be satisfied. You can see that you will regret it if you drink the toxin and that you won't regret it if you don't drink the toxin.[23] The No-Regret condition distinguishes the original toxin case from "temptation cases," in which the temptation to reconsider merely marks a transient preference shift and ought rationally to be resisted.

But in the revised scenario, you'll regret it if you don't follow through on your intention to drink. It isn't, of course, that you'll regret not drinking the toxin; rather, you'll regret what led you not to drink the toxin, namely, reconsidering. (Again, there are heavy redeliberation costs in this case.) If you fail to drink because you've merely forgotten your intention—without reconsidering—you'll regret *that* because, not having reconsidered, you weren't in a position to appreciate the reasons against drinking.[24] And, of course, you won't regret it if you follow through on your intention, since follow-through without deliberation is necessary for you to focus your attention on the all-important lecture.[25]

22. Bratman, "Toxin, Temptation, and the Stability of Intention," sec. 8.

23. More specifically, you expect that you'll abandon any intention to drink once you get the payoff for forming the intention to drink. But should you fail to abandon the intention, you'll nonetheless regret drinking, since you'll see that you ought to have abandoned it.

24. We can imagine that it matters to you enormously that you follow through on your intentions, except where you've redeliberated carefully. Maybe that's a touch too rationalist for realistic psychology, but in the decision-theoretic context at hand, that sort of rationalism is the default stance.

25. The 'of course' marks the fact that we're merely stipulating that the case has this feature. One might object that there could be no such case, because redeliberation costs could never outweigh the costs of making yourself sick. There is no reason to believe this general claim, however. Perhaps it's hard to imagine how staying focused on your lecture

So Bratman's No-Regret condition is satisfied: in this scenario, if you drink you won't regret it, and if you don't drink you will regret it. That is, this is what you expect in the scenario. One might conclude that these expectations make it possible to form the intention to drink.

That conclusion is mistaken. You can no more form the intention to drink in the revised scenario than in Kavka's original puzzle.[26] Though Bratman's No-Regret condition is satisfied, there simply isn't the right connection between the basis of your intention and your expectation of a future free from regret.

The easiest way to see this is to note that the revised scenario is, in effect, two toxin puzzles: Kavka's puzzle targeting the formation of your intention to drink, plus a new puzzle targeting your nonredeliberative retention of that intention as the time arrives to act on it. As we saw, the heavy redeliberation costs in the second puzzle can be rendered vivid by imagining a second eccentric billionaire eager to reward them. We might try to imagine you forming the intention at midnight in order to get the first payoff and then retaining it at noon (with that first million in the bank) in order to get this second payoff. In each instance, you intend to drink the toxin with an eye toward autonomous benefits of so intending, as opposed to benefits of the intended act. The drinking itself figures in your calculations only insofar as you're willing to undergo the costs of making yourself sick to get each payoff. In each instance, you're principally aiming at a benefit of being in the state of intending to drink, not at any benefit of actually drinking. If for this reason you cannot thus form the intention to drink the toxin—as many philosophers, including Bratman, argue—then you equally cannot thus retain it. But you satisfy Bratman's No-Regret condition in the revised scenario. The No-Regret condition cannot, therefore, explain how such toxin cases differ from temptation cases.

How might we diagnose the problem? If you cannot form the intention to drink in the revised scenario, it must be because you cannot thereby count as exercising executive authority—the authority distinctive of intention—over your conduct. Let's back up and ask what it is about such authority that could make this so.

What prevents you from forming the intention in Kavka's original scenario is that you think it unlikely that your having formed the intention will have rational bearing on your conduct when the time comes

---

preparations could be more important to you than the prospect of illness, but we could add details to ensure this feature.

26. For an opposing view of the revised toxin scenario, see Hinchman, "Trust and Diachronic Agency," 43. I've changed my mind about the case.

to act.[27] You think it indeed likely that when the time comes to act on the intention to drink, should you form it, your having formed it will have no rational bearing on what you should do. You foresee that, if rational, you'll have redeliberated and changed your mind by then. And here we make the first step toward a diagnosis. It is, quite generally, impossible to form an intention to φ at t while believing it unlikely that your having formed the intention will have any rational bearing on your conduct at t.

What rational bearing on your conduct must you expect your intention to have? You needn't think it unlikely that you'll have occasion to redeliberate the matter before acting. Circumstances change, after all, and in forming an intention you needn't deem yourself omniscient. You can form an intention to φ at t even while thinking it likely that unforeseen circumstances will arise between now and t, thereby making it rational to redeliberate. (Of course, if you foresee these circumstances, you have to take them into account.) You must expect, however, that your intention to φ at t will have a rational bearing on your conduct at t if circumstances emerge as you now expect them to emerge. What, again, is this rational bearing? You expect that you'll reasonably follow through on the intention without redeliberating.[28] The executive au-

27. If this isn't obvious, let me elaborate. (For greater elaboration, see ibid.) What's distinctive of Kavka's original scenario is that you believe that forming the intention to drink will create an intrapersonal predicament in which your acting self won't reasonably let itself be influenced by your—that is, by its own—executive capacity. The billionaire thus puts you in a predicament that you can see will engender a pathological self-relation. It isn't merely that you don't believe that you'll in fact drink. It may be that forming an intention entails believing that you'll succeed in doing what you intend to do (or at least not believing that you won't succeed). (I don't think that's quite right: see n. 29 below.) But the problem here isn't that you don't expect to succeed in drinking the toxin or that you do expect not to succeed. (We needn't imagine that you expect to gag.) The problem is rather that you don't believe that the intention to drink will have rational bearing on your conduct when the time comes to act. You don't expect that rationality will require you even to try to drink the toxin. Note that this is not like the difficulty a soldier emerging from his trench would face in forming or retaining the intention to carry out his mission behind enemy lines. His doubt that he'll even get a chance to try to carry out his mission may prevent him from categorically intending to carry it out. But he can form the conditional intention to carry it out if given the chance. Yet we needn't imagine you fearing that some external obstacle will prevent you from even trying to drink.

28. Note one thing I'm not saying here: that you must credit yourself with an accurate grasp on what will make it reasonable to follow through on your intention. I'm not saying you must expect that how you are reasonable in following through on the intention—whether by your lights at follow-through or simply in fact—will match how you now conceive yourself to be reasonable in forming it.

thority of intention rests on the presumption that such a self-relation—such self-influence—is reasonable.[29]

The question, then, is whether that's how you project a future in the revised scenario. Do you view yourself, as viewed from the perspective of your future acting self, as having legitimate executive authority over its conduct? It seems clear that if you cannot claim legitimate executive authority in Kavka's scenario, then you equally cannot in the revised scenario. The revisions do not, after all, change the fact that if you were to redeliberate when the time comes to drink, you would push the toxin away with relief that there's no reason whatsoever to drink it. All that has changed is that you expect that you reasonably won't redeliberate when the time comes to drink. That expectation makes it reasonable for you to form an intention now about what to do then, but it does not make it reasonable to form the intention specifically to drink.

The problem is the same as in Kavka's scenario. You cannot form an intention to φ at t unless you expect that you'll be reasonable in doing specifically this, so described: following through on that intention at t. In the revised scenario, you expect that you would be reasonable in not redeliberating when the time came to act on the intention in question. But you don't expect that you would be reasonable in following through on the intention. Of course, the two act descriptions would corefer—if you could form the intention and act on it. But you cannot form the intention, precisely because you don't expect that you would be reasonable in satisfying the latter description.

IV

Let's diagnose how Bratman's No-Regret condition permits this counterexample, so that we may reformulate the condition to exclude it. The problem in the revised toxin scenario appears to be this: what explains your lack of regret has nothing to do with either the content or the deliberative basis of the intention on which you're following through. In the revised scenario, you would intend to drink the toxin because you would believe that you thereby—that is, by forming or by retaining the intention—ensure that you receive a payoff. But what explains your lack of regret at following through on this intention has nothing to do with that rationale. Let's begin by considering how this

29. We can perhaps resolve the long-standing debate over whether you can form the intention to φ when you believe you will not φ by focusing on this 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 Hinchman, "Trust and Diachronic Agency," sec. 5, an account elaborated in Sec. IV below.

counterexample to Bratman's condition on the stability of intention engages our recent remarks on the executive authority of intention.[30]

How could a backward-looking attitude such as regret play a role in regulating the capacity for practical commitment? From a backward-looking perspective, practical commitment takes the form of a self-trust relation. Since you do not count as simply following through on an intention insofar as you are redeliberating, to follow through on an intention you must act on the basis of trust in the self of yours that formed it. This is not at all to say that you act 'blindly'—merely that you exercise the relevant 'sightedness' (into your reasons and rational requirements) in deliberation and (perhaps also, albeit differently) in intention formation, not when you merely follow through on the intention. Whatever we say about the executive authority of intention, therefore, we must also say about your presumption that your intending self is worthy of your trust in this way. It thus makes sense to interdefine the two attitudes informing your forward-looking perspective when you intend: to hypothesize that the executive authority of intention just is this presumption of trustworthiness. Adapting Bratman's insight, we may add that you cannot exercise this authority if you do not expect the exercise to be stable in the way captured by a No-Regret condition. It would follow that the freedom from regret that figures as a criterion of stability looks back not primarily at your action but at the self-trust relation whereby you performed it.

We thus arrive at my proposal. I propose that we reformulate Bratman's No-Regret condition in terms of a species of what I'll call *trust regret*: regret specifically that you instituted the self-relation—a self-trust relation—whereby you followed through on the intention. We can dis-

30. It is not clear that Bratman himself really explains the relation between the stability and the executive authority of intention. Bratman's original account of his No-Regret condition rested with noting a pragmatic need for diachronic stability in planning agency. (See Bratman, "Valuing and the Will," 56. This briefly restates how he justified the No-Regret condition in "Toxin, Temptation, and the Stability of Intention," where he first proposed the condition.) Clearly there is such a need. But the question is why the stability should take precisely that form. More recently, Bratman has elaborated his account of the No-Regret condition in terms of his broader account of *agential authority* (Bratman, "Temptation Revisited"; for the notion of agential authority, see Sec. V below). This strategy appears to run the explanation in the wrong direction. An account of agential authority would explain a dimension of your (real) authority over your action, when you have it: it would explain what makes it the case that you're in charge. But an account of the stability of intention should explain the nature of your claim of authority—how you can presume to guide yourself in this way, counting on your intention as an antidote to weakness—without presupposing that the authority is genuine. At the very least, it would be simpler to explain the nature of stability—the claim or presumption of authority, of treating yourself as this species of authority—without assuming that we understand what grounds this authority or makes it genuine. That is the approach that we're pursuing here (and that I pursue more fully in Hinchman, "Narrative and the Stability of Intention").

tinguish toxin cases from temptation cases by asking whether in forming the intention you can—without ignorance, confusion, or an external mechanism—posit a future free from regret specifically targeting the self-relation that you thereby expect to institute.

You cannot form the intention to drink in the revised scenario because you would expect that you'd regret not the mere act of drinking extensionally individuated (which, by hypothesis, you would not in context regret), but the diachronic self-relation that you would thereby institute between your intending and acting selves. You would regret not what you would have done—namely, drink the toxin (in the context of your need not to redeliberate)—but the executive self-relation you would thereby institute. After all, you would expect that if you did redeliberate at noon tomorrow, you would see that you'd have no reason whatsoever to drink the toxin. This expectation does not alone ensure that you cannot form the intention. In a temptation case, you may also expect that redeliberation would lead you to change your mind. The expectation is important because of what it reveals about your self-relations: when you expect that you would change your mind upon redeliberation, you expect that you would not appear to be a trustworthy executive to your acting self. The expectation of such an appearance is not a problem in a temptation case, since you also expect that you will not regret the self-relation whereby you thus influenced yourself. But the revised toxin case is different: since the appearance of untrustworthiness is not caused by a preference shift, you cannot expect it to go away as your preferences shift back. Unlike in a temptation case, you would expect that the perspective from which you would act would persist as the perspective from which you would look back with unsettling regret. This is the perspective from which you would feel the pinch of the recognition that you got nothing from intentionally making yourself sick.

One may try to object that you don't get nothing for performing the act, since you get the payoff for not redeliberating. The problem, however, is that 'not redeliberating'—the description you must satisfy to get the payoff—is not a description of the action you intend to perform. We return to the distinction emphasized at the end of the previous section. While it isn't wrong to say that you won't regret drinking because of this payoff, that would be tendentiously imprecise. You likewise would not regret doing nearly anything you deemed necessary to get the payoff at noon that day—as long as you preferred getting the payoff to doing that thing (with its consequences). In the sort of case we're considering, there are comparatively few things that we couldn't imagine to meet this description. So what you would regret or not has nothing directly to do with what you've specifically chosen to do at noon that day. It

therefore has nothing directly to do with how you've specifically committed yourself to act.

The case gives us our first glimpse of one key difference between the intelligibility constraints on the two species of practical commitment. If the intelligibility-constraining role of your twice-future self were exactly like that of a co-conspirator, you could form the intention to drink in the revised toxin scenario simply by 'recruiting' that self to your instrumental manipulation of your once-future acting self. "Don't regret the manipulation," you could (as it were) say to your twice-future self: "Manipulating that other self to drink is in your self-interest." The absurdity of that attitude toward your future selves reveals that the role of your twice-future self is, in a key respect, unlike the role of a co-conspirator in constraining intelligibility. Unlike in the interpersonal case, all three of the perspectives defining the intrapersonal case *are yours*. One key difference between the intelligibility constraints lies in the fact that the executive authority of intention serves to organize these perspectives into that of a responsible—because it is appropriately self-responsive—single agent, whereas the deliberative authority of practical judgment serves by contrast to integrate that judgment into a rational order defined by its accessibility to multiple agents. I'll explain this difference more fully in Section V. For now, note merely that this fact about the authority of intention affects how and when you can expect you'll regret following through on an intention. Regret targeting the self-relation whereby you followed through—what I'm calling trust regret—is partly regret at having failed in the executive task of self-constitution. There is no such task in the interpersonal case.

Elaborating the self-relation from the forward-looking side of intention, I propose that the relation whereby you commit yourself amounts to an invitation of self-trust: you expect and desire your acting self to follow through on the commitment simply because of that desire. You've formed or endorsed a desire to φ at t in a way that counts as closing deliberation—for now at least—and you expect to φ at t simply because you're thus committed. When you follow through on the commitment, you'll do so in the spirit of self-trust. That doesn't mean, again, that you expect you will follow through. It means merely that if you don't, you expect it will be because you—by your current lights, irrationally—refused the invitation to trust.

The proposal needs elaboration on several fronts. Here are a few preliminary remarks.[31] Trust, we can say generally, is a species of willed

31. For more elaboration, see Edward S. Hinchman, "Receptivity and the Will," *Noûs* 43 (2009): 395–427, and "Regret and Responsible Agency" (unpublished manuscript, University of Wisconsin–Milwaukee, 2010). Hinchman, "Trust and Diachronic Agency," contains a defense of the basic idea that intending is inviting self-trust.

dependence, where the dependence is under appropriate guidance of a counterfactual sensitivity to evidence of untrustworthiness in the trusted. 'Appropriate guidance' means that you would not trust if you had evidence that the trusted is not worthy of your trust. 'Evidence' does not mean conclusive evidence. Even inconclusive evidence of untrustworthiness can undermine trust. Beyond that, we may leave open what precisely counts as trustworthiness, untrustworthiness, or evidence thereof. What is important for our purposes is that insofar as you actively weigh positive evidence of trustworthiness, you act not from trust but from your own deliberative judgment—and the question remains whether to trust that judgment. What makes it possible to follow through on an intention nondeliberatively is that trust can be reasonable with no active assessing of your intending self, as long as the latter is trustworthy. You need merely be (and, counterfactually, have been) disposed to notice and respond appropriately to evidence of that self's untrustworthiness, should there be (or have been) any. Expecting that you will have trusted reasonably, by your future lights—that is, expecting not to regret the trust relation—would simply manifest an expectation that you will exercise the counterfactual sensitivity that is part of the definition of trust.

We are now speaking entirely from the perspective of the intending self. The claim is that when you intend, you posit a regret-free future with respect to an aspect of your execution of the intention. You expect that you will not regret entering into the self-trust relation that you establish when you follow through on the intention. You acknowledge that you may well regret the action you thereby perform—since, like any action, it may have unforeseen consequences. The proposal indeed builds on this distinction: you can deem the action regrettable (say, because of unforeseen harms) but not the trust relation, or you can deem the trust relation regrettable but not the action (say, because of unforeseen benefits). Again, trust regret is not mere regret at having performed the action.

The reformulated No-Regret condition suggests an alternative approach to questions of practical intelligibility. When you regret not your act as such but having trusted your intending self, you manifest a failure to remain intelligible to yourself. "What was I thinking?" you may exclaim. "How could I have done that?" You'll tend to feel rather creepy about yourself—not exactly ashamed, but perplexed and disoriented. Your predicament is like that of parties to a cause that has gone sour. It isn't merely that the cause now looks wrongheaded. It's that parties to it cannot understand how that group, so constituted, could ever have spoken for them. When gripped by trust regret, you cannot understand how the trust relation to which you were party could speak for you. It does, of course, speak for you, since it is the form in which you unified

yourself as an agent. Over and above the act itself, it is this fact that you cannot understand when gripped by trust regret. We might think of the creepy feeling that goes with this failure to understand your agency as the sanction of trust regret.[32]

What I've said so far suggests that in intending you expect you'll act in a way that will avoid your own later de facto regret. But that can't be right, since we sometimes expect our future selves to regret unfairly. Trust regret appears to function as an intrapersonal reactive attitude, both because it comes with a sanction and because—as we'll see more fully in Section VI—its appropriateness conditions contain a norm of fairness. The fairness norm is internal to the regret in the way that it is internal to all reactive attitudes: if it would be unfair to apply the sanction, then it would be wrong to adopt the attitude.[33] And if you expect that it would be wrong to regret your having followed through on an intention, then you can form the intention even if you expect to regret following through on it. Just as regret can reveal that an intending self or other should not have been trusted, so trustworthiness in an intending self or other can reveal that trust should not be regretted.[34]

32. Since the sanction amounts to a failure to make the right sort of sense of that self-relation, we can agree with Velleman and say (if we choose to talk thus) that the 'constitutive aim' of practical commitment is indeed a species of self-intelligibility. I'm merely rejecting Velleman's interpretation of this thesis on one key issue: the precise respect in which self-intelligibility matters, I'm arguing, lies in the claim of executive authority inherent in intention. On this view, such self-intelligibility determines the core not, as Velleman argues, of self-knowledge but of volition. For Velleman, your intention to φ is merely an expectation that you will φ, an expectation that guides you via your drive to make sense of your behavior, which entails a desire to make such expectations true. That account is wrong, I would argue, for a reason sketched in n. 37 below. Since I cannot give a full argument here, I'll present my account of the basis of this intelligibility constraint as a mere alternative to Velleman's account. What I'm positing, again, is not a drive to make sense of your behavior but a need to make sense of how you are authorizing your behavior—to the extent that you are. You authorize your behavior to the extent that it is governed by the claim of executive authority inherent in your choice or intention. To make this claim of authority, it has to be intelligible to you why you should follow through on the intention, a form of intelligibility that requires you to posit a basis for the authority in your own status as trustworthy. You cannot presume such trustworthiness if you expect that your twice-future self will fairly regret your having followed through on the intention.

33. I here follow Gary Watson's analysis of responsibility in "Two Faces of Responsibility," in his *Agency and Answerability* (Oxford: Oxford University Press, 2003), 260–88. One might worry that this appeal to fairness raises a question of justification in a way that undermines the distinction between justification and intelligibility that we're about to emphasize. I reply to this worry in Sec. VI.

34. Still, can't you form an intention and then act on it, all the while sighing under your breath, "I'm going to regret this"—but nonetheless not regard the anticipated regret as unfair? We need not deny the possibility of such akrasia when intention and act are more or less synchronic. Akrasia sometimes takes that form, as when we plump for the drink, dessert, or more complicated temptation impulsively or wantonly. But that's different

V

I've suggested that developing this approach to commitment rests on distinguishing two species of practical authority. In interpersonal commitment, what is at issue is whether you can make sufficient sense of what you're doing for it to count as the deliverance of your deliberative faculty. In intrapersonal commitment, by contrast, the question is whether you can make sufficient sense of what you're doing for it to count as manifesting a claim of executive authority over your conduct. These are two different but interrelated requirements. Since the basis of each intelligibility constraint lies in its function, the constraints differ insofar as they are constraints on different kinds of things, and the things in question appear to be distinct species of authority. The first constraint qualifies the deliberative authority of judgment, while the second qualifies the executive authority of choice or intention.

We can grasp the importance of this distinction in authority by observing that practical rationality spans two species of akratic gap. Akrasia can break the link between your judgment (all things considered) that you should φ and your choice or intention to φ: having deliberated, you conclude that you ought to quit smoking before your next checkup, but you never actually choose or form an intention to quit. Call this species of akrasia 'incontinence'. A different species of akrasia breaks the link between intending to φ and actually φing: you intend to quit smoking after you finish the pack of cigarettes in your pocket, but then you break down at the gas station and buy another pack anyway. Call this species 'weakness'.[35] Incontinence breaks the link between practical judgment and practical commitment, while weakness breaks the link between practical commitment and action. A constraint on deliberative authority is therefore a constraint on the form of practical authority

from presuming to exert executive authority over a future acting self even while expecting that your twice-future self will regret this influence. It is doubtful that you can exert such authority—at least, when the expected gap between intention and execution leaves room for reconsideration. This is a denial not of akratic action but of akratic intending. For a fuller defense of this denial, see Hinchman, "Regret and Responsible Agency," which also explains how this view is compatible with other forms of perverse agency, since it doesn't entail the thesis that practical commitment transpires under the guise of the good.

35. For a clear statement of how weakness differs from merely acting against your better judgment, see Richard Holton, "Intention and Weakness of Will," *Journal of Philosophy* 96 (1999): 241–62. For a view of akrasia specifically as (what we're calling) incontinence, see Christopher Peacocke, "Intention and *Akrasia*," in *Essays on Davidson: Actions and Events*, ed. Bruce Vermazen and Merrill B. Hintikka (Oxford: Oxford University Press, 1985), 51–73. The present conception of practical commitment as mediating judgment and action is similar to a conception recently articulated by T. M. Scanlon ("Structural Irrationality," in *Common Minds: Themes from the Philosophy of Philip Pettit*, ed. G. Brennan, R. Goodin, F. Jackson, and M. Smith [Oxford: Oxford University Press, 2007], 84–103), but there are also some crucial differences (see Edward Hinchman, "Reasons and Rational Coherence" [unpublished manuscript, University of Wisconsin–Milwaukee, 2010]).

that would rationalize—that is, bring into accord with rational norms—your transition from judging that you ought to φ to choosing or intending to φ. And a constraint on executive authority is a constraint on the form of practical authority that would rationalize your transition from choosing or intending to φ to φing (or to judging that you ought to ψ, where you view ψing as a necessary means to φing).

Why, then, is there an intelligibility constraint on each species of authority? Begin with deliberative authority. The task of practical judgment is to close deliberation—not merely to end deliberation but to reach a conclusion in accord with applicable norms. One of these norms has a forensic dimension: when you bring deliberation to a conclusion, judging that you ought to φ, there must be something you could say to someone who challenged your rational entitlement to this conclusion. Even when you can get away with a reply along the lines of "I just felt like drawing that conclusion," this will fly only because of special features of your deliberative context.[36] Both the reasons and the aspects of context that provide a medium for adjudicating these exchanges must be interpersonally accessible. And it is this accessibility requirement that creates the intelligibility constraint. The basis of the constraint on deliberative authority is therefore forensic, deriving from the requirement that you be able to make sense not of your action simpliciter, but of your having drawn that deliberative conclusion. Why not, in sum, keep deliberating? If you don't expect you could make your rationale for concluding intelligible to a potential co-conspirator, that would undermine the presumption of deliberative trustworthiness required to rationalize your transition from judgment to commitment.[37]

36. You can often close deliberation about whether to spend another hour in bed in this way—that is, because this special context gives 'how you feel' an unusual normative standing. Most decisions would be unintelligible if formed on that basis—with nothing more you could say even to a potential co-conspirator.

37. This leaves it open that you may act on the basis of someone else's deliberative authority—that is, on their judgment rather than your own. The requirement applies not to the action as such but to the mental act or attitude whereby you authorize it. If you act by simply trusting another's judgment, the intelligibility constraint applies to that person's judgment. 'Trusting another's judgment' does not mean trusting her advice, which can influence you only through your own judgment, but means accepting her invitation to share an intention, which can bypass your judgment. To say that the intelligibility constraint applies to this other person's judgment does not entail that how the context entitles her to that judgment must be intelligible to you. The intelligibility constraint applies both to the other person's act of judging and to her act of inviting you to trust her judgment—that is, to her act of inviting both her own trust and (through her invitation that you share the intention she would thereby form) your trust as you act together. The forensic basis for the latter intelligibility constraint lies in your entitlement to defer specific justificatory challenges to her. (For argument, see Edward Hinchman, "How to Settle on a Shared Intention" [unpublished manuscript, University of Wisconsin–Milwaukee, 2010]. See also n. 32 above.)

The basis of the constraint on executive authority, by contrast, is not forensic but metaphysical. The task of choice or intention is not to close deliberation but to affirm the judgment whereby deliberation is closed. The possibility of incontinence ensures that practical commitment need not accompany deliberative closure. Even if they occur simultaneously, judging that you ought to φ is logically distinct from choosing to φ.[38] The task of intrapersonal practical commitment is to organize deliberation, judgment, commitment, and action into an intelligible whole such that we can identify in that whole the agent to whom the action is attributable. This is not, in general, a forensic enterprise because such attributability is not the same as accountability: as Susan Wolf and Gary Watson have argued, it is possible for an action to be attributable to you without your being accountable for it.[39] It is instead the metaphysical enterprise that Bratman calls the constitution of "agential" authority: "For the agent to direct thinking and acting is for relevant attitudes that guide and control that thinking and action to have authority to speak for the agent."[40]

It seems better to call the authority "executive" because that makes clearer the contrast with deliberative authority by locating the rational force of this authority in the second transition (from commitment to follow-through) rather than in the first (from deliberation, via judg-

38. We'll continue to focus on the diachronic case involving intention and follow-through, but everything we say about the presumption of executive authority in that case is also true of the synchronic case in which your choice to φ and your φing occur simultaneously (or close enough for it not to make sense to distinguish commitment from follow-through). Even synchronically, we can distinguish judgment from choice and choice from action, though spelling out the precise nature of the intrapersonal relations is complicated. See Hinchman, "Receptivity and the Will," sec. 7, for analyses of judging that you ought to φ and choosing to φ that make the distinction clear even in a synchronic case.

39. Watson, "Two Faces of Responsibility." Watson is refining a thesis from Susan Wolf's *Freedom within Reason* (Oxford: Oxford University Press, 1990), chap. 2.

40. Michael Bratman, "Introduction," in *Structures of Agency*, 4. Gary Watson has claimed that Bratman posits a "metaphysical imperative" to maintain one's identity (see Gary Watson, "Hierarchy and Agential Authority," in *Free Will: Critical Concepts in Philosophy*, ed. John Fischer [New York: Routledge, 2005], 4:94–95), 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)" (Bratman, "Three Theories of Self-Governance," in *Structures of Agency*, 246; Bratman says in a footnote that he is replying to Watson's claim). Does deliberative authority then map onto what Bratman calls "subjective normative authority" (Bratman, "Introduction," in *Structures of Agency*, 4–5)? No. Though there are common elements, these are not two names for the same notion. Since Bratman doesn't make any theoretical use of the distinction between incontinence and weakness and therefore tends to regard deliberation as issuing directly in practical commitments (not distinguishing between judgment and choice), there is no room for our distinction between deliberative and executive authority within his framework.

ment, to commitment). The more fundamental contrast lies in the fact that the intelligibility constraint on executive authority derives from the exigencies of self-constitution underlying the attributability of an action and not in any straightforwardly forensic requirement. It is crucial to intention that you don't have to convince your future self to act; all you have to do is remain convinced. Any forensic challenge that you may face retrospectively is a challenge to your deliberative authority—"Why did you draw *that* conclusion?"—rather than to your executive authority. Your executive authority presupposes that you have deliberative authority on the question of what you should do and adds to that authority the act or attitude of commitment to the deliverance of this authority that we can most easily (if incompletely) understand as what an incontinent agent lacks. The authority behind this commitment thus includes the responsibility to others that figures in deliberative authority but adds a responsibility to self that is not fundamentally forensic. What you "owe" yourself is not a justification of what you're doing but the self-consistency and resolution that would suffice for you to count as doing it—both causally (you can't go weak) and constitutively (the action must be attributable to you).

When you've committed yourself to act, the executive self-relation unfolds only insofar as you remain intelligible to yourself, where you manifest this self-intelligibility only insofar as you expect to avoid your future trust regret—regret, specifically, that you instituted this very self-relation. Trust regret is a mark of your failure to remain intelligible to yourself. Your aim to avoid it is precisely, then, the aim to understand what you're up to when you commit yourself. The question of intelligibility as you form and retain the intention is not whether you actually have made relevant sense of yourself but whether you expect to be able to make relevant retrospective sense of yourself—just as the question of intelligibility informing deliberative authority is not whether you have made sense of yourself in drawing the conclusion but whether you expect you can or could make sense of yourself in justifying your conclusion to a possible co-conspirator. In each case, the question of self-intelligibility anticipates a reaction from a third perspective: not your present self, not the one you aim to influence, but a further party standing after or athwart your conduct. Your expectation that what you're doing will make sense to this third party just is your ongoing sense that what you're doing is relevantly intelligible.

The intelligibility to others that functions as a constraint on forming a judgment and the intelligibility to yourself that functions as a constraint on making a choice or forming an intention are thus functionally distinct but formally parallel. We may put the parallel like this. In each dimension, the 'constitutive aim' of agency—should we go in for that sort of talk—is to make yourself intelligible from a perspective that

stands to one side of the thrust of your agency. Committing yourself puts you into relation with the future self that will follow through on the commitment, and this relation must be intelligible to a twice-future self in position to regret it. Following through on the commitment—that is, performing the action in question, which was the subject of your deliberation—in turn puts you into relation with a world that may contain other agents to whom your action must, by its own lights, be and remain inscrutable. So the second test of intelligibility, as a constraint on deliberation, lies in potential co-conspirators—that is, in those whom the action does not target in this way. Of course, if there is no deception, then there is no need to appeal to a potential co-conspirator: you can test for intelligibility by appeal to any external perspective. The test case comes when there is deception. When there is deception, you can test for intelligibility by appeal only to the possible perspective of someone who shares your project. As we've seen, there is no reason to worry that this restriction compromises the intelligibility of your action.

VI

We've agreed not only with Velleman that there is an intelligibility constraint on the presumption of deliberative authority but also with Korsgaard that there is an intelligibility constraint on the presumption of executive authority. Why not prefer a Kantian explanation of our reformulated No-Regret condition? In Section II, we saw that Korsgaard's explicit account fails to answer key questions. Let's now see why a Kantian account is bound to give unsatisfying answers to those questions. Seeing this will enable us to appreciate why each intelligibility constraint should function by requiring not that you make sense of what you're doing here and now but that you expect that sense can or will be made of what you're doing from another perspective.

Note first that our non-Kantian account explains the Kantian distinction between being moved by a mere incentive or inclination and self-governance. Putting our conclusion in Kantian terms, we can say that what distinctively moves you when you govern yourself is your aim to avoid your own future trust regret. This is a different way of saying what we may agree with Korsgaard is true: that you constitute yourself as an agent by unifying your inclinations into responsible—because they are self-responsive—wholes. The point is that you can do that without identifying with a principle of choice. No doubt you often commit yourself to a course of action because you identify with a principle that favors it, then follow through on the commitment because your allegiance to the principle has not wavered. Principles do often mediate the self-relations constitutive of agency. The point is merely that, *pace* Korsgaard,

identification with a principle of choice—a principle you expect to share with your acting self—is not essential to those self-relations.[41]

We are pressing a twofold objection to Korsgaard's Kantian account. First, it does not explain how principles can mediate these self-relations merely to note that the self-relations often involve your identification with a principle of choice. Second, you can achieve the unification whereby you commit yourself without identifying with a principle of choice. You can do it simply by being responsive to the normative status of a key subset of the self-regarding attitudes that you expect yourself to have (in Bratman's useful expression) at plan's end.[42] Those who act with a particularistic will can nonetheless manifest practical commitment—all it takes is for them to intend and then act with a concern to avoid their own future regret at the self-relation they thereby realize. Even an unprincipled agent may recoil at the prospect that his self-relation will later creep him out.[43] This shows that unprincipled agents can hold themselves to a standard of objective intrapersonal practical intelligibility, confirming again that objective intelligibility is not the same thing as universal intelligibility.

It is important to emphasize that the argument we're pursuing does not generate a nonnormative approach to practical commitment. We noted in Section IV that the No-Regret condition applies to expectations of *fair* trust regret, since you might expect that your plan's-end self will have come to regret where it should not—that is, where you do not deserve such treatment. Does that reveal an explanatory need for principles after all? Since we cannot pursue a broader critique of Kantian ethics here, assume that the appeal to fairness should be given a Kantian

41. Note well one limit of this argument: it does not question Kantian claims about what would follow if practical commitment did necessarily involve identification with a principle of choice. (Perhaps in that case lying would always be impermissible.) It questions specifically this necessity claim.

42. Bratman, "Toxin, Temptation, and the Stability of Intention," 86. For present purposes, we may follow Bratman in viewing "plan's end" as the point beyond which you will not change your attitude toward the action. In some cases, you'll expect this to be right after the action is performed; in others you may expect it to be years later.

43. What if someone fails to recoil? Such a person seems rather like Harry Frankfurt's 'wanton' (see his "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68 [1971]: 5–20; and also [for a pertinent comparison] Korsgaard's *Sources of Normativity*, 99 n. 8). But we needn't pursue this idea at this point in the dialectic, since the present claim is merely that someone can fail to act from a principle of choice while nonetheless being concerned to avoid trust regret. Could there be someone who failed to care about trust regret and yet nonetheless managed to commit himself? I haven't defended the account against such counterexamples. My starting point here is Bratman's No-Regret condition. For a defense of the approach against such broader worries, see Hinchman, "Regret and Responsible Agency."

interpretation.[44] Assume, in other words, that the No-Regret condition applies by requiring that, as you intend, you expect that your plan's-end self will spare you the sanction of trust regret, in part by sharing principles with you concerning when such regret would be unfair.[45] That would yield a relation between your intention-forming self and your plan's-end self similar to the relation between intending and acting selves required by Korsgaard's account of practical commitment. There is nonetheless this crucial difference: the shared principle would not figure in what a Kantian would deem the maxim of your action. The approach would yield a Kantian account not of practical commitment but of a key aspect of the self-relation through which our non-Kantian account of practical commitment is realized.

There is, moreover, a reason to reject that account of the self-relations that does not derive from any broader worry about Kantian ethics. The problem is that a Kantian emphasis on principles will always make questions of intelligibility look like questions of justification. Why should it matter to you that a later self of yours share your principles of fairness except in a context of 'reasoning with' that self about whether something—for example, regretting a trust relation—would be fair? The issue between your intention-forming self and your plan's-end self is not one of justification. Indeed, a simple way to explain why all these self-relations pose questions of intelligibility rather than questions of justification is to note that self-justification is a species of reasoning with yourself and that the process of reasoning with yourself ended when you made up your mind to act. If you're still in the business of self-justification—even about whether it would be 'fair' to regret following through on this or that choice—then the question of what to do is still open for you, and you haven't really made a choice or formed an intention. If, having closed the matter and committed yourself to act, you nonetheless find yourself rehearsing the considerations in favor of acting, it cannot be for the benefit of your future self—unless, of course, you're merely trying not to forget them for some other purpose, such as the justification of your action to others. In sum, the question has to be one of self-intelligibility rather than self-justification because there is no longer any justifying to do in the intrapersonal dimension that unifies you as an agent.[46]

44. For a full—and firmly un-Kantian—account of such 'fair' intrapersonal dealing, see Hinchman, "Narrative and the Stability of Intention."

45. The assumed account would, of course, need to supply details about this species of fair dealing, but those details do not matter to the present point.

46. Of course, people often do second-guess themselves by 'reopening' deliberations after—sometimes long after—they have acted. But since the question of what to do is no longer open, this is not really deliberation. Such nondeliberative second-guessing falls

What if a Kantian conceded this point but insisted that her appeal to a principle of fairness shared between intending and twice-future selves functions to confer only intelligibility, not justification? Could such a Kantian motivate a prohibition on deception? It seems not. The prohibition on deception that we're considering is a prohibition on attempting to influence someone through a specific kind of deception: by insincerely representing yourself as committed to a course of action—as we framed it in the prologue, by interpersonally committing yourself where you are not intrapersonally committed. No prohibition on this sort of deception could be derived from an argument that generalized from the relation between intending and *twice*-future selves. The interpersonal analogue of such an intelligibility-testing principle of fairness would be a principle shared between agent and co-conspirator. And as we've seen, the intelligibility-testing role played by a co-conspirator is not specifically that of being influenced by the commitment in question.[47] This intelligibility-testing perspective need not even be actualized, and if it is, you equally need not be in the business of influencing your co-conspirator. It is for two reasons, then, that we cannot generate a prohibition on deception from this constraint on self-intelligibility: (i) the species of self-intelligibility at issue models this dimension of your self-relations on your relations not with a possible victim of your act but with possible co-conspirators, and (ii) the prohibition on deception at issue is not a prohibition on deceiving your co-conspirators.[48]

On the present approach, intrapersonal practical commitment has two distinctive features—or rather, a single distinctive feature with two sides. From one angle, the self-relation whereby you expect to make retrospective sense—by looking back from plan's end—is not the re-

---

among the pathologies that you need not expect you'll be spared in order to get deliberation closed.

47. In any case, if you actually give assurances to a co-conspirator in the course of deceiving someone else, this second dimension of possible deception will need intelligibility testing by appeal to the possible perspective of a further co-conspirator—in a conspiracy targeting not the original victim but your original co-conspirator—and so forth, for as many levels of conspiracy as you can actually undertake.

48. Could the nature of practical commitment itself generate a prohibition on deception that was not a prohibition on deceptive influence? A prohibition on deception would have to be a prohibition on something that we could understand as violating a norm that we could understand as worth enforcing. Could the nature of agency, perhaps backed by the concessively Kantian interpretation of the fairness requirement we've been considering, generate a requirement that you be 'honest with yourself' that might in turn ground a requirement that you be honest, say, with possible co-conspirators—apart, that is, from relations of possible influence? It is hard to know how to answer. Everyday experience renders each of us familiar with how our agency is compatible with the everyday self-deceptions—about how well one is loved, say, or about one's prospects for success in a difficult enterprise—necessary for going forward in life. And for related reasons we don't expect co-conspirators—coworkers, say, or spouses—to be transparent to one another.

lation whereby you expect your commitment to exert an influence. From the other angle, the self-relation whereby you expect your commitment to exert an influence is not a relation that you need expect will then make sense to you—since your sense-making capacity may briefly be impaired by akrasia. We've just considered an implication of the first angle; let's now consider an implication of the second. This will make it clear that the present approach does not in fact challenge Kantian accounts of the universalization implicit in judgment, whether theoretical or practical. The break with Kantian approaches lies not there—we may agree with Kantians that practical judgment is implicitly universal in these dimensions—but solely on this question of the nature of practical self-influence.[49]

I'm proposing a view of practical self-influence that specifically rejects the Kantian assumption that practical thinking is essentially self-conscious or reflexive. On the view we've developed here, when you judge that you ought to φ and then make a choice or form an intention to φ on that basis, the acting self who stands as object to your judgment, choice, or intention does not do so purely reflexively. Of course, the 'you' who judges, and so on, is the same person as the 'you' who acts and counts as the same agent in that respect, a respect that codifies issues of attributability and responsibility in addition to purely metaphysical questions of personhood. But the need to bridge the two akratic gaps discussed in Section V reveals that these three practical perspectives—judgment, choice or intention, and action—are crucially different. In this respect—in respect of how agency unfolds, whether logically or temporally, through functionally distinct perspectives— the 'you' who judges is not the same as the 'you' who chooses, nor is the 'you' who

49. My account does not, for example, challenge either of the universality claims that Stephen Engstrom characterizes as follows: "A judgment by a particular subject has subjective universal validity in that any subject that can grasp its concept would, if in the conditions of the judging subject, share the same judgment (the same predicate, so to speak) and so be in agreement with the judging subject. A judgment about a particular object has objective universal validity in that any object to which its concept can be applied would, if in the conditions of the object judged, share the same predicate and so be in agreement with the judgment's object" (*The Form of Practical Knowledge* [Cambridge, MA: Harvard University Press, 2009], 116). As we're about to see, my account challenges only Engstrom's further claim—the linchpin in deriving anything like a Categorical Imperative from the universality claims—that practical judgment is in its efficacy "essentially self-conscious" (ibid., 120), thereby ensuring that its subject is also its object. Engstrom observes (ibid.) that this further claim is not a claim that Kant ever quite defends: Kant argues merely that the object of a practical judgment must not be "given from elsewhere," not that it must be identical with the subject of the judgment. In Edward Hinchman, "Must Practical Self-Influence Be Reflexive?" (unpublished manuscript, University of Wisconsin–Milwaukee, 2010), I show how Kant could be right on the 'not given from elsewhere' formulation without that yielding the linchpin claim or any other route to a Categorical Imperative.

intends the same as the 'you' who acts on the intention. Simply put, the former in each instance judges or intends for the latter, not for itself. There are constraints on how it may do this, but these are constraints on deliberative or executive authority—that is, the authority that the former holds over the latter.[50]

We might view my principal dissent from Kantian approaches as expressing a dissatisfaction with Kantian blindness to the theoretical implications of akrasia. I am not arguing for an externalist approach to either species of akrasia, on which judgment would have no internal link to choice and choice no internal link with action. I agree with Kantians that, through the medium of choice, practical judgment rationally necessitates action.[51] But it does so, I'm arguing, specifically through the medium of the agent's self-relations. It is because akratic choice is rationally unintelligible from the perspective of judgment, and akratic action rationally unintelligible from the perspective of choice, that the intelligibility conditions on deliberative and executive authority must look to a third perspective. Simply put, agents do not constitute themselves as committed purely reflexively. It is fine to say that agents constitute themselves as intrapersonally committed. But they do so by bringing themselves into a three-place self-trust relation that we can model on an interpersonal relation.[52]

VII

How, then, is giving others 'your word' like and unlike having a 'word' to give in the first place? The interpersonal and the intrapersonal run in parallel insofar as each implicitly appeals to an onlooking perspective from which you posit the intelligibility of your act, a perspective that may diverge from the perspectives of those—whether your own acting self or another person—whom you aim to influence. You address your claim of deliberative authority not to the target—perhaps the victim—of your act but to a possible co-conspirator. And you address your claim of executive authority past your acting self—whom you may expect will be deafened by temptation—to the self marking a horizon at plan's end.

50. Readers not yet convinced by my characterization of the intelligibility constraints on deliberative and executive authority may note that we began with a case that most philosophers believe reveals an intelligibility constraint on executive authority. The Toxin Puzzle reveals a constraint on executive authority by revealing how the intending self must be in the business of doing justice—whatever this exactly involves—to the circumstances of the acting self (as it understands them), circumstances defined as different from its own.

51. For my full account of this internalism, see Hinchman, "Receptivity and the Will."

52. For an account of the interpersonal analogue of this self-relation—the relation whereby distinct agents can settle on a shared intention, with one initiating and the other responding—see Hinchman, "How to Settle on a Shared Intention."

The parallel breaks down insofar as the species of intelligibility needed for the deliberative authority of practical judgment differs from the species of authority needed for the executive authority of intention.

In one respect, it is the intrapersonal relation that enables you to misrepresent your interpersonal relations. You can deceive another about your commitments only because you can expect you'll retrospectively adopt a perspective toward yourself that is analogous to the perspective that others adopt toward you when they help you pull the deception off. You and your plan's-end self do not, of course, act together. In fact, there is no action that you count on your plan's-end self to perform. But you nonetheless could not lie about your commitments if you could not expect this future self to turn a deaf ear to the deception—thereby withholding the sanction of trust regret. In this respect, at least, practical commitment requires conspiring as much with yourself as with possible others.