Trust, Mistrust, and Autonomy

Edward Hinchman and Andrea Westlund

Florida State University

[To appear in Mark Alfano & David Collins (eds.), *The Moral Psychology of Trust* (Lexington Books, 2023)]

Is autonomy compatible with trust? To pose the question in a stark form that we'll discuss: is governing yourself compatible with letting yourself be governed by trust in another? There are two broad reasons to believe that autonomy must be compatible with trust. One is that we regard trust relationships as vital to a life well lived, and it would be odd if you could pursue such a life only at a cost to your autonomy. Another reason is that it is common to see people in the grip of egotism and self-obsession 'get in their own way' or even 'trip over themselves' while pursuing a life well lived, an impairment that seems best repaired by trust in another – by letting another, or another's judgment, guide you out of your impasse, even if the trust does not blossom into anything worth calling a 'relationship.' It would be odd if self-governance required you to block or trip over yourself rather than letting another help you on your way. The two reasons are interrelated: trust relationships are vital to a life well lived partly because they provide a context for such assistance. But, even apart from any fully formed trust relationship, such trusting acquiescence to the will of another seems fundamental to the practice of autonomy. Autonomy – that is, governing yourself as you lead your life – seems to require such openness to influence.

One risk of such openness to the influence of others is that it renders you vulnerable to manipulation. Is autonomy compatible with this risk? One of us (Westlund 2003, 2009) has argued that the self-governing agent treats herself as answerable for her action-guiding

commitments, where answerability requires openness to the rational influence of external, critical perspectives on those commitments. Taking responsibility for your action-guiding commitments by holding yourself answerable for them implicitly requires you to trust your interlocutors enough to treat their critical interventions as worthy of response. It requires trusting them enough, that is, to allow their practical perspectives to exert an influence on your own – at least to extent of prompting you to review and reconsider the reasons you take yourself to have. But a manipulative or disingenuous interlocutor might exploit such answerability to his own ends, causing you to doubt yourself where you shouldn't and to abandon commitments you should, instead, hold dear.[1] In such cases, openness to correction might appear to undermine autonomy instead of supporting it.

Does an antinomy lie at the heart of autonomy? Does autonomy require an openness to influence difficult to distinguish from heteronomy? Though autonomy works in part through a capacity for trust, we argue that autonomy is undermined only by inappropriate trust, wherein the the truster fails to be appropriately responsive to evidence that the trustee is unworthy of that trust. The problem isn't that autonomy requires dispensing with trust, but rather that trust must itself be governed by appropriate responsiveness to evidence of untrustworthiness.

Our view confronts two interrelated challenges. One challenge lies in explaining why you shouldn't treat your vulnerability to manipulation as a general reason to let your trusting disposition lapse into an untrusting one. We argue that there is no such general reason to let your responsiveness to evidence of untrustworthiness be supplanted by a suspicious disposition to seek and require positive evidence of trustworthiness. A second, and deeper, challenge arises

[1] Ebels-Duggan (2015) and Perez de Calleja (2019) both raise versions of this worry.

from the fact that in some especially sophisticated cases of manipulation there is no evidence of untrustworthiness to which you could respond. We argue that you cannot rationally cope with even this unresolvable risk of exploited trust by trusting only yourself. Autonomy requires appropriate trust, we argue, because an openness to rational influence must be apportioned appropriately between trust in self and trust in others.

I. How trust might be exploited

Anyone who accepts that autonomy is compatible with trust must therefore explain how autonomy is compatible with the possibility that trust will be exploited. Exactly how does the possibility of exploited trust pose a challenge to autonomy?

There are three different ways your trust might be exploited, and only one of them points to an apparently worrisome tension between autonomy and trust. It poses no difficulty if your trust is exploited by someone whom you trusted mistakenly from your own perspective. Say an impartial onlooker with full knowledge of your circumstances would say that you should not trust B because there is evidence available to you that B is not worthy of your trust. If you trust B, despite this evidence available to you that B is not worthy of your trust, that's on you, since your trust is mistaken from your own perspective, which includes this available evidence. (To say that the evidence is 'available to you' is to say that it figures as part of your perspective.) It does not tend to show that autonomy is incompatible with trust to note that trust that is, from your own perspective, *mistaken* can be exploited. Again, by 'mistaken' trust we mean trust that is not appropriately responsive to evidence of untrustworthiness, and we assume that evidence of

the trustee's unworthiness of your trust is available to you. Trust is not reasonable in this case, and we only mean to argue that autonomy is compatible with (and requires) *reasonable* trust.

Many cases of exploited trust are cases of this first sort, in which evidence of untrustworthiness is available but mistakenly overlooked or ignored. But not all cases are like that. Sometimes, trust that is not from your own perspective mistaken is nonetheless exploited. Imagine a case in which you've done 'your best': you've governed yourself as well as anyone in your circumstances – that is, from the perspective defined by those circumstances – possibly could. No evidence of your interlocutor's untrustworthiness is available to you, and you proceed in trusting him. Yet your trust is nonetheless exploited, and your agency is in that respect manipulated. If this trust is reasonable, then it might appear that autonomy is threatened by the very capacity we take it to require. But this case is less worrisome than it might at first appear. Governing yourself as well as you possibly could in the circumstances does not "fireproof" you against violations of autonomy, including even violations that take advantage of a capacity that is required for autonomy. Coercion and blackmail, for example, both take advantage of a capacity at the heart of autonomy, namely, the capacity to respond rationally to threats to things that you value (including threats of various sorts to your own well-being). Manipulation that works by co-opting capacities required for autonomy may *violate* your autonomy without pointing to any deep tension between autonomy and the exercise of those capacities.

This brings us to the deepest, and in our view, most interesting sort of case in which an agent's capacity to trust may be targeted. In this third type of case, we argue, autonomy *is* more deeply threatened, but in a way that our view would predict and explain. Consider the example of

indoctrination discussed by Mirja Pérez de Calleja (2019).[2] Drawing on recent studies of youth radicalized by online exposure to ISIS propaganda, Pérez de Calleja observes that some victims of indoctrination are targeted in ways that appear to prey precisely on their responsiveness to alternative points of view. When they encounter online recruiters, the youth on whom she focuses appear to hold themselves answerable for their commitments, according normative standing to justificatory challenges coming from different perspectives (Pérez de Calleja 2019, 205). Indeed, their responsiveness to critical challenge is precisely what draws them into dialogue with online recruiters, whose tactics gradually induce them to distance themselves from family and friends and from their own previous worldview. Their transformation appears to be replete with justificatory dialogue. And yet, from the point of view of family, friends, therapists, and even ex-converts themselves, it looks anything but autonomous. In short, it is a case in which openness to external, critical perspectives seems to be precisely what draws the agents into a state of heteronomy.

Now, if the recruit is simply ignoring signs of untrustworthiness in the recruiters that are available to her, then her trust is unreasonable, and the case collapses into the first type discussed above. If, on the other hand, the recruiters are so skilled that they leave no trace of untrustworthiness, and the recruit is simply "doing her best" to govern herself in the circumstances, then it collapses into the second type. But something more complex seems to be going on in the cases described by Pérez de Calleja. The indoctrination seems to target the agent's capacity for reasonable trust itself, employing tactics that undermine her ability to judge what counts as a "red flag" and thereby undermining her ability to detect and respond appropriately to signs of untrustworthiness. Pérez de Calleja argues that what undermines the

[2] This example is discussed more briefly, but in broadly similar terms, in Westlund (2022).

autonomy of the apparently self-answerable converts in her example is not the viciousness of the views they adopt, but the fact that they are drawn into justificatory dialogue in the context of an environment of intense emotional pressure, which cultivates a high degree of fear and anxiety.

We think that the problem is not the fear and anxiety itself, but a problem created by the fear and anxiety – a problem with one's capacity for reasonable trust. This interpretation accords with accounts cited by Pérez de Calleja: sociologist Dounia Bouzar (2017) notes that indoctrinators typically use conspiracy theories to undermine their targets' trust in family, friends, and authority figures (cited in Pérez de Calleja 2019, 196) – a process that leads the converts to develop an unusual degree of emotional dependence on the indoctrinator. The indoctrinator, in effect, exploits the receptiveness of the agent to undermine her confidence in her own judgment – including, specifically, judgments pertaining to evidence of untrustworthiness – and then replaces her previous commitments with a ready-made, self-enclosed world-view that is impenetrable to further doubt. By the end of the process, the convert trusts only the indoctrinators. *This* trust is not reasonable – but the agent is no longer in a position to assess it as such. We cannot straightforwardly say that evidence of untrustworthiness either is or is not available to the agent, because the indoctrination has worked by undermining the agent's ability to detect and respond to such evidence, creating an emotional dependence on the manipulator to make such assessments for her.

This form of exploitation of trust is in obvious respects akin to gaslighting. Gaslighting, on Kate Abramson's influential account, is a form of emotional manipulation that aims to eliminate the very possibility of disagreement with the gaslighter, by undermining the target's

capacity to judge that she has been wronged (Abramson 2014).[3] The indoctrinator's manipulation of the recruit is similar, insofar as it aims to create a kind of emotional dependency on and deference to the manipulator by undercutting an aspect of the target's autonomy. In the indoctrination case, the manipulator proceeds by directly targeting the agent's capacity for reasonable trust and subverting it to their own ends. Since our view is that autonomy requires reasonable trust, it is to be expected that tactics that undermine the capacity for reasonable trust will undermine autonomy.

II. How exploited trust is a challenge to autonomy

In the previous section, we identified three different ways in which an agent's trust might be exploited. In the first type of case, evidence of untrustworthiness is available to the agent, but she does not respond appropriately. Her trust is thus unreasonable. In the second, no evidence of untrustworthiness is available to the agent, so her trust is reasonable but unfortunate. In the third, the agent's capacity for reasonable trust is itself compromised by the manipulation – and there is, therefore, no comparably straightforward answer to the question of whether her trust is reasonable or unreasonable. Her autonomy is impaired because a capacity on which it depends (the capacity for reasonable trust) has been compromised.

A quick reply to the challenge of reconciling autonomy with trust would observe that the first two forms of exploited trust don't engage any question about the truster's autonomy, while the third engages a question about the truster's autonomy in a way that confirms our relational

[3] One of us (Hinchman 2022) has defended an alternative approach to gaslighting that is compatible with Abramson's account on this core point.

view that trust is crucial to autonomy. We have already explained why the first two cases pose no threat to autonomy: the first is a misfiring exercise of autonomy (you try to be autonomous but fail), and the second a merely misfortunate exercise (your exercise of autonomy runs aground through no fault of your own). If exploited trust poses a threat to autonomy, that threat must arise in the third case. But in that case the trustee exploits precisely the truster's autonomy, on the relational view of autonomy that we develop in this paper. It is because autonomy has this relational nature that trust can be exploited in this way – a confirming instance of our view.

We hold that this quick reply dispenses with the challenge, but we concede that it does so by assuming our relational view of autonomy. It does not engage the opponent who rejects our view of autonomy, and a fuller defense of that view must begin from a less tendentious set of assumptions. We offer this fuller defense by focusing not on the actuality of exploited trust but on its mere possibility. We imagine our opponent to be worried about a fourth case, in which the trustee does not actually exploit the trust but might have done so. Imagine a case just like the third exploitation case but in which the trustee does not exploit the trust. To make it vivid in a crude case (we'll consider a better case presently), imagine that the trustee plans to exploit the trust, in the autonomy-compromising way that we described, but is pricked by conscience to abandon the plan just before the time comes to execute it – good fortune for the truster. This possibility – one not delineated by available evidence of untrustworthiness (since it mimics the third case) – appears to make the would-be exercise of autonomy rest, from the truster's perspective, on good fortune or, less optimistically, on sheer dumb luck. How is it compatible with autonomy – with governing yourself – that the successful realization of your practical stance should thus rest on fortune or luck?[4] By the 'successful realization' of your stance, we do

[4] Note that this does not parallel 'Frankfurt cases' (1969). The question here is not whether A acts

not mean success in obtaining a sought-for result. We mean realization of the autonomy that you presume that you are exercising. How could the realization of autonomy depend on the happenstance that one whom you trust does not, in the third way we described, exploit your practical stance?

Let us offer a better case to illustrate this possibility. We approach this fourth kind of case, in which trust is not exploited, by asking what it would take for this trusted agent to count as worthy of your trust. If she is not worthy of your trust, we might map the case onto our second exploitation case: just as you can exercise your autonomy misfortunately, so you can exercise it through an accidental good fortune that reduces to sheer dumb luck. The case that worries our opponent most sharply is not that of an untrustworthy trustee who fortunately abandons her plan to exploit your trust, but that of the trustworthy opponent who would never plan to exploit your trust but whose fidelity to your trust in her is nonetheless, in a key respect, accidental. To make this vivid, imagine that she plans fidelity, then is tempted to exploit your trust in a way that would compromise your relational autonomy, but then rights herself, resisting this temptation and remaining faithful. Your trust is not exploited, and – unless we embrace the absurd view that trustworthiness requires an insusceptibility to exploitative temptations – the trustee is worthy of your trust. But your relational autonomy appears to rest on a problematic species of fortune or luck. It makes sense to presume that what it rests on is the trustee's worthiness of your trust, but the question is how this grounding relation works. This element of happenstance is not like, say, the stuff that has to happen in your body for your psychology to realize autonomous agency. We're talking, not about the *causal* ground of your autonomy, but about what constitutes it. On

for reasons of her own but how she can get a 'reason of her own' through trust. Moreover, the threat of exploited trust does not target A's ability to act otherwise than she does.

our relational view, your autonomy is partly constituted by your reasonable trust in others. Your trust here is reasonable, but the apparently accidental element in the trust relation seems incompatible with relational autonomy. What might remove the appearance of happenstance?

What we need to reply to this challenge is a fuller understanding of the nature of trustworthiness. All three of the cases in which trust is exploited involve untrustworthy interlocutors who are concerned not with the autonomy of the agent, but rather with the advancement of ends of their own. They challenge their targets' deliberative perspectives without regard to the effects these interventions will have on the targets' responsiveness to their own practical reasons. We take this observation about *un*trustworthiness as a springboard for a more thorough elaboration of the nature of trustworthiness itself. A *trustworthy* partner in justificatory dialogue, we argue, takes a 'custodial' attitude toward the truster's autonomy, by manifesting concern for the truster's responsiveness to apparent evidence of untrustworthiness. Even non-exploiters can be untrustworthy, insofar as they are not sufficiently attentive to the quality of the trust relation between them and their interlocutors. But when they are thus attentive, it is this custodial element in trustworthiness that ensures that the truster does not realize relational autonomy through sheer good fortune or plain dumb luck.

III. Why relational autonomy requires trust

In this section we explain why a relational view of autonomy rests, in part, on a view of trust. Our relational view of autonomy rests on an understanding of trust that places equal emphasis on each side of the trust relation. When A trusts B, A relies on B in the way that risks betrayal of the trust, and B is worthy of A's trust only if B takes a kind of responsibility for A's responsiveness

to evidence that B is *un*worthy of that trust. In the case that provokes the challenge, in which A 'does his best' to be appropriately responsive to such evidence, B will count as worthy of A's trust only if B shows appropriate concern to put A in touch with such evidence. As we'll see, "appropriate concern" doesn't mean that B must address all available evidence of her untrustworthiness; it means only that she must address evidence that is especially salient or especially relevant to the case at hand. The core point is that B mustn't leave the burden to sort out such evidence entirely on A's shoulders. Our approach aims to vindicate the slogan: relational autonomy requires relational trustworthiness. Other accounts of trustworthiness fail to capture a key dimension of trustworthiness: a trustworthy agent is in the business of ensuring that those who trust her do not thereby lapse into heteronomy by having their trust exploited. To be trustworthy, in sum, you must be invested in the truster's autonomy in trusting you.

That last formula begins to articulate the deep connection between trust and autonomy that we aim to develop: a trustworthy agent is in a key respect invested in the autonomy of the trustee. That connection itself makes autonomy relational, but even without thematizing it, we can motivate a 'relational' view of autonomy by considering how an autonomous agent holds herself answerable to trusted critics and advisors. We now develop that dimension of the nature of autonomy, before returning to the challenge of reconciling autonomy with trust.

How exactly does autonomy require you to be answerable to critics and advisors? Since your judgment is fallible insofar as it may fail to track your reasons in a given case, a dogmatic insistence on the probity of your own judgment does not manifest autonomy, although you are in that case literally and emphatically treating your judgment as your law. Even if your self-relation does not degenerate into egotism or self-obsession, such an over-confident narrow-mindedness manifests a pathology of autonomy: the reflexive element in autonomy has become rigid and to

that extent unresponsive to your reasons. Since the role of judgment in autonomy is to manifest appropriate responsiveness to your reasons, this pathology of judgment amounts to a pathology of autonomy, not a way to govern yourself through judgment. The antidote to this rigidity lies in your capacity to tap into the supplementary and potentially critical perspectives that others have on your agency.

Holding yourself answerable to the perspectives of potential critics and advisors is crucial to autonomy because without it your judgment cannot count as genuinely responsive to your reasons. You need not, of course, be responsive to *all* critics and advisors. You need not be responsive to *any* in your local community – if members of your local community are not worthy of your trust. You need merely to be open to potential critics and advisors who are worthy of your trust – with an acknowledgement that a critic or advisor may be worthy of your trust even when you (mistakenly) judge that they are not. That acknowledgement simply registers the fallibility of your judgment on the question of who is worthy of your trust: you may fail to trust B by responding to what you mistakenly treat as evidence of B's untrustworthiness. Though we are arguing that trust is governed by responsiveness to evidence of untrustworthiness in the trustee, it does not follow that a failure to trust cannot manifest a mistake about such evidence. Your judgments of untrustworthiness are defeasible, and you may get further information that counters the initial evidence of untrustworthiness and returns you to the default stance of (equally defeasible) trust.

If trusted critics and advisors figure in the practice of autonomy by supplementing or correcting your judgment, it cannot be solely up to your judgment to tell you whom to trust. It is for this reason crucial that your capacity for reasonable trust work not through evidence of a critic's or advisor's trustworthiness, as assessed by your judgment, but through evidence of her

untrustworthiness. If there is (or were) evidence that this person is not trustworthy as a critic or advisor for you in these circumstances, then you will cease trusting her (or would not have trusted her). You thus govern yourself through your judgment without relying on your judgment to tell you whom to trust. Your judgment tells you only whom *not* to trust, through its responsiveness to evidence of untrustworthiness, not whom to trust. If there is no evidence of B's untrustworthiness, it may be reasonable to trust B even if B's criticism or advice runs contrary to the verdicts of your own judgment.

For autonomy to work like this, obviously, the fact that B's criticism or advice runs contrary to your judgment cannot *eo ipso* generate evidence of B's untrustworthiness. And that seems right, since the mere fact that someone disagrees with you does not give you a reason to distrust them. In certain circumstances, such disagreement might be worrisome: for example, when you have additional evidence that your interlocutor is not taking the problem at hand seriously, or is not giving due regard to your perspective, or seems to be motivated by an agenda of their own rather than by the question of what you have reason to do or believe. But in an ordinary case, where you have no such reason to distrust another, taking seriously the alternative view they offer manifests an appropriate sense of one's own fallibility and concern for the quality of one's own judgment. This is simply to reaffirm the point from which we began, that self-governance does not simply reduce to being governed by one's own judgment. The role played by one's own judgment is more complex than that formulation would suggest. Governing yourself requires a default presumption of trustworthiness in others. This presumption is monitored by your judgment insofar as the presumption is defeasible and responsive to evidence of untrustworthiness in others.

The question, then, is how you hold yourself appropriately answerable to external perspectives. We have assumed that such answerability includes a disposition to trust, but we should pause to consider why that is so. Why should taking a critical perspective seriously, in the way now at issue, involve a disposition to trust the deliverances of that perspective?[5] Can't you merely expect that the critic will prove usefully adept at reminding you of considerations that you have for the moment forgotten but that you could have weighed correctly had you happened to think of them yourself? Of course, you can; in that case, you rely on the critic but do not trust them in any substantial respect – any more than you trust your alarm clock when you rely on it to jog your memory by 'reminding' you of an appointment. The question is whether every case of being guided by a critic or advisor must look like that. Say you rule your fiefdom with exceptionless authority, never letting anyone supplement or correct your verdicts, but you sometimes lose track of which new law you've been planning to implement, so you hire an amanuensis to help you keep track of the plans through your disposition to let him correct you about the intentions of your former self. Your willingness to let yourself be thus corrected does not dissipate the appearance of autonomy-undermining overconfidence (if not of egotism or self-obsession). The problem is that you are not disposed to let your 'critic' correct you from *his* perspective – to let him offer any challenge to your judgment. A disposition to do that is the only thing that could help dissipate the appearance of overconfidence.

An emphasis on answerability in relational autonomy thus gives way to an emphasis on trust and trustworthiness. What then is it to be trustworthy? The next section explains why

[5] To trust the deliverances of another perspective is not necessarily to treat them as correct, but rather to treat them as worth entertaining in a way that you understand might change how you see and respond to the situation at hand. It is to be open to the rational influence of another perspective on your own, even though at first blush it might strike you as mistaken. Holding yourself answerable to a trustworthy other includes subjecting your own contrary judgments to critical scrutiny and self-questioning that might not otherwise have seemed warranted.

trustworthiness is not a mere disposition to do what you are trusted to do but depends on an understanding of the point of the trust. Section IV explains why trustworthiness includes a disposition to take responsibility for the trust in light of that understanding. Since, as we've just seen, an aspect of the truster's relational autonomy depends on trust, we conclude, in Section V, that the trustworthy trustee's disposition to take responsibility for the truster's trust ensures that relational autonomy is not undermined by the possibility that trust is exploited.

IV. How trustworthiness requires concern for the truster's autonomy

The first thing to say about trustworthiness is that it involves more than mere reliability. Legend has it that Kant was reliable in taking his afternoon walks: at a certain time each afternoon he reliably passed in front of a given neighbor's house. That neighbor could have relied on Kant to set her timepiece; Kant was reliable in that respect. Though perfectly reliable, Kant was not thereby worthy of the neighbor's trust. Why not? One plausible thought is that Kant's reliability does not amount to trustworthiness because Kant does not have the right attitude toward this person, or toward others in a position to rely on him in this respect. Simply because he is oblivious toward her reliance on him, he lacks any relevant species of goodwill, or concern, or commitment toward her or toward what she needs from him. But what exactly is this attitude? It turns out to be far from easy to give a general account of it. From this general observation, Karen Jones (2012) draws a conclusion that we believe is on the right track: what distinguishes trustworthiness from mere reliability lies not in the nature of the trustee's attitude toward the truster but in the trustee's attitude toward the fact that the truster is relying on the trustee.

Jones (2012) distinguishes trustworthiness from mere reliability not in terms of the trustee's attitude but in terms of the trustee's responsiveness to reasons: if trustworthy, you treat the truster's trust in you as a reason to do what you are trusted to do. That account contains an insight but is falsified by the form that trustworthiness takes in the interpersonal and intrapersonal normativity of commitment. If trustworthy, you will not regard yourself as having a reason to do what you are trusted to do if performing that action would defeat the understood point of your commitment. Say A trusts you to water her patio plants while she's out of town, but it unexpectedly rains enough that giving them more water would harm them. So, you don't water them. Does that make you untrustworthy?

On Jones's account, you would be trustworthy only if you treat, or are disposed to treat, A's trust in you to water her plants as a reason to water her plants. When you see that the plants already have plenty of water and that it would harm them if you added more, however, you don't have to deliberate, weighing a reason grounded in A's trust against other reasons, as you would if confronted instead by a lack of available water. Fleshing out this alternative scenario, imagine you discover that A has left you no supply of water: the spigot you'd expected to use is dry, you can't find any nearby source of water, and it would be a huge inconvenience to fetch water from elsewhere. You decide that the plants can survive till A returns, so you don't water them. This would not manifest untrustworthiness. If A complains that it does, add details to make the expectations informing A's trust more unreasonable. At some point, these details will ensure that your failure to water does not reveal you as unworthy of A's trust in you to water.

The original scenario is not at all like that: in the original case, you see immediately not only that there is no need to water but that watering would be positively harmful. We might assume that if A could see what you see, she'd agree. But we're imagining that she does not have

this information, and that without it she still trustingly expects you to water her plants. You've disappointed that trust in the sense that you have not done what she trusts you to do, but you have not in the slightest betrayed her trust. We can imagine a case in which there's a dispute about what 'living up to her trust' requires of you: A feels betrayed, but you believe that her feelings of betrayal reveal a confusion about the situation. Perhaps in the present case it's a confusion about the biology of plants: perhaps A believes that it is impossible to give a plant too much water. Or perhaps A never conceptualized her trust as about the plants themselves but only about your willingness to do what she believes she needs from you. In prioritizing the health of the plants, you have betrayed her trust in you to sacrifice the plants in order to show your fidelity to her. In this elaborated case, the confusion lies not simply on A's side but between you.

If there is room for such confusion between you, then there must be something worth calling an understanding between you. Sometimes part of this understanding will be made explicit. You might, for example, ask: Is A trusting you not to overwater? Is she trusting you to show an attitude of care toward her – perhaps, in unexpected circumstances, regardless of the plants? You might feel uncertain what exactly A trusting you to do, and you may need to resolve that uncertainty by working from a more precise description of that act. Perhaps in interpreting A's trust as prioritizing the health of the plants you reveal your untrustworthiness on a different matter: whether she can trust you to kill living things if that's what it takes to give her what she needs from you. In the latter case there will likewise be resources to develop a basis for the sort of confusion we're probing. What if you believe her confused about what she needs from you? Say you recognize one of A's plants as, unbeknownst to her, a rare orchid worth thousands of dollars. Even if you interpret her trust in the way she wants you to, as not addressing the health of the plants but instead as some sort of loyalty test, is it consistent with that loyalty to destroy

this plant simply because she does not know how much this plant is worth? One might assume that, in trusting you, she owes it to you to tell you what she expects from you, but human nature is too complex and conflicted for that. The understanding that informs trust can only be so explicit about what the trustee is expected to do. What it leaves implicit is not thereby excluded from the content of the trust.

Jones takes trustworthiness to lie in the trustee's responsiveness to a reason given by trust, but she misses how trustworthiness lies in fidelity to an understanding. The trustee's responsiveness to the trust-given reasons that Jones emphasizes requires that the trustee understand what is relevantly at stake for the truster, an understanding that the truster's expectations of the trustee may violate. In this respect, trust mediated by a promise or some other form of agreement is paradigmatic. Annette Baier (1994, 137) has influentially depicted promising as a "peculiar" instance of trust. From our perspective, however, the fiduciary agreement at the core of a promise reveals how trustworthiness serves its principal normative function. Insofar as you are trustworthy, you are governed by an interpersonal understanding.

The next step in appreciating the role of trust and trustworthiness in autonomy lies in seeing how a trustworthy agent is governed by an understanding shared with the truster. We have seen that it is not enough simply to do what you're trusted to do. A trustworthy agent can be counted on to do it only if doing it serves the understood point of the trust. As in the earlier cases that we considered, the important thing is not merely to get the job done but to get it done in a way that vindicates the truster's trust. And we may now say the same with our new emphasis on understanding. Even if doing what she is trusted to do serves the understood point of the trust, a trustworthy agent does not rest with doing what she is trusted to do if doing it would, due to a misunderstanding of the situation, leave the truster feeling betrayed. A trustworthy agent aims to

serve the understood point of the trust in part by serving the truster's autonomy, which includes the truster's felt sense of his own reasonability in trusting her. Since trust is reasonable only when there is no compelling evidence that the trustee is unworthy of it, a trustworthy agent will take some responsibility for the trust by being concerned to dispel whatever evidence of her untrustworthiness may be available. She'll think that such evidence is misleading, and if she knows that the truster is in position to be moved by it, she owes it to him to explain why the evidence is misleading. This obligation can be met without much fanfare; the explanation need not be very full. The point is that it is incompatible with being trustworthy to let such evidence move the truster, with no counterweight, into regretting the trust. That is incompatible with trustworthiness because it shows a lack of concern for the truster's relational autonomy.[6]

V. Autonomy as aptly managing the risks of trust

Taking stock, we can summarize the foregoing by saying that a relational view of autonomy is a view of how you manage the risks of trust. On one simple version, you manage the risks of trust by apportioning your trust to available evidence of others' trustworthiness. The problem for this version, as we saw, is that it makes it hard to see how what it calls 'trust' is really trust, since it views your reliance as guided by your grasp of evidence. Even if you acknowledge that your evidence of B's trustworthiness is inconclusive – as evidence of trustworthiness typically is – a stance of seeking or needing evidence of B's trustworthiness betrays an attitude toward B that runs contrary to the spirit of trust in B. But if we codify this observation by allowing that you may reasonably trust even when there is no evidence of trustworthiness, we need to explain how

[6] For a conception of trust that emphasizes such shared understanding, see Hinchman 2017, 2021.

you could thereby count as managing the risks of trust. We have explained this by specifying that your trust is guided by your counterfactual sensitivity to evidence of untrustworthiness: if there is (or were) evidence of untrustworthiness, you will stop trusting (or would not have trusted). That helps by ensuring that your reliance will count as genuine trust, but it hurts by leaving open the possibility that your reasonable trust will be exploited by the trustee in a way that makes your agency heteronomous. How is that possibility compatible with idea that you yourself manage the risks of your trust? Even if your trust is not actually exploited, how can you be autonomous if you leave it up to the trustee to manage the risk you run that it will be?

We can thus refine the challenge to our view of relational autonomy, codified by the fourth type of case described in Section II. The threat to your autonomy created by the possibility that your reasonable trust will be exploited lies in what it shows about that presumption of reasonability. The presumption is grounded in the fact that there is no evidence available to you that the trustee is unworthy of your trust. But the possibility of exploitation entails the possibility that such evidence will become available. And here's the rub: if your trust is exploited, in the third way that we described in Section II, the exploitation will prevent you from discovering this evidence of untrustworthiness until it is too late. That isn't like the possibility that the evidence is not now available to you through sheer happenstance but will become available later. In that case, no one is responsible for having prevented you from considering that evidence. In the case we're considering, that's exactly what the trustee does, or might do: prevent you from considering evidence that will or would later lead you to regret your trust. Framing the challenge in these evidential terms helps us resolve one lingering issue: which bits of evidence of her untrustworthiness would a trustworthy trustee help you appreciate as misleading? The answer is: the bits that realistically might become available to you and lead you to regret your trust.

How, then, does the trustworthiness of the trustee figure in what is intuitively *the truster's own* management of the risks of trusting? Say A trusts B to φ. If we view B's trustworthiness as merely a reliable disposition to φ, then this cannot be an instance of A's own risk management, since whether B has this disposition or not is at best up to A only in a causal sense (say, in therapeutic trust, wherein your trust causes the trustee to be worthy of it). If we complicate matters by adopting Karen Jones's more complex view of trustworthiness – saying that B must also be disposed to treat the fact that A is relying on him to φ as a reason to φ – we avoid some of the problem cases but can't avoid others, since whether B has this more complex disposition is not up to A either (except perhaps in a therapeutic case). So, how can we get A 'in on' this status – that is, 'in on' B's worthiness of her trust – in a way that lets us regard A, in her own capacity as an autonomous agent, as thereby managing the risks of trust?

To solve this problem, we proposed adding further conditions to Jones's treatment of trustworthiness. To be a trustworthy influence on A, B must not only treat A's reliance as a reason to do what A is relying on B to do, and in a way that does justice to a shared grasp of the point of the trust, but must also be responsive to evidence of her own untrustworthiness that is not currently available to A. Our idea, in sum, is that a trustworthy B aims to prevent A's trust from being 'Gettierized' (adapting that epistemic concept to this new context). A's trust would be Gettierized if (i) A does his best to manage the risks of trust with available evidence, and (ii) A's trust is successful in the sense that B does not betray or exploit it, but (i) does not at all contribute to the explanation of (ii). In an easy case in which A's trust is not Gettierized, (i) contributes to the explanation of (ii) insofar as A trusted B because there was no evidence that B was likely to betray her trust, though A would not have trusted B if there had been such evidence. But in the case we're considering, there is evidence that B is likely to betray A's trust,

and A trusts B – successfully, as it happens, but the success is therefore not to A's credit.[7] What we need, then, is for B to be disposed to ensure that (i) does contribute to (ii). How can B do that? As we've seen, a trustworthy B will manage the risks of trust on A's behalf.

What, in general, is it for A to manage the risks of trust? To manage the risks of trust, A must be disposed to cease trusting in response to available evidence of B's untrustworthiness. In the case that worries us, there is evidence of B's untrustworthiness, but it is not available to A. So here is how B can manage the risks of trust *for* A: in some appropriately loose sense, B *gives* A that evidence, then helps A to see that it is misleading. This is evidence that would have led A not to trust B, if it had been available to A. And this evidence will lead A to regret the trust if it becomes available only later. Even if the trust is successful insofar as B does what A trusts B to do, A may regret the trust as unreasonable from A's perspective at the time: there was evidence of untrustworthiness that ought to have led A not to trust B. But A would have reversed course and trusted B if A had come to see that the evidence of untrustworthiness was misleading. And A will not regret the trust on the basis of misleading evidence of B's untrustworthiness. So now B takes A through that dialectic – or at least manifests a disposition to do so, should the circumstances demand it of B. What's crucial, in sum, is that B does not try to exploit the fact that this evidence is not available to A. We thus, in effect, build an anti-exploitative and therefore autonomy-promoting disposition into our account of B's trustworthiness. In the epistemology-influenced shorthand that we've adopted, it is a disposition to do what it would take to ensure that A's trust is not Gettierized in this way.

[7] This point could also be put in terms of an analogue of 'safety': for all A's efforts at managing the risks of trust, A could easily have been betrayed.

It does not vindicate trust if the trustee does what she is trusted to do with no regard for evidence of her untrustworthiness. There will sometimes be little she can do to address the problem, but showing that she cares about what such evidence implies about the reasonability of the truster's trust itself helps to make that trust more reasonable. If she manifestly does not care at all, that will tend to make it look like the trust was vindicated through dumb luck or sheer good fortune. The point, again, is not that she is untrustworthy if she does not counter every piece of evidence of untrustworthiness that may emerge. Some evidence of untrustworthiness may derive from her past conduct, some from her present conduct, and some from the conduct of others who relevantly resemble her. In each case, there is only so much that she can do to show that the evidence is misleading about her worthiness of trust in the present instance. The point is that she must show concern to counter such evidence, not that she must succeed in countering it.

Trust is indeed at odds with autonomy when success (the trustee does what she is trusted to do) merely accompanies the truster's exercise of relevant competence (the trust is reasonable). But trust is not only compatible with but manifests autonomy when the truster succeeds *by* exercising that competence. In epistemology, that 'by' marks a relation for which Ernest Sosa coined the term 'aptness.'[8] Applying it here, we see that trust is compatible with autonomy if and only if it is apt: not just successful *and* reasonable but successful *because* reasonable. What links the success with the reasonability of trust lies in the relational nature of trustworthiness: that a trustworthy agent does what she is trusted to do in a way designed to vindicate the reasonability of the trust. The relational view thus explains how your reasonable trust serves your relational autonomy: you manage the risks of your trust through your receptivity to the trustee's worthiness

[8] For a recent deployment of the concept, see Sosa 2021, 18-25.

of it – that is, through your trusting but prudent responsiveness to evidence that the trustee might not manage those risks on your behalf.

References

Abramson, Kate
2014 "Turning Up the Lights on Gaslighting," *Philosophical Perspectives* 28, 1-30.

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

Bouzar, Dounia
2017 "A Novel Motivation-Based Conceptual Framework for Disengagement and De-Radicalization Programs," *Sociology and Anthropology* 5:8, 600-602.

Ebels-Duggan, Kyla
2015 "Autonomy as Intellectual Virtue," in Harry Brighouse and Michael MacPherson (eds), *The Aims of Higher Education: Problems of Morality and Justice* (Chicago: University of Chicago Press).

Frankfurt, Harry
1969 "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 66:23, 829–839.

Hinchman, Edward
2017 "On the Risks of Resting Assured: An Assurance Theory of Trust," in Paul Faulkner and Tom Simpson (eds), *The Philosophy of Trust* (Oxford: Oxford University Press).
2021 "Disappointed Yet Unbetrayed: A New Three-Place Analysis of Trust," in Kevin Vallier and Michael Weber (eds), *Social Trust* (New York: Routledge).
2022 "The Role of Assurance in Judgment and Memory," in Sanford Goldberg and Stephen Wright (eds.), *Memory and Testimony: New Essays in Epistemology* (Oxford: Oxford University Press).

Jones, Karen
2012 "Trustworthiness," *Ethics* 123:1, 61-85.

Perez de Calleja, Mirja
2019 "Autonomy and Indoctrination: Why We Need an Emotional Condition for Autonomous Reasoning and Reflective Endorsement," *Social Philosophy and Policy* 36:1, 192-210.

Sosa, Ernest
2021 *Epistemic Explanations* (Oxford: Oxford University Press).

Westlund, Andrea
2003 "Selflessness and Responsibility for Self: Is Deference Compatible With Autonomy?" *Philosophical Review* 112:4, 483-523.
2009 "Rethinking Relational Autonomy," *Hypatia* 24:4, 26-49.
2022 "Agency and Autonomy," in Luca Ferrero (ed.), *The Routledge Handbook of Philosophy of Agency* (New York: Routledge).