11
TRUST AND WILL

Edward Hinchman

We might ask two questions about the relation between trust and the will. One question, about trust, is whether you can trust "at will." Say there is someone whom you would like to trust but whose worthiness of your trust is not supported by available evidence. Can you trust despite acknowledging that you lack evidence of the trustee's worthiness of your trust? Another question, about the will, is whether you can exercise your will at all without trusting at least yourself. In practical agency, you act by choosing or intending in accordance with your practical judgment. Self-trust may seem trivial in a split-second case, but when the case unfolds through time – you judge that you ought to φ, retain your intention to φ through an interval, and only at the end of that interval act on the intention – your self-trust spans a shift in perspectives that mimics a relation between different people. Here too we may ask whether you can trust "at will." Can you enter "at will" into the self-relation that shapes this diachronic exercise of your will? What if your earlier self does not appear to be worthy of your trust? If you cannot trust at will, does that entail – perhaps paradoxically – that you cannot exercise your will "at will"?[1]

In this chapter, I explore the role of the will in trust by exploring the role of trust in the will. You can trust at will, I argue, because the role of trust in the will assigns an important role to trusting at will. Trust plays its role in the will through a contrast between trust in your practical judgment and a self-alienated stance wherein you rely on your judgment only through an appreciation of evidence that it is reliable. When you have such evidence, as you often do, you can choose to trust yourself as an alternative to being thus self-alienated. When you lack such evidence, as you sometimes do, you can likewise choose to trust, provided you also lack significant evidence that your judgment is not reliable. In each case, you exercise your will by trusting at will. You regulate your trust, not through responsiveness to positive evidence of your judgment's reliability, but through responsiveness to possible evidence of your judgment's unreliability: if you come to have significant evidence that your judgment is not reliable, you will cease to trust, and (counterfactually) if you had had such evidence you would not have trusted.[2]

The key to my approach lies in distinguishing these two ways of being responsive to evidence. On the one hand, you cannot trust someone – whether another person or your own earlier self – whom you judge unworthy of your trust. Though you can for other reasons rely on a person whom you judge untrustworthy, it would be a mistake to

describe such reliance as "trust." On the other hand, trust does not require a positive assessment of trustworthiness. When you trust, you are responsive to possible evidence that the trustee is unworthy of your trust, and that responsiveness – your disposition to withhold trust when you encounter what you assess as good evidence that the trustee is unworthy of it – makes trust importantly different from a "leap of faith." You may lack sufficient evidence for the judgment that your trustee is worthy of your trust, but that evidential deficit need not constrain your ability to trust. Even if you have such evidence, that evidence should not form the basis of your trust.

I thus take issue with Pamela Hieronymi's influential analysis of trust as a "commitment-constituted attitude."[3] Your trust is indeed constrained by your responsiveness to evidence of untrustworthiness. But you need not undertake an attitudinal commitment to the trustee's worthiness of your trust: you need undertake no commitment akin to or involving a judgment that the trustee is trustworthy. An attitudinal commitment to a person's worthiness of your trust creates normative tension with your simply trusting her. If you judge that she is worthy of your trust, you need not trust her; you can rely, not directly on her in the way of trust, but on your own judgment that she will prove reliable. That sounds paradoxical. Are you thereby prevented from trusting those whom you deem most worthy of your trust? There is no paradox; there is merely a need to understand how we trust at will. A volitional element in trust enables you to enforce this distinction, trusting instead of merely relying on your judgment that the trustee is reliable.

Since the contrast between these two possibilities is clearest in intrapersonal trust, I make that my focus, thereby treating the question of "trust and will" as probing both the role of the will in trust and the role of trust in the will. We can see why trust is not a commitment-constituted attitude by seeing how trust itself plays a key role in the constitution of a commitment. In order to commit yourself to φing at *t*, you have to expect that your future self will at *t* have a rational basis for following through on the commitment not merely in a spirit of self-reliance but also, and crucially, in the spirit of self-trust.

11.1 Why Care about Trusting at Will?

What then is it to form a commitment? And how might it matter that the self-trust at the core of a commitment be voluntary? We can see how it might matter by considering the alternative. Say you intend to φ at *t* but just before *t* learn that context makes it imperative that you either assess your intending self as trustworthy before you act on the intention or redeliberate whether to φ from scratch. Imagine that it is too complicated to redeliberate from scratch but that materials for assessing your trustworthiness are available in the form of evidence that you were indeed reliable in making the judgment that informs your intention. If you proceed to act on that judgment, having made this assessment, you do not simply trust your judgment. Without that trust, your judgment that you ought to φ does not inform your intention to φ in the normal way.

You are not simply trusting your judgment if you require evidence that it is reliable. To say that you simply trust your judgment that you ought to φ is to say that you reason from it by, e.g. forming an intention to φ, or act on it by φing, without explicitly redeliberating. When you follow through on your intention to φ without redeliberating, your follow-through is not mediated by an assessment or reassessment of the self that judged that you ought to φ. There is, of course, the problem that you cannot keep assessing yourself – assessing your judgment that you ought to φ, assessing *that*

judgment, then *that* judgment, ad infinitum (or however high in this hierarchy you are able to formulate a thought). But my present point is different: there is an important contrast between (i) the common case in which you trust your judgment that you ought to φ by forming an intention to φ and then following through on that intention and (ii) the less common but perfectly possible case in which you feel a need to assess your judging or intending self for trustworthiness before feeling rationally entitled to follow through on it.

That distinction, between trusting yourself and relying on yourself through an assessment of your reliability, marks an important difference between two species of self-relation. The difference is functional, a matter of how the two stances ramify more broadly through your life. Do you second-guess yourself – forming a practical judgment or intention but then wondering how trustworthy that judgment or intention really is? In some regions of your life such self-mistrust may be perfectly appropriate – when you are learning a new skill or when a lot is at stake. But in the normal course of life you must exercise this virtue of temperance: to intend in a way that is worthy of your trust, and to act on that intention unless there is good evidence that you are not worthy of that trust. When there is no significant evidence of your untrustworthiness in intending, or any good reason to believe that circumstances have changed in relevant ways since you formed the intention, then you should trust yourself and follow through. Evidence of your own untrustworthiness constrains your capacity to trust yourself. But in the absence of such evidence you may avoid self-alienation by exercising your discretion to trust yourself at will.

In what respects is it "self-alienating" to rely on your responsiveness to evidence of your reliability instead of trusting yourself? One respect is simply that doing so does not amount to making a decision or choice, or to forming an intention. But why care about those concepts? What would you lose if you governed yourself without "making choices" or "forming intentions" but instead simply by monitoring the reliability of your beliefs about your practical reasons? One problem is that your beliefs about your reasons may pull you in incompatible directions, so you need the capacity to settle what to do by forming the "practical judgment" that you have conclusive or sufficient reason to do A, even though you may also believe you have good reasons to do incompatible B.[4] Another problem is that, because you have limited evidence about the reliability of your practical judgments, you will be unable to govern your reactions to novel issues. But what does either problem have to do with "self-alienation"? The threat of self-alienation marks the more general datum that our concepts of choice and intention enable us to govern ourselves even when we lack evidence that our beliefs about our reasons are reliable. Instead of responding to evidence of your reliability, your choice or intention manifests responsiveness to the normative dynamic of a self-trust relation.

What is that normative dynamic? And how does your responsiveness to it ensure that you are not self-alienated? We can grasp the distinctive element in self-trust by grasping how the distinctive element in trust lies in what it adds to mere self-reliance. In any form of reliance, including trust, you risk disappointment: the trustee may not do what you are relying on her to do.[5] But in trust, beyond mere reliance, you also risk betrayal – in the intrapersonal instance, self-betrayal.[6] We can understand the precise respect in which self-trust embodies the antidote to self-alienation by grasping how the risk of betrayal shapes the normative dynamic of a trust relation.

How exactly does trust risk betrayal? Annette Baier set terms for subsequent debate when she argued that the risk of betrayed trust is distinctively moral. What is most fundamentally at stake in a trust relation, Baier argued (1994:137), is not simply

whether the trustee will do what you trust her to do – on pain of disappointing your trust – but whether she thereby manifests proper concern for your welfare. I agree with Baier's critics that her approach over-moralizes trust.[7] But these critics link their worry about moralism with a claim that I reject: that the risk of betrayal adds nothing, as such, to the risk of disappointment. Baier is right to characterize the distinctive risk of trust as a form of betrayal, but the assurance that invites trust targets the trustor's rationality, not the trustor's welfare or any other distinctively moral status. I do not emphasize rationality to the exclusion of morality; I claim merely that the rational obligation is more fundamental: while it does not follow from how trust risks betrayal that trust is a moral relation, it does follow from how trust risks betrayal that trust is a rational relation.

I elsewhere defend that claim about interpersonal trust (see also Potter, this volume), arguing that the assurance at the core of testimony, advice or a promise trades on the risk of betrayal (Hinchman 2017; see also Faulkner, this volume). I review that argument briefly in section 11.4 below. In the next two sections I extend my argument to intrapersonal trust. To the objection that an emphasis on self-betrayal over-moralizes our self-relations, I reply that this species of betrayal is rational – not, as such, moral. To put my thesis in a single complex sentence: you betray yourself when, in undertaking a practical commitment, you represent yourself as a source of rational authority for your own future self without manifesting the species of concern for your future self's needs that would provide a rational basis for that authority. Such betrayed self-trust shapes the self-relations at the core of diachronic agency – of your forming and then later following through on an intention – by serving as a criterion of normative failure. The prospect of self-betrayal reveals how trust informs your will: when you exercise your will by forming an intention, you aim not to influence yourself in a self-alienated manner, through evidence of your reliability, as if your later self were a different person, but to guide yourself through trust, by putting yourself in position to treat your worthiness of that trust as itself the rational basis of that guidance. Trust could not thus inform your will if you could not trust at will, thereby willing a risk of self-betrayal.

11.2 How Betrayed Self-Trust Differs from Disappointed Self-Trust

How then does betraying your own trust differ from disappointing it? And how does the possibility of self-betrayal figure in the exercise of your will? When you form an intention, you institute a complex self-relation: you aim that you will follow through on the intention through trust in the earlier self that formed it. Such projected self-trust rests on a rational capacity at the core of trust: your counterfactual sensitivity to evidence of untrustworthiness in the trustee. Within this projection, if there is evidence that your earlier self is unworthy of your trust, you will not trust it, and if there had been such evidence, you would not have trusted it. Our question is what that sensitivity is a sensitivity *to*: what is it to be thus unworthy of trust? My thesis is that you are on guard against the prospect of betrayed, not of disappointed, self-trust. In a typical case of self-trust, as in a typical case of trust, you run both risks at once and interrelatedly. But we can learn something about the nature of each by seeing how they might come apart – in particular, how you might betray your self-trust without disappointing it.

First consider a general question about the relation between disappointed trust and betrayed trust. If A trusts B to φ, can B betray A's trust in her to φ without thereby disappointing that trust? If A's trust in B to φ amounts to something more

than her merely relying on B to φ, then we can see how B might betray A's trust even though she φs and so does not disappoint it. Perhaps B φs only because someone – perhaps A himself – coerces her into φing. Or perhaps B φs with no memory of A's trust in her and with a firm disposition not to φ were she to remember it. In either case, B betrays A's trust in her to φ, though she does φ and in that respect does not disappoint A's trust – even if A finds it "disappointing" (in a broader sense) that his trust has been betrayed. We are investigating the distinction specifically in intrapersonal or "self"-trust. Our challenge is to explain how there are analogues of these interpersonal relations in the relations that you bear to yourself as a single diachronically extended agent.

The first step toward meeting the challenge concedes a complexity in how you would count as "betraying" your own trust. In an interpersonal case, the trustee can simply betray the trustor's trust – end of story. But it is unclear how there could be a comparably simple story in which an individual betrays his own trust. Can we say that the individual is both the active "victimizer" and the passive "victim" of betrayed self-trust? If he worries that he is being "victimized," there is something he can do about that – stop the "victimizing"! As we will see, this is precisely where the concept of betrayed self-trust does its work: the subject worries that she is betraying her own self-trust and responds by abandoning the judgment that invites the trust. When she abandons the judgment, her worry about betrayal is thereby resolved. But the resolution reveals something important about intrapersonal trust: that the subject resolves this question of trust by responding to a worry, not about disappointed self-trust, but about betrayed self-trust. The following series of cases reveals how it is in the context of such a worry – and of such a resolution – that intrapersonal trust may figure as undisappointed yet betrayed.

Consider first a standard case of akrasia:

> *Tempted voter.* Ally is a firm supporter of political candidate X based on an impartial assessment of X's policies. She thereby judges that she has conclusive reason to vote for X, rather than for X's rival Y, and forms an intention to vote accordingly. While waiting in line to vote, however, she overhears a conversation that reminds her how X's policies will harm her personally, which in turn creates a temptation to vote for Y. Despite still judging that she has conclusive reason to vote for X, the temptation "gets the better of her" and she votes for Y. She almost immediately regrets her vote.

How should Ally have resolved this moment of akratic temptation? Her regret reveals that she ought to have resolved the akrasia in a "downstream" direction: by letting the judgment that she retains, even while tempted, guide her follow-through.[8] Does she betray her trust? We might think there is no intrapersonal trust here, since Ally fails to act on her intention, but that would overlook how she has trusted her intention to vote for X for weeks before election day. Imagine that she has campaigned for X, partly on the basis of her intention to vote for X. She thereby treats her trustworthiness in intending to vote for X as a reason to campaign for X – not as a sufficient reason unto itself, but as one element in a set of reasons that, she judges, suffices for campaigning. Simply put, if she had not intended to vote for X, she would not have regarded herself as having sufficient reason to campaign for X. And she does betray her trust in herself in that respect; she betrays her self-trust while also disappointing it.

We get one key contrast with a case that lacks this akratic element:

> *Change of Mind*. Amy arrives at the voting place judging that she has conclusive reason to vote for candidate X rather than the opposing candidate Y. But while in line to vote, she overhears a discussion that re-opens her deliberation whether to vote for X. After confirming the accuracy of these new considerations via a quick Internet search on her phone, Amy concludes that she has conclusive reason to vote for Y instead of X and marks her ballot accordingly.

Unlike Ally's worry, Amy's worry targets the deliberation informing her judgment. Ally does not worry about her deliberation whether to vote for X; she remains confident that she has decided the matter correctly – despite the fear of personal harm that generates her temptation to rebel against that judgment. But Amy does worry about her deliberation: specifically, she worries that she may have made a misstep as she conducted that deliberation, misassessing the considerations that she did assess (including evidence, practical principles, or anything else that served as input to her deliberation), or ignoring considerations available to her that she ought to have assessed. If, like Ally, Amy has campaigned for X partly on the basis of her intention to vote for X, then she disappoints her trust – but, unlike Ally, without betraying it. It is no betrayal if you fail to execute an intention that you come to see you ought to abandon.

Consider now a case with this different normative structure:

> *Change of Heart*. Annie, like Ally and Amy, arrives at the voting place judging that she has conclusive reason to vote for X rather than Y. But while in line she overhears a heartfelt tale of political conversion, wherein the speaker recounts her struggles to overcome the preconceptions that led her earlier to support candidates from X's party. Annie recognizes herself in the speaker's struggles; she was likewise raised to support that party uncritically. She wonders how this political allegiance might lead to similar regret – but, even so, the preconceptions are *her* preconceptions, and she finds it difficult to shake them. Though shaken by the felt plausibility of the hypothesis that she is untrustworthy, she continues to judge that she has conclusive reason to vote for X. When her turn to vote arrives, she stares long and hard at the ballot, unsure how to mark it.

Unlike Ally's worry in Tempted Voter, Annie's worry targets her judgment that she has conclusive reason to vote for X. But unlike Amy's worry in Change of Mind, Annie's worry does not target the deliberation informing her judgment – or, at least, not in the way that Amy's does. Annie does not worry that she has made a misstep as she conducted that deliberation. She is perfectly willing to take at face value her confidence that she has correctly assessed all the considerations that she did assess, and that she did not ignore any consideration available to her that she ought to have assessed within that deliberation. Her worry instead targets the "sense of" or "feeling for" what is at stake for her in the deliberative context that informs how she is guided by this confidence, her broader confidence not merely that she has correctly assessed everything she did assess, among those considerations available to her, but that she has considered matters well and fully enough to permit drawing a conclusion. She worries that her feeling of conclusiveness – her sense that she has considered matters long enough and well enough to justify this conclusion – may not be reliably responsive to what is really at stake for her. Though she cannot shake this sense of the stakes, she worries that she

ought to try harder to shake it. As long as she thereby retains the judgment without letting it guide her, Annie counts as akratic. But her akrasia is crucially unlike Ally's in Tempted Voter. Whereas Ally betrays her trust while also disappointing it, Annie fears she will betray her trust by failing to disappoint it. If the metaphor for Ally's predicament is *weakness*, the metaphor for Annie's predicament is *rigidity*.

We might thus describe the phenomenological difference between the three cases. But what are the core normative differences? The first difference is straightforward: Change of Heart generates a second-order deliberation, whereas Change of Mind generates a first-order deliberation. Here are two possible bases for the second-order deliberation in Change of Heart: Annie may worry that impatience makes her hasty, or she may worry that laziness makes her parochial. Whichever way we imagine it, the fundamental target of Annie's worry is her feeling for what is at stake: specifically, her sense of how much time or energy she should devote to the deliberation informing her judgment. Each addresses not her truth-conducive reliability but, to coin a term, her *closure-conducive* reliability. She does not re-open her first-order deliberation as Amy does, by suspending her earlier presumption that she is truth-conducively reliable about her reasons. Unlike Amy, she continues to judge that she has conclusive reason to vote for X. What Annie questions is *whether* to suspend her presumption of truth-conducive reliability – that is, whether to re-open her first-order deliberation. In asking this question, she suspends the presumption that she is closure-conducively reliable, the presumption that informs her sense that she is entitled to treat that first-order deliberation as closed.

How does Annie undertake this higher-order species of reflection? What is it to question one's own closure-conducive reliability? This leads us to the second normative difference between Change of Heart and Change of Mind. How could Annie adopt a mistrustful higher-order perspective on whether to trust her own first-order deliberative perspective?

11.3 How the Prospect of Betrayed Self-Trust Plays Its Normative Role

Annie's higher-order perspective on her first-order judgment projects a broader future for her, insofar as it crucially involves an attitude toward her own future regret. Unlike reflection on her truth-conducive reliability, reflection on her closure-conducive reliability represents her agency as extending not merely to the time of action but out to the horizon that Michael Bratman calls *plan's end*, the point beyond which she will no longer think about the action.[9]

We can codify this forward-looking reflection as follows. When Annie judges that she has conclusive reason to vote for X, she projects a future, out beyond election day, in which:

(Down) (a) she will not regret having voted for X, and
(b) she will regret not having voted for X.

But when Annie worries about the trustworthiness of this judgment, thereby deliberating whether to redeliberate, she projects a future, out beyond election day, in which:

(Up) (a) she will regret having voted for X, and
(b) she will not regret not having voted for X.

I have labeled the projection that emerges from the perspective of judgment "Down" because it points downstream: Annie will have nothing to regret if she trustingly commits herself to this judgment and then acts on the commitment. This is how Ally struggles with temptation: her viewing it as "temptation" rather than an occasion to change her mind derives from the downstream-pointing projection of her judgment. She believes that she will regret giving in to the "temptation" because she expects that it will amount to a merely transient preference reversal. By contrast, I have labeled the projection that emerges from Annie's mistrust in her judgment 'Up' because it points upstream: she will regret it if she lets herself be thus influenced by her judgment, and she will not regret it if she does not let herself be thus influenced. This regret does not mark a merely transient preference reversal within the projection but expresses her settled attitude toward the self-relation that she manifests in thus following her judgment.

Why should the concept of regret play this role in structuring the two projections? Here is my hypothesis: regret plays this normative role as the intrapsychic manifestation of betrayed self-trust. As others have emphasized,[10] betrayal finds its natural expression in reactive attitudes, engendering contempt or resentment in the betrayed toward the betrayer. If regret functions as an intrapersonal reactive attitude, that enables the concept of betrayed trust to shape self-governance in prospect, as referring not to something actual but to something to be kept non-actual – on pain of regret. It could not play this role if it – that is, betrayed self-trust experienced as regret that you trusted your judgment – were not something with which we are familiar in ordinary experience. Such regret is common in two sorts of case, in each of which the subject is concerned for her intrapersonal rational coherence, not merely for her welfare.[11]

First, we do sometimes make bad choices that we regret in this way. "What could I have been thinking?" you ask yourself at plan's end, appalled that you trusted a judgment that now seems manifestly unworthy of your trust. This experience is crucially unlike merely being displeased by the results of following through on a judgment. You may well be displeased with the results of trusting your judgment yet not regret the self-influence as such. You may think you did your best to avoid error yet fell into error anyway. Or you may temper your self-criticism with the thought that no evidence of your own untrustworthiness – including your closure-conducive unreliability – was then available to you. If there was no evidence of untrustworthiness available to you when you trusted, then your trust was not unreasonable, however displeased you may be with the results. In an alternative case, however, you may think that there was evidence of your own untrustworthiness available, and that you trustingly followed through on your judgment through incompetence in weighing that evidence. That case motivates a deeper form of regret that targets your self-relations more directly.

Here then is the second source of everyday familiarity with betrayed self-trust. As we mature, we do much that we wind up regretting in this way: you judge that you have conclusive reason to φ, trust that judgment because you are too immature to weigh available evidence of your untrustworthiness, then later realize your mistake. Your question is not: "What could I have been thinking?" It is all too clear how immaturity led you to deliberative error. One of our developmental aims is to learn to make judgments that will prove genuinely authoritative for us.

How might Annie's judgment fail to be authoritative? Here, again, is my answer: she fears that her judgment will, looking back, appear to have betrayed her own trust. The answer presents the case in all its diachronic complexity, wherein the subject looks ahead not merely to the time of action but all the way out to "plan's end."

Annie fears that the deliberative perspective informing her judgment does not manifest the right responsiveness to her ongoing – and possibly changing – needs. When she reasons "upstream," she aims to feel the force of these needs from plan's end, by projecting a retrospect from which she would feel relevant regret.[12] We thus return to the idea from which we began: though reactive attitudes are the key to distinguishing trust from mere reliance, they need not be moral. Our focus on reactive attitudes reveals not their moral but their rational force: they target the subject's rational authority and coherence. Annie's projection out to plan's end serves as a reactive-attitudinal retrospect on her planning agency, not because it represents her as planning through that entire interval, but because it represents her as having settled the question of her needs in more local planning. The local planning that informs her voting behavior, with its implications for broader planning, requires that she view her judgments as rationally adequate to that exercise of self-governance. And her reactive-attitudinal stance from plan's end settles whether her judgments were indeed thus adequate insofar as they avoided self-betrayal in the way they presumed. Her self-mistrustful attitude in the voting booth both projects this verdict and uses the verdict as a basis for assessing the presumption.

When you reason "upstream" – abandoning your judgment because you mistrust it – you show responsiveness to the possibility of betrayed self-trust. Such self-mistrust does not entail betrayed self-trust, since it is possible that you do care appropriately about what is at stake for you in your deliberative context and therefore that your self-mistrust is mistaken. But it is possible that your self-mistrust is not mistaken: perhaps you really have betrayed the invited self-trust relation. The responsiveness at the core of trust is a rational responsiveness because it targets the possibility that your trust in this would-be source of rationality has been betrayed.

11.4 Inviting Others to Trust at Will

How does this intrapersonal normative dynamic run in parallel with an interpersonal dynamic? The intrapersonal dynamic unfolds between perspectives within the agency of a single person, as the person acts on an aim to bring those perspectives into rational coherence. The interpersonal dynamic, by contrast, engages two people with entirely separate perspectives that cannot, without pathology, enter into anything like that coherence relation. Interpersonal trust must therefore engage an alternative rational norm – but what norm? It helps to reflect on a parallel between intending and promising: just as you invite your own trust when you form an intention to φ, so you invite the trust of a promisee when you promise him that you will φ. In neither case does the invitation merely prompt the invitee to respond to evidence of your reliability in undertaking the intention or promise. In each case, you aim that the recipient of your invitation should trust you and feel rationally entitled to express that trust through action – following through on your intention in the first case, performing acts that depend on your keeping your promise in the second – even in the absence of sufficient evidence that you are worthy of the trust, as long as there is no good evidence that you are unworthy of it. And we can make similar remarks about other forms of interpersonal assurance – say, testimony and advice. The parallel reveals something important about the value of a capacity to trust at will. Both intrapersonally and interpersonally, a capacity for voluntary trust makes us susceptible to rational intervention – whether to preserve our rational coherence or to give us reasons we would not otherwise have.

As in the intrapersonal case, the rational influence unfolds through two importantly different perspectives. Take first the perspective of the addressee, and consider the value in trusting others, beyond merely relying on your own judgment that another is relevantly reliable. If someone invites your trust by offering you testimony, advice, or a promise, and evidence is available that the person is relevantly reliable, you can judge that she is reliable on the basis of that evidence and on that basis believe what she testifies, or do what she advises, or count on her to keep her promise – on the basis, that is, of *your* evidentially grounded judgment that she will prove to be or have been relevantly reliable. But what if no such evidence is available? Or what if, though the evidence is available, there is insufficient time to assess it? Or what if she would regard your seeking and assessing evidence of her worthiness of your trust as a slight – as a sheer refusal of her invitation to trust? You might on one of these bases deem it preferable to trust without seeking evidence of her worthiness of your trust – as long as you can count on your capacity to withhold trust should evidence of her unworthiness of your trust become available. Here again we see why it might prove a source of value to be capable of trusting at will. Though you could not trust if there were evidence that the would-be trustee is unworthy of your trust, if there is no such evidence you can decide to trust merely by disposing yourself to do so.

Is this a moral value inhering in the value of the trust relation? As in the intrapersonal case, that over-moralizes trust. Say, after asking directions, you trust the testimony or advice of a stranger on the street. Or say you trust your neighbor's promise to "save your spot" in a queue. Do you thereby create moral value? Moral value seems principally to arise on the trustee's side, through whatever it takes to vindicate your trust. Setting morality aside, a different species of value can arise on your side of the relation. Assuming the trustee relevantly reliable, you can acquire a reason that you might not otherwise have – a reason to believe her testimony, the follow her advice or to perform actions that depend on her keeping her promise. This reason is grounded partly in the trustee's reliability and partly in the (counterfactual) sensitivity to evidence of the trustee's unreliability that informs your trust: if you have (or had) such evidence you would cease trusting (or would not have trusted). The latter ground marks the difference between a reason acquired through trust and a reason acquired through mere reliance. Sometimes you cannot trust a person on whom you rely, because evidence of her unreliability forces you to rely, not directly on her, but on your own judgment that relying on her is nonetheless reasonable. But when you lack such evidence you can get the reason by choosing to trust her – even if you have evidence that would justify relying on her without trust.

How could a reason be grounded even partly in your trusting sensitivity to evidence of the trustee's unworthiness of your trust? The key lies in understanding how the illocutionary norms informing testimony, advice and promising codify your risk of betrayal. The normative basis of your risk of disappointment lies in your own judgment: you judge that the evidence supports reliance that would incur this risk, so the responsibility for the risk itself lies narrowly on your side – whatever else we may say about responsibility for the harms of disappointing your reliance. When you trust, however, responsibility for the risk of betrayal you thereby undergo is normatively distributed across the invited trust relation. What explains this distribution? In the cases of assurance at issue, you trust by accepting the invitation that informs the trustee's assurance, which is informed by the trustee's understanding of how that response risks betrayal. You thus respond to the trustee's normative acceptance of responsibility for that risk – something that has no parallel in mere reliance. You can

trust "at will" because you can choose to let yourself be governed by the trustee's normative acceptance of responsibility for how you are governed, an exercise of will that may give you access to reasons to do or believe things that you would not otherwise have reason to do or believe, but at the cost of undergoing the risk that this trustee will betray you, When I say that the trustee accepts "normative" responsibility, I mean that she thereby commits herself to abiding by the norms that codify that responsibility. If she is insincere, she flouts those norms and in that respect does not even attempt to live up to the responsibility she thereby incurs. This is one principal respect in which your trust risks betrayal.

The normative nature of this exchange emerges more fully from the other side. When you offer testimony, advice or a promise, do you merely "put your speech act out there," aiming to get hearers to rely on you for whatever reasons the evidence available to those hearers can support? If that were your aim, your testimony would not differ from a mere assertion, your advice would not differ from a mere assertion about your hearer's reasons, and your promise would not differ from a mere assertion of intention. What is missing in these alternative acts is the distinctive way in which you address your testimony, advice or promise: you invite your addressee's trust. In inviting his trust, you engage your addressee's responsiveness to evidence of your unworthiness of his trust, and thereby to the possibility that his trust might be betrayed. But you more fundamentally engage your addressee's capability to draw this distinction in his will: to trust you *instead* of merely relying on you through an appreciation of positive evidence that you are reliable. If he can do the first, then he can also do the second – perhaps irrationally (if there is insufficient evidence that you are reliable). Why should he trust? The simplest answer is that that is what you have invited him to do. In issuing that invitation, you aim at this very exercise of will – that he should trust you at will. You take responsibility for the reason you thereby give him (assuming you reliable) by inviting him to rely on your normative acceptance of responsibility for the wrongfulness of betraying the trust you thereby invite.

What if you do not believe that you *can* engage your addressee's capacity for trust, because you believe that there is good evidence available to this addressee that you are not worthy of it?[13] You thereby confront an issue that you can attempt to resolve in either of two ways. You can attempt to counter the appearance that this evidence of your untrustworthiness is good evidence, thereby defusing its power over your addressee's capacity to trust you. Or you can shift to the alternative act, attempting instead to get your addressee to rely on you for reasons of his own – including perhaps a reason grounded in evidence that you are, on balance, relevantly reliable. On this second strategy, you now longer invite the addressee's trust. The only way to invite his trust – without insincerity or some other normative failure – is to counter the appearance that you are unworthy of it. You must counter this appearance because without doing so you cannot believe that your addressee will enter freely – "at will" – into this trust relation, by accepting your invitation, not by responding to positive evidence of your reliability but by relying on you in the way distinctive of trust. In inviting trust, you aim at willed trust.

By this different route we again contrast the intimacy of trust with a form of alienation. In intrapersonal and intrapersonal cases alike, governance through trust contrasts with an alienated relation mediated by evidence. To the worry that an emphasis on betrayal over-moralizes trust, I reply that the norms informing each relation are rational, not moral. Appreciating the rational force of the trustee's invitation to trust helps us grasp the parallel species of intimacy at stake in the intrapersonal relation. As

a self-governing agent, each of us sometimes encounters Annie's predicament in Change of Heart: she fears that follow-through on her voting intention may amount to self-betrayal. In that worst-case scenario for her rational agency, as she followed through she would manifest trust in her intending self but betray that trust by not having adequately served, in the judgment that informs her intention, the needs of her acting self – as she will learn when she regrets from plan's end. You run that risk whenever you follow through on an intention: you risk betrayal by your own practical judgment. Like Annie, you can address the risk by being open to a change of heart. But every time you form an intention you are already like Annie in this respect: you are responsive to the possibility that you ought to undergo such a change of heart, and you aim to avoid that possibility. Your aim as you commit yourself looking downstream thus acknowledges not merely the psychological but also the normative force of upstream-looking self-mistrust. As you judge or intend, you thereby acknowledge the normative bearing of your capacity to trust at will.[14]

Notes

1 Self-trust has been a topic in recent epistemology (e.g. Foley 2001 and Zagzebski 2012) and in discussion of the moral value of autonomy (e.g. McLeod 2002). My angle on self-trust is different: I am interested in its role in action through time, without any specifically moral emphasis.
2 Nothing in what follows turns on any difference between trustworthiness and reliability: by "reliability" I mean the core of what would make you worthy of trust.
3 For the view that trust is constituted by a commitment (by a commitment-constituting answer to a question), see Hieronymi (2008) and McMyler (2017). For more general treatments of "commitment-constituted attitudes," see Hieronymi (2005, 2009).
4 For more on this, see Watson (2003).
5 To cover trust in testimony or advice, we can modify this to include the trustee's not being as you trust her to be (viz. reliable in relevant respects).
6 Many philosophers join me in holding that trust distinctively risks not mere disappointment but betrayal. See, e.g. Baier (1994: chapters 6–9); Holton (1994); Jones (1996, 2004); Walker (2006: chapter 3); Hieronymi (2008); McGeer (2008); McMyler (2011: chapter 4); and Hawley (2014).
7 For example, Hardin (2002: chapter 3); Nickel (2007: section 6); and Rose (2011: chapter 9).
8 I take the "stream" metaphor from Kolodny (2005: e.g. 529).
9 Bratman (1998, 2014). Though I am indebted to Bratman for the idea that projected regret is crucial to the stability of intention, in Hinchman (2010, 2015, 2016) I dissent from some details in how he develops it.
10 See note 3 above, especially Holton (1994).
11 One background issue: does your capacity to serve as a source of rationality for your future self license illicit bootstrapping, whereby you get bumped into a rational status "for free" merely by forming an intention? For developments of this worry, see Bratman (1987:23–27, 86–87); and Broome (2001). Smith (2016) offers a dissenting perspective. I treat the issue in Hinchman (2003, 2009, 2010, 2013, 2017: sections II and IV). For present purposes, it does not strictly matter how I reply to the bootstrapping challenge, both because my present argument could work in conjunction with the weaker view that we operate under an ("error-theoretic") fiction that we can give ourselves the rational status (following Kolodny 2005: section 5) and because my core claim is not that a trustworthy intention to φ gives you a reason to φ (which does look like bootstrapping) but that it (a) gives you "planning reasons" to do other things – things that you would not have sufficient reason to do if you were not trustworthy in intending to φ – and (b) more generally serves as a source of rational coherence.
12 I offer a much fuller defense of upstream reasoning (replying to the objections in Kolodny (2005), 528–539) in Hinchman (2013: section 3). The most fundamental challenge lies in explaining how it is possible for you to judge that you ought to φ while also mistrusting that judgment. If this is impossible, then reasoning must always point "downstream," since mistrusting a judgment would simply amount to abandoning it.

13 This is not precisely the question of "therapeutic trust" (see e.g. Horsburgh 1960; Holton 1994; Jones 2004; McGeer 2008). But it raises a question about whether you can *invite* therapeutic trust.

14 Thanks to Ben McMyler, Philip Nickel, and Judith Simon for stimulating comments on an earlier draft. I develop this view of intrapersonal trust more fully in Hinchman (2003, 2009, 2010, 2016). And I develop this view of interpersonal trust more fully in Hinchman (2005, 2014, 2017 and forthcoming).

References

Baier, A. (1994) *Moral Prejudices*, Cambridge, MA: Harvard University Press.

Bratman, M. (1987) *Intention, Plans, and Practical Reason*, Cambridge, MA: Harvard University Press.

Bratman, M. (1998) "Toxin, Temptation, and the Stability of Intention," reprinted in his *Faces of Intention*, Cambridge: Cambridge University Press, 1999.

Bratman, M. (2014) "Temptation and the Agent's Standpoint," *Inquiry* 57(3): 293–310.

Broome, J. (2001) "Are Intentions Reasons? And How Should We Cope with Incommensurable Values?" in C. Morris and A. Ripstein (eds.), *Practical Rationality and Preference*, Cambridge: Cambridge University Press, 98–120.

Faulkner, P. and Simpson, T. (eds.) (2017) *The Philosophy of Trust*, Oxford: Oxford University Press.

Foley, R. (2001) *Intellectual Trust in Oneself and Others*, Cambridge: Cambridge University Press.

Hardin, R. (2002) *Trust and Trustworthiness*, New York: Russell Sage Foundation.

Hawley, K. (2014) "Trust, Distrust, and Commitment," *Noûs* 48(1): 1–20.

Hieronymi, P. (2005) "The Wrong Kind of Reason," *Journal of Philosophy* 102(9): 437–457.

Hieronymi, P. (2008) "The Reasons of Trust," *Australasian Journal of Philosophy* 86(2): 213–236.

Hieronymi, P. (2009) "Controlling Attitudes," *Pacific Philosophical Quarterly* 87(1): 45–74.

Hinchman, E. (2003) "Trust and Diachronic Agency," *Noûs* 37(1): 25–51.

Hinchman, E. (2005) "Advising as Inviting to Trust," *Canadian Journal of Philosophy* 35(3): 355–386.

Hinchman, E. (2009) "Receptivity and the Will," *Noûs* 43(3): 395–427.

Hinchman, E. (2010) "Conspiracy, Commitment, and the Self," *Ethics* 120(3): 526–556.

Hinchman, E. (2013) "Rational Requirements and 'Rational' Akrasia," *Philosophical Studies* 166 (3): 529–552.

Hinchman, E. (2014) "Assurance and Warrant," *Philosophers' Imprint* 14, 1–58.

Hinchman, E. (2015) "Narrative and the Stability of Intention," *European Journal of Philosophy* 23 (1): 111–140.

Hinchman, E. (2016) "'What on Earth Was I Thinking?' How Anticipating Plan's End Places an Intention in Time," in R. Altshuler and M. Sigrist (eds), *Time and the Philosophy of Action*, New York: Routledge, 87–107.

Hinchman, E. (2017) "On the Risks of Resting Assured: An Assurance Theory of Trust," in P. Faulkner and T. Simpson (eds.), *The Philosophy of Trust*, Oxford: Oxford University Press.

Hinchman, E. (forthcoming) "Disappointed yet Unbetrayed: A New Three-Place Analysis of Trust," in K. Vallier and M. Weber (eds), *Social Trust*, New York: Routledge.

Holton, R. (1994) "Deciding to Trust, Coming to Believe," *Australasian Journal of Philosophy* 72 (1): 63–76.

Horsburgh, H.J.N. (1960) "The Ethics of Trust," *Philosophical Quarterly* 10(41): 343–354.

Jones, K. (1996) "Trust as an Affective Attitude," *Ethics* 107(1): 4–25.

Jones, K. (2004) "Trust and Terror," in P. DesAutels and M. Walker (eds), *Moral Psychology*, Lanham: Rowman & Littlefield, 3–18.

Kolodny, N. (2005) "Why Be Rational," *Mind* 114(455): 509–563.

McGeer, V. (2008) "Trust, Hope, and Empowerment," *Australasian Journal of Philosophy*, 86(2): 237–254.

McLeod, C. (2002) *Self-Trust and Reproductive Autonomy*, Cambridge: MIT Press.

McMyler, B. (2011) *Testimony, Trust, and Authority*, Oxford: Oxford University Press.

McMyler, B. (2017) "Deciding to Trust," in P. Faulkner and T. Simpson (eds.), *The Philosophy of Trust*, Oxford: Oxford University Press.

Nickel, P. (2007) "Trust and Obligation-Ascription," *Ethical Theory and Moral Practice* 10(3): 309–319.
Rose, D. (2011) *The Moral Foundation of Economic Behavior*, Oxford: Oxford University Press.
Smith, M. (2016) "One Dogma of Philosophy of Action," *Philosophical Studies* 73(8): 2249–2266.
Walker, M.U. (2006) *Moral Repair*, Cambridge: Cambridge University Press.
Watson, G.(2003) "The Work of the Will," in S. Stroud and C. Tappolet (eds.), *Weakness of Will and Practical Irrationality*, Oxford: Oxford University Press.
Zagzebski, L. (2012) *Epistemic Authority*, Oxford: Oxford University Press.