Demoralising Trust*

Abstract

What do we expect of those whom we trust? Some argue that when we trust we are confident the trusted will act on moral motivations. But often we trust without appraising the trusted's moral qualities, and sometimes trust expects more than morality demands. I argue for a non-moral commitments account: when we trust a person we expect they will be motivated to act a certain way by a commitment that we ascribe to them. My alternative accommodates an expanded typology of trust's vulnerabilities, including tragic disappointments that are as painful as betrayal, but without the recompense of moral complaint.

1 Introduction

Philosophers sometimes moralise trust. It is not uncommon for theories of trust to assume that if we are let down by those we trust we are thereby victims of betrayal, a distinctive kind of wrong that is made possible when someone puts their trust in another.[1] It is also sometimes held that when we trust a person we expect they will do what we trust them to because of moral motivations or reasons. On this view, betrayed trust indicates a moral failure.[2] But this picture of trust is misleading. Sometimes we trust others without depending on their moral qualities, such as when we trust colleagues to do a good job. Sometimes we trust others to do more than morality demands, particularly in personal and intimate forms of trust. And sometimes our trust can fail to be satisfied, despite the fact that the person trusted has committed no moral wrong. In this paper I make the case against accounts of trust that moralise its expectations, and in favour of an alternative commitment-based account. I argue that when we trust we expect that the person trusted will act as we wish them to because they have a commitment to something – an action, goal, value, project, other people, etc. – that motivates them to act this way.

The scope of my argument is limited. I will not discuss trust in political and social institutions, self-trust, or therapeutic trust i.e. the communication of dependency to another person in the hope that that person will cultivate trustworthiness such that I might trust them in future.[3]

I am concerned only with interpersonal trust, and take trust in intimate relationships as my primary, but not exclusive, focus. I focus on such cases because they most clearly bring into relief the non-moral risks integral to trust. I thus approach this from a different angle to others who have taken issue with moralised accounts. Some have suggested that accounts of trust that look to the moral qualities of the trusted do so because they focus narrowly on trust in personal relationships, and neglect the less sentimental, more prosaic forms of interpersonal trust that involve no moral judgements of the trusted.[4] Conversely, I argue that it is precisely the most intimate forms of trust that are insufficiently understood by moralized accounts.[5]

My complaint about moralising theories of trust also differs from extant arguments against theories of trust that deny cases of immoral trust, that is, immoral people trusting one another to do immoral things.[6] What I call moral-motivation theories moralise the reasons that ground the truster's confidence in the trusted, maintaining that when X trusts Y to Φ, X's willingness to rely on Y's future Φ-ing is secured by X's optimism that Y has moral motivations that will lead Y to Φ, even where Φ is not an all-things-considered moral act. In this respect, my argument against moral-motivation theories shows that the category of theories that illegitimately moralise trust is broader than previously understood.

I first outline and identify the problems that arise for moral-motivation theories of trust (sections 2 and 3). My alternative account begins by considering the variety of disappointments we make ourselves vulnerable to when trusting (section 4). Both the shortcomings of moral-motivation theories and the variety of risks trust exposes us to – including a number of negative outcomes usually absent from trust theories – give us reason to prefer what I call a commitments theory of trust (sections 5 and 6). I maintain that when we trust a person we depend on them to Φ and we expect they will Φ because they will be motivated to do so by a commitment we ascribe to them. My alternative at once lowers trusters' moral expectations, raises the bar a person must clear in order to be trustworthy, and acknowledges the tragic possibility of trusting relationships that falter without culpability. When we trust we sometimes expect more than moral goodness,

and this exposes us, I argue, to disappointments that can be as painful as betrayal, yet without the recompense of moral complaint.

2 Moral motivations

Moral-motivation theories are a subset of motivation-based theories of trust. Motivation-based theories usually begin by distinguishing their accounts of trust from what Karen Jones calls 'risk-assessment accounts'.[7] Risk-assessment accounts understand trust as staking something important on the actions of others, a gamble that is based on an assessment of the potential benefit relative to the likelihood that a person will act as we want them to. We trust in this sense because we are persuaded of the high likelihood that a person will act as we hope, or because the potential payoff is worth the risk they will not, or because of an adequate combination of predictability and payoff.[8]

What risk-assessment theories call trust could also be called reliance. But this, the motivation-based theorist will argue, is a problem, for trusting is not the same as relying. When we rely on something we make predictions about its future performance. Sometimes this is a judgement about functionality, such as when I rely on my car to start in order to get to work; sometimes it is an assessment of regularity, as when Kant's neighbours set their clocks by his precise and consistent daily routine.[9] We can rely on inanimate objects and on persons, and in both cases reliance involves the judgement that the object of my reliance is capable of doing whatever it is I rely on it to do. Trust too, according to the motivation-based account, involves a judgement of competence – I would not trust a well-meaning but incompetent babysitter – but trust differs from reliance insofar as it also involves an attitude about the motivations of the person trusted. The car on which I rely has no motivations; I need not think anything of Kant's motivations to rely on him to tell the time. But if instead I trust Kant to be on time for, say, a mutual appointment, my confidence in him is based on a judgement about his reasons or motivations.

Not all motivations are eligible for trust's expectations. Fear can make behaviour predictable, but I expect fear to motivate someone when I, for example, manipulate them through

threat of force, but not when I trust them.[10] What kind of motivation is ascribed to the trusted? One option, first suggested by Annette Baier and developed by Jones, is goodwill towards the trusted.[11] On Jones' goodwill-based account, trust is an attitude that combines the affective – optimism about the trusted's goodwill – and the cognitive – an expectation, grounded in that optimism, that the trusted will respond favourably to my dependence on them. On this account when I e.g. trust a friend to babysit I expect their goodwill towards me to provide them with motivation enough to care for my child.[12]

This optimism about another's goodwill requires an attitude towards the trusted that supports this optimism. One candidate attitude is belief that the person considers me a friend; if I am confident that they do consider me a friend, then I am more likely to be optimistic about their goodwill, insofar as "goodwill" is understood as the kindness we extend towards friends. But Jones rules this out, noting that goodwill based on friendship is too restrictive for trust.[13]. Though Jones does not expand on why she thinks this too restrictive, the suggestion appears to be that not everyone we trust is a friend. I might, for instance, trust my doctor without befriending her. We could also observe that sometimes fulfilling trust requires that we do precisely what kindness inclines us not to do. A doctor overly concerned with kindness might be reluctant to be honest about a patient's ill health for fear of distressing them, but patients trust their doctors to be forthcoming with their diagnosis.[14]

Instead, Jones maintains that trust's optimism about the goodwill of another depends on a supporting ascription of at least one of a number of relevant character traits, examples of which include benevolence, integrity, honesty, and moral decency.[15] Our optimism about the goodwill of those we trust is thus made possible, on this view, by assessments of their moral qualities, such that we must think sufficiently well of a person's character if we are to trust them. This is not to say that Jones thinks we need think especially highly of the trusted's moral qualities. Sometimes, when the risk of disappointment is low, we need only think that the trusted is minimally moral, without malice or ill will; Jones suggests, for example, that this is the level of confidence in the

character of others that we need to trust that we will be safe in public spaces.[16] By contrast, where we are pessimistic about goodwill we suspect people will be malicious and foresee a significant risk of wilful harm from others.[17] Thus for Jones though trust need not expect saintliness it does require that we think a person moral enough to motivate them either to extend goodwill toward us in the relevant domain, or at least to not bear us ill will.

Not everyone has been convinced that goodwill correctly distinguishes trust from reliance. McLeod, for instance, complains that 'goodwill' is too vague a term to rule out its colloquial usage as 'kindly feeling', and that it thereby fails to correctly track the moral concern for others that we ascribe to those we trust.[18] McLeod proposes instead that trust ascribes moral integrity, where moral integrity means doing what one thinks is the morally right thing to do.[19] And because we do not trust those we believe to be radically morally mistaken – those whose moral integrity would lead them to do what we consider to be deplorable – McLeod suggests that we also expect 'that what the trusted person stands for morally speaking is similar enough to what we stand for…that we can count on that person to do what we trust her to do'.[20] Though McLeod's account is narrower in its focus than Jones', identifying a single moral quality ascribed by trust, both share the view that if we are to trust a person we must think well enough of their moral character and expect that moral character to motivate a particular action.

McLeod and (early) Jones propose what I call moral-motivation theories, which maintain that when I trust someone to Φ, I depend on them to Φ and I am confident that their motivations to act will be sufficiently morally good to ensure that they will indeed Φ. On this picture, trust requires the truster make an at least implicit moral judgement of the motivations of the trusted. This means that these theories moralise the psychology of trust, but do so in a way that is different to what might otherwise be understood by the term 'moralised theory of trust'. One might understand a moral theory of trust to be a theory that maintains that any legitimate case of trust requires that the truster expects moral action from the trustee. As many have observed, it is difficult to defend such a view, for it would have to deny, implausibly, the possibility of trust

among the immoral to do immoral things (networks of political corruption would provide examples of this kind of trust).[21] But moral-motivation theories can allow trusters to expect immoral action from trustees.

There are two ways a truster could meet the description of a moral-motivation theory without expecting moral action. First, it could turn out that the truster is mistaken about the goodness of the action they expect (Φ). In such cases the truster ascribes moral motivations that secure their confidence in Φ, which they think is a moral act, but they are objectively wrong about the moral worth of Φ. Second, a truster could ascribe moral motivations that secure their confidence in Φ, even though they do not think Φ a moral act. A person could, for instance, trust another to conceal a wrongdoing, and accept that doing this would be immoral, yet nonetheless depend on that person to do this because they think the trusted is a person of sufficient fidelity to conceal the wrong. Such trust still meets the moralising condition of moral-motivations theories because it relies on a judgement of moral character independent of the attitude held toward the action that is entrusted. In short, whereas conventionally moral theories of trust stipulate that trust expects objectively moral action, moral-motivation theories stipulate that trust expects subjectively (according to the truster) moral motivations.

Some other theorists of trust have taken a different route to reach the same moral-motivations conclusion. Reactive-attitude accounts of trust also begin by distinguishing trust from reliance, but do so by considering what attitudes are appropriate to a failure of trust rather than a failure of reliance.[22] A range of responses are fitting when my car won't start, but betrayal is not of them. Betrayal is, however, appropriate to failed trust, and reactive attitudes accounts will leverage this observation to explain what distinguishes trust from reliance: when we trust a person we count on them to Φ in such a way that warrants feeling betrayed if the person inexcusably fails to Φ. Building a reactive attitude into our theory of trust also has the advantage of solving the problem of the confidence trickster.[23] The problem is that though the trickster expects their good-natured victim to be responsive to the trickster's dependence on them – and thus meets the

descriptions of trust provided by some theories – this is a case of exploitation, not trust. Stipulating that trust warrants feeling betrayed is said to solve this insofar as a disappointed confidence trickster could not legitimately feel betrayed.

Not all reactive attitude accounts moralise trust, because not all explicitly treat betrayal as a moral emotion.[24] But some do.[25] We might ask: why is betrayal appropriate to failed trust, but not to a failed con? More than one answer is available here. We could invoke the same considerations of moral character proposed by Jones or McLeod. McLeod, for instance, maintains that feelings of betrayal are *prima facie* appropriate when trust is disappointed because, since trusters expect moral integrity, a failure to live up to this trust will be perceived by the truster as a *prima facie* moral failure.[26] Alternatively we might adopt an obligation-responsiveness account. Phil Nickel, for instance, has argued that a necessary condition for trust is that the truster believes that the trusted has a moral obligation to Φ, and that the trusted will be motivated by this obligation.[27] As for the source of this obligation, we might follow Cogley in suggesting that trusters take themselves to be entitled to Φ in virtue of their relationship to the trustee, or Cohen and Dienhart in suggesting that trusters take themselves to be entitled to Φ in virtue of the very fact that they entrust Φ with the trustee.[28] These views share the claim that trusters can be warranted in feeling betrayed where con-artists cannot because of a moralised feature of trust's expectations: if I trust a person to Φ, I believe they are morally obligated to Φ, and am entitled to react appropriately if this obligation is not discharged.[29] They also share the view that obligation-ascription can explain the confidence we have in others when we trust them; trust's confidence, on this view, is supported by the belief that the trusted is responsive to their obligations.

I submit, then, that the literature on trust gives us a variety of accounts that take one of two different routes to the same moral-motivations principle. (Early) Jones and McLeod tell us that trusters are optimistic that the trusted will Φ because they are confident in the trusted's moral qualities, and believe that these moral qualities will motivate the trusted to Φ. Reactive attitudes theories tell us that trust is distinguished by the appropriateness of feeling betrayal, and obligation-

ascription theories moralise this focus on betrayal by stipulating that trusters ascribe a moral obligation to the trusted which, the truster believes, both motivates the trusted and warrants feeling betrayed if the they fail to act as expected. Thus either via an account of the hopes of trust, or an account of the attitudes appropriate to unfulfilled trust, these theories reach the distinctive principle of a moral-motivation theory: that when X trusts Y to Φ, X's willingness to rely on Y's future Φ-ing is secured by X's optimism that Y has moral motivations that will lead Y to Φ.

3 Problems with moral motivations

The problem for these theories is that the ascription of moral motivations is neither necessary nor sufficient for trust. That this ascription is not necessary is a strong reason to reject moral-motivation theories. That it is insufficient is less of a problem – these theories make no claim to an exhaustive account of trust – but the insufficiency invites discussion of what is missing. I take up that invitation in sections 4-6.

Amy Mullin has already suggested a number of examples of trust that do not require confidence in the moral qualities of the trusted. Perhaps I have a regular chess partner whom I trust will be a good sport when playing.[30] This involves expecting, for instance, that she is a competent player, and that she will resist the temptation to play carelessly or lazily. I can also trust an adversary to be pleasant towards me in front of mutual acquaintances out of respect for norms of civility (ibid.). Trust in these cases need not ascribe moral qualities to the trusted. *Pace* McLeod, my confidence in the honourable play of my chess opponent need not rely on a judgement that they will act with moral integrity. Similarly, I need not ascribe moral integrity to the adversary I expect to abide by civil norms because those civil norms need not be morally grounded. Though Jones allows for a broader variety of character traits that could ground the confidence of the truster, she nonetheless maintains that trust requires optimism about the goodwill of the trusted, supported by the relevant positive appraisal of their character. In Mullin's examples, the truster need ascribe neither goodwill towards me nor moral qualities motivating their actions.

Obligation-responsiveness accounts are better placed to accommodate Mullin's examples. Perhaps trust in my chess partner requires that I believe they are subject to obligations of fair play, and certainly I will think they are subject to the norms of the game. Similarly, it appears that the trust in my adversary to abide by civil norms requires I am confident that my nemesis will recognise and respond to the obligations generated by those norms. Nonetheless, obligation-responsiveness theories meet problems of their own when we consider cases where we trust someone to do things that are not obligatory. In many lines of work colleagues must trust each other to do their job, but when we are lucky we can trust our colleagues to do a *good* job. This involves expecting our colleague will do more than what duty demands, perhaps because we have confidence in their commitment to making the most of a particular project. If I trust a conference co-organiser to do a good job, it could be because I am confident that they see value in running a good conference, and that this will motivate them to go beyond what is minimally demanded of them (e.g. taking responsibility for communicating travel and accommodation details to all delegates, giving significant thought to speaker introductions, or showing sincere interest and enthusiasm in work presented by participants). By contrast, I may withdraw my trust if I come to think my co-organiser is a jobsworth, someone chronically averse to the supererogatory. The conference-organisers case is a problem for obligation-responsiveness theories because it brings into relief the way that trust can sometimes expect more than just what the trusted is obligated to do.

The obligation-responsiveness theorist may respond that even if I trust my colleague to go above and beyond, I still think they are subject to the obligations of running the conference (say, ensuring sufficient diversity of speakers), and I will not trust them if I do not believe they are responsive to these obligations. In this respect, the expectation of obligation-responsiveness is still necessary to trust, even if trust expects more than what is required. Nonetheless, as the trust-in-colleagues example shows, sometimes the confidence we have in those we trust is secured by more than just the trusted's responsiveness to obligations. Sometimes trust's confidence in another is secured by non-moral features of the trusted.

One further reason to think that trust sometimes expects more than moral qualities is that it is possible for our trust to go unfulfilled despite the trusted meeting all relevant moral expectations. Take the following case. Two close friends – I will call them Andrea and Ben – grew up together in the same town, where they still live, and share many of the same interests and values. They confide in one another with very personal and sensitive matters, and they trust each other to support them through difficulty and join them in celebrating good news. One day Andrea leaves to spend a year abroad. Ben stays home. The two stay in contact in an effort to sustain their friendship over distance. But when Andrea returns, she finds that Ben's character has changed, not so much as to make Andrea suspect something unusual has happened to Ben, but enough for her to think he is a different person to who he was before she left. His tastes have changed, he no longer finds Andrea's jokes funny, and he has lost interest in the hobbies they used to share. Ben has also made new friends, whom Andrea thinks are nice enough but do not have a lot in common with her. In the months that follow Ben is less inclined to accept Andrea's invitations. His shared interests with Andrea are vanishing, and his warm feelings for her as a friend have cooled. Andrea finds that conversations by phone, text, or email are increasingly prompted and led by her, and she has a growing feeling that Ben is no longer interested in talking with her. Eventually Ben turns down all of Andrea's invitations to spend time together, a long time passes without any communication between them, and the friendship is effectively over.

Ben and Andrea's friendship, while it lasts, includes forms of trust that are most commonly found in intimate relationships. Trusting relationships with loved ones involve trusting that the other will prioritise us enough to come to our aid when we need it, take pleasure in our happiness enough to celebrate our good news, keep our confidence, act with sensitivity regarding whatever it is that pains us emotionally, and be charitable with us, less quick to judge negatively what we say and do. Andrea trusted Ben to prioritise her enough to accept at least some of her invitations to spend time with her, to take interest in her and her life, and to show willingness to talk as friends. Ben disappoints this trust. But – crucially for my purposes in this paper – such trust can be

disappointed despite the fact that the truster's expectations about the moral qualities of the trusted are met. In the fading friendship case, Ben disappoints Andrea's trust, but he does not fail to meet any moral expectations her trust might have of him and his motivations.

This last claim needs some defence. If Ben has shown any failure of goodwill, integrity, benevolence, honesty, moral decency, or obligation-responsiveness, then this is not a counterexample to moral-motivation theories. Has Ben shown himself lacking moral quality? I argue not. Ben's values and interests have changed while Andrea is away, and continue to change when she returns, in a way that makes their friendship difficult to sustain. The differences between them become significant enough to render friendship possible only if forced, performed without sincerity, or maintained out of duty rather than out of interest. The moral thing for Ben to do would not be to fulfil Andrea's trust in him to continue the friendship, because to do so would be dishonest and patronising. The moral thing to do would instead be to withdraw from the friendship, despite the fact this will let Andrea down, provided he does so as kindly and sensitively as possible. (If we think I am letting Ben off the hook too easily, then we can build into the case that he has tried everything he can to sustain and later to rebuild the friendship, and that he has explained his change of heart to Andrea with sympathy and honesty, all out of concern for the hurt it will cause Andrea when they drift apart.)

One might respond that Ben has not failed to act with good intentions, but he has nonetheless failed to meet his obligations as Andrea's friend. This could mean that obligation-responsiveness accounts will accommodate the case. Part of this response I will grant for the sake of argument: insofar as Ben and Andrea's friendship is sustained, they have obligations of friendship to one another. But the reason that Ben withdraws from Andrea is precisely that their friendship is fading. That which grounds whatever obligations one might think are part of friendship no longer applies to Ben. In other words, so long as Ben is Andrea's friend he may owe her his time, attention, and care, but he does not owe her his friendship.

Alternatively one might think that Ben's new behaviour reveals that he was never truly committed to the friendship. If this is the case then we might say that Ben does not disappoint Andrea because she was mistaken to trust in the first place. But though waning affection sometimes reveals that the friendship was unstable from the beginning, this need not be the case for Ben and Andrea; people and their friendships can change significantly over time without invalidating their earlier commitment to each other. We can assume for the sake of argument that Ben was at no point deliberately devious about his fondness for Andrea, and that Andrea is perceptive enough to know whether a person who is acting like a friend, or whom she wants to be a friend, really is a friend. Granting this, we can assume their friendship was at some point genuinely mutual, and Andrea's trust was not always mistaken. Genuine friendships also fade, and the fading of a friendship does not retrospectively delegitimate it.

Nonetheless, given the fading friendship, there will come a point at which continuing to trust Ben would be a mistake because his behaviour has provided enough evidence for a sufficiently perceptive person to recognize that he is no longer committed. And it is not implausible that Andrea might fail to recognize this despite her general ability to understand whether a friendship is authentic (perhaps the friendship has come to mean so much to her that her perception is clouded when it comes to Ben). There are thus two plausible versions of their story, both of which appear problematic for me. Perhaps Andrea fails to recognize Ben's withdrawal, and trusts where she should not. We might infer from this that Ben does not disappoint her trust because her trust is invalid. Or perhaps Andrea does recognize Ben's fading commitment to the friendship, and she continues to reach out to him not because she trusts he will reciprocate but because she hopes he will. We might infer from this that Ben disappoints hope, not trust.

But though both versions of the story are plausible, neither undermine the fact that before they reach the stage at which further trust would be a mistake, Andrea's previous trust in Ben has already been disappointed. This prior disappointment can be explained by two features

of the kind of trust found in friendship. The first is its scope. Entrusting a friend with an invitation to dinner has a relatively narrow scope: its fulfilment depends solely on what the friend does with the invitation. But this narrow entrusting is supported by a more general trust that we have in friends: that they will care about what is important to us, make time for us when we need them to, want to spend time in our company, and that they will do all of this because of a genuine commitment to our friendship. Andrea experiences a disappointment of this broader trust, and her trust is disappointed regardless of her attitude to the invitations she extends to Ben after the friendship has disintegrated.

Second, Andrea's trust is future-oriented. In this regard Andrea not only trusts that Ben will make time for her because he values their friendship, but she also trusts Ben to continue to do this in future. This future-orientation of trust explains the dependency that trust involves, and the associated risks to which that dependency exposes us. When we trust we make plans that depend on taking certain things for granted, and our intentions for the future are sometimes structured by the dependencies secured by trust; the depth of trust involved in strong friendships will sometimes support dependencies that run very far into the future. This kind of trust involves trusting not just that a person will care about us now, but that they will continue to do so in future, and our trust in friends is not qualified by a time-limit. Andrea's trust in Ben is disappointed precisely because she depends not only on his friendship in the present, but also on its endurance.

To recap: moral-motivation theories maintain that when X trusts Y to Φ, X's willingness to rely on Y's future Φ-ing is secured by X's optimism that Y has moral motivations that will lead Y to Φ. Such theories accurately capture some instances of trust; sometimes it is indeed the case that trust's confidence is secured by the moral qualities the truster ascribes to the trusted. But the problem is that moral-motivation theories do not extend to many common instances of trust, including the cases considered in this section. In some cases (the chess players, the civil adversary, the conference co-organiser) trust does not involve the ascription of moral motivations and therefore such an ascription cannot always be what supports a truster's confidence in the trusted.

And in trust between friends, the confidence in the trusted must be secured by something other than or in addition to their moral qualities, because the hopes of the truster can be disappointed despite the trusted showing no moral failing.

4 Trust's vulnerabilities

My aim in the rest of this paper is to argue for an alternative account of interpersonal trust that can explain both trust that ascribes moral motivations and trust that does not. Moreover, I will argue that my account is better able to accommodate the additional vulnerability introduced by trust in personal relationships, as exemplified in the fading friendship case. I will begin this positive account of trust by expanding further on the varieties of trust's vulnerability.

Philosophical work on trust tends to be sensitive to only two ways in which trust can be disappointed. The first is betrayal. The second is when the trusted's failure to do as we trust them to shows not that we have been betrayed, but that our trust was initially misplaced.[31] This could be because the relevant assessment of the trusted's competence was erroneous. Were I to entrust the care of an infant with their 5-year-old brother, I could not legitimately blame the brother if he fails; the mistake was mine. I can also misjudge whether a person is capable of completing a task by underestimating the difficulty of the task rather than overestimating the capacity of the trusted.

Trust can also be misplaced because it is incorrectly communicated. If the trusted is to be accountable for fulfilling the trust of another it must be the case that she knows or should have known that she has been entrusted with something, and accordingly the truster is subject to any norms we think usually apply to successful communication of expectations. Precisely how one articulates these norms (I will not give an account of this) will depend on how one understands a number of important variances across the range of situations in which we trust one another, including different levels of explicitness needed in different cases. When I trust another to keep a secret, what I tell that person will partly determine whether I can reasonably take for granted that they will appreciate this is just between them and me (consider the contrast between telling

someone the details of a traumatic event from my childhood and telling someone I am waiting for a bus). Similarly, who I am talking to will also partly determine how explicit I need to be about whether I am confiding in them (friends ought to be better at judging this for themselves).

A complete typology of trust's outcomes must include more than just satisfied, betrayed, or misplaced trust. Consider first the various ways in which we can culpably fail to satisfy the trust of another. Betrayal is only one version of this. Perhaps I culpably disappoint another's trust simply because I am maliciously indifferent to the concerns of others. Or perhaps I culpably disappoint trust not out of malintent, but culpable negligence e.g. I fail to show up to accompany a relative to a hospital appointment, despite knowing they depend on me, out of laziness or distraction or weakness of will in the face of more pleasurable alternatives. Culpable negligence can also extend to some cases of trust that fails to be satisfied because the trusted is not competent to the task. Not all failures of trust through incompetence are cases of misplaced trust. Sometimes a person will let down another because they were incompetent to a task that they should have been capable of doing.

The fact that misplaced trust is possible means that sometimes fault lies with the truster, not the trusted. It is my fault that my trust is not satisfied when I entrust care of an infant with someone clearly not competent to the task. Hence either the truster or the trusted can be culpable for disappointed trust. But it is also possible for trust to be disappointed without culpability on either side of the trusting relation. Cases of what I will call innocent disappointment of trust come in three kinds. The first is very similar to the culpably mistaken truster, but involves excusing factors that render the truster non-culpable for the mistake they have made in trusting. The 5-year-old baby-sitter hypothetical involves a mistake in a situation in which the stakes are potentially very high. More intelligible mistakes, in less risky situations, might be considered non-culpable. Say I trust my housemate to clean the bathroom this weekend, quite reasonably decide there is no need to communicate this explicitly because they have cleaned the bathroom regularly in the years

we have lived together, but do not know that they have to leave town for a family emergency. My ignorance explains my mistake in trusting as I do in this instance, but I am not culpable.

The second and third kinds of innocent disappointment of trust both involve changes in circumstances that alter what the trusted has most reason to do. The second kind of innocent disappointment involves changes in external circumstances that provide overriding reasons for the trusted to act in ways that prevent the satisfaction of another's trust. My hypothetical housemate's family emergency could be an example of this. Until they receive the news of the emergency, the most compelling reason relevant to their deliberation about whether to clean might be generated by the fact that I trust them to do so. But news of the emergency presents a new reason to abandon the cleaning in order to deal with much more important matters. It may be tempting to describe such changes in circumstances as excuses that negate the otherwise-culpability of the trusted, but this would be to presume guilt unless and until a case for the defence is successful. To think of trust in this way would be to render it a kind of coercive power, capable of imposing prima facie obligations on others simply by entrusting something with them and shifting the burden of proof for the defeat of the obligation onto the trusted.

Changes of circumstances can also lead to dissatisfied trust by undermining the reasons that direct someone to fulfil the expectations of trust. This is the third kind of innocent disappointment of trust, and it is the kind we find in the case of fading friendship. The reasons Ben has to spend time with Andrea, to celebrate her happiness, and to give her the care and support we reserve for friends in need, are grounded in their friendship. That friendship is itself grounded in Ben and Andrea's commitment to it, in the values and interests they share, and in the affection they have for one another. These grounds for Ben's reason to act as Andrea trusts him to are vulnerable to change. As the friendship fades, the reasons Ben previously had for acting in this way lose their grounds. Once his commitment to their friendship breaks down altogether, his reasons to prioritise Andrea as a friend would no longer apply.

Thus one relevant contrast between the second and third kinds of innocent disappointment lies in the normative consequences of changes of circumstances, the way that changes can alter the reasons relevant to whether a person fulfils another's trust. Sometimes our reasons to do as trusted remain, as it were, but are overridden. Sometimes they no longer apply. A second relevant contrast lies in the circumstances that change. The family emergency differs from Ben's change of heart insofar as the latter is a change in the psychology of the trusted – a change in values, interests, and attitudes toward his friend – whereas the former is a change in circumstances external to the trusted. Changes either internal or external to the psychology of the trusted can alter the reasons they have for fulfilling trust, and those normative alterations can result in the disappointment of another's trust without culpability. This possibility is particularly important for the further details of my account of trust in the following sections.[32]

Why do changes in circumstance render trust disappointed, and not mistaken? We might say that when my housemate learns of the family emergency, the trust is no longer valid; it would be foolish, perhaps callous, to expect them to clean the flat rather than deal with the emergency. Similarly, we might say that when Ben's affection for Andrea changes, she should not continue to trust him. But though the changes mean that continued trust would be misplaced, this does not retrospectively invalidate the trust prior to the change, as I have argued above with regard to Ben and Andrea. In both the housemate case and the fading friendship example, things turn out differently to what the truster hoped for and expected from the trusted; this is a disappointment of the original trust. Moreover in Andrea's case, the reason why continued trust would be a mistake – that Ben is no longer her friend – is part of the explanation for why her previous trust has been let down.

5 Commitments

Moral-motivation theories are, at best, incomplete. Where else might we turn to better account for the vulnerabilities of trust, while avoiding the problems outlined in section 3? One popular

alternative candidate is a trust-responsiveness account. Trust-responsiveness accounts maintain that trust expects not moral qualities, but responsiveness to another's dependence, and moreover responsiveness to that dependence for its own sake and not because I, for example, fear what you have threatened to do if I do not comply with your expectations. This is, for example, Jones' later position, an explicit departure from her earlier account.[33] The trust-responsiveness account provides a way of distinguishing trust from reliance – I do not expect my car to be trust-responsive – without moralising trust's expectations. But there are two problems facing this account.

First, it also accurately describes coercive communication of dependence. Say that I depend on a colleague to cover for an inexcusable absence, and I know that I can successfully guilt them into helping me ("I could lose my job if you don't make excuses for me"). Here I rely on my colleague's help, and if they are soft-hearted enough I can count on the fact that they will respond favourably to my communication of that reliance. But this is not trust but the exploitation of another's good nature. The coercive colleague is thus similar to the confidence trickster discussed earlier; they too communicate their dependence on their victim as part of their manipulation, and stake their nefarious plans on the responsiveness of their victim to that communication.

We might think that the problem has been caused by dropping goodwill from the picture. Jones' earlier position was that trust expects responsiveness to trusting that is motivated by goodwill. The guilted colleague, by contrast, is motivated to respond to my dependence on them by conscience, a motivation that could be ruled out by a goodwill-plus-responsiveness model. But the most troublesome aspect of the case is not that the colleague's response to my trust is motivated the wrong way, but rather that my attitude towards them is not one of trust. This is most evident in the case of the con-artist, who might rely on both the trust-responsiveness and goodwill of their victim, but does not thereby trust their victim. We must look elsewhere to distinguish trust from manipulation; I will return to this in my own account in section 6.

The second problem for the trust-responsiveness account is that it cannot capture what Andrea expects of Ben in virtue of their friendship. If Andrea expected Ben to simply be

responsive to her dependence on him, then she should be satisfied by Ben e.g. coming to dinner because he does not want to disappoint Andrea, despite the fact he would rather not spend time with her. Indeed, if Andrea's trust expects only that Ben will try not to let her down, then her trust could still be satisfied after the friendship is faded; Ben could keep up a pretence of friendship, or they could alter their relationship from sincere friendship to unsentimental dependence, with Ben regularly accepting invitations not out of fondness for Andrea but because he knows she depends on his company. Presumably, however, this is not going to satisfy Andrea, because she expects that Ben will spend time with her because he sees her as a friend and not because she is in need, as if out of charity.

It is more accurate to say that Andrea expects Ben to act in certain ways because of his commitment to their friendship, which suggests that we might be more successful with a commitment account of trust. One example of such an account has been defended by Katherine Hawley, who maintains that when we trust a person to Φ we understand the trusted to be committed to Φ and we rely upon them meeting that commitment.[34] We might suppose that we ought also to say that trust expects a person to be motivated by their commitment, but Hawley denies this.[35] However without attention to motivations a commitment account cannot distinguish trust from cases like the following. Say that Charlie has committed to joining me for dinner on Thursday, but she is forgetful and also commits to dinner with a mutual acquaintance, David. David tells me of Charlie's forgetfulness, but I do not worry because I know that David was in fact inviting Charlie to the same dinner that I was inviting her to. Thus I know that Charlie will act in accordance with her commitment to me, but this accordance will be unintentional, and I would not consider this evidence of her trustworthiness.

Hence a commitment account ought to also stipulate an expectation about the trusted's motivation. The most intuitive version of such an account, I submit, would be the following: when we trust a person to Φ we are confident that they will be motivated to Φ by a commitment we

ascribe to them. This is the account I defend in what follows, albeit with a number of qualifications.[36]

I will begin these qualifications with the ambiguous term "commitment". Sometimes we use "commitment" to refer to a normative demand. This is the sense of the term when it is used to talk about undertaking and succeeding or failing to meet commitments. Call these normative commitments. At other times we use "commitment" to refer to a particular kind of psychological attachment a person has to a wide variety of possible objects of commitment: actions, goals, values, projects, other people, etc. Call these psychological commitments.

Whether normative commitments apply to me does not depend on anything about my psychology. If I undertake a normative commitment through promising, whether that commitment applies to me in future does not depend simply on whether I want to fulfil it. Sometimes I do not even know that I have undertaken a normative commitment, e.g. if I fail to understand that something I have said has led another to reasonably expect something from me. Normative commitments function much like obligations, insofar as it is not up to me whether a commitment applies to me, or whether I have fulfilled it. And normative commitments often generate obligations; by promising to Φ I commit myself to Φ-ing and thereby generate an obligation to Φ. But normative commitments differ from obligations. Consider an example suggested by Hawley: I meet someone in a lawless desert in frontier America, and deliberate about whether I should shoot before the other draws their gun on me.[37] We treat each other with suspicion because neither of us have signalled a commitment to letting the other live, nor can such courtesy be presumed in this setting, yet nonetheless we each have a moral obligation not to murder the other. It is thus possible for us to be obligated without being committed. This is because normative commitments, unlike obligations, are always generated by something we have done or said – through, for instance, promising or contracting – whereas obligations are only sometimes generated this way.

Psychological commitments play a role in deliberation that make them significant for the expectations of trust. When a person has a psychological commitment to C they have an internal reason to A, where A is an action that sustains or contributes to the achievement of C, and an internal reason to A is a reason that is contingent upon my having what Bernard Williams called the relevant 'subjective motivational set': desires, evaluative dispositions, curiosities and interests, aversions etc.[38] Internal reason to A *per se* is not sufficient for commitment to C because internal reasons are not necessarily known to the person who has them. More specifically, then, when I am committed to C I have and know that I have internal reason to A.

Trust in personal and intimate relationships expects psychological commitments. Andrea trusts Ben to prioritise her, support her, give her the benefit of the doubt etc. because he is committed to their friendship, not out of obligation, but because he values their friendship and this valuing gives him reason to act in ways that support and sustain the friendship. Expectation of psychological commitments is also part of trust in less intimate, non-sentimental relationships. I can trust my chess opponent to be a good sport because I have confidence in their commitment to fair play, and I can trust my conference co-organiser to do a good job because I have confidence in their commitment to the value of a well-run conference. Generally, when trusting involves the expectation of psychological commitments it involves the judgement that the person trusted will possess the relevant subjective motivational set that gives them reason to act in the way I trust them to act. I also expect that the trusted will be aware of the rational relation between their motivations and the actions I want them to take. This involves, but is not exhausted by, a judgement of competence similar to other judgements of competence involved in trust.[39] Trust expects a person is generally capable of at least basic practical reasoning, but it also expects the person to understand that they have reason to perform the desired action in particular. Trust also expects that the person will take the reasons to perform this particular action to be stronger than countervailing reasons (we do not trust a person to Φ when we expect them to conclude that they should not Φ).[40]

It may seem a strong presumption for a truster to make if they are to ascribe to the trusted a rational complex of commitment, reasons generated by commitment, responsiveness to those reasons, and resistance to countervailing factors. But it is precisely the demandingness of this expectation that explains the variety of risks we are exposed to when we trust. The second and third forms of innocent disappointment of trust (section 4) are particularly salient here. Changes to circumstances can change the rational landscape for the trusted such that what they previously had greatest reason to do, because of the relevant psychological commitment, is no longer their first priority (consider again the housemate, committed to cleanliness in the flat, now dealing with a family emergency). And a change of heart can eliminate the commitment that previously gave them reason to act in the way expected by the truster, as in the fading friendship case. Both kinds of change threaten to disappoint the expectations of the truster because both challenge the normative priority that a relevant psychological commitment occupies in the deliberation of the trusted.

Does trust ever expect a person to be motivated by normative commitments? We might think that it is sometimes enough to trust someone if I expect that person will be motivated to Φ because they ought to, rather than because they will have internal reason to. Here the distinction between normative commitments and obligations becomes particularly helpful, because without this distinction a commitments account that includes normative commitments risks becoming too inclusive, failing to distinguish trust from similar ways in which we can be confident in others. Consider, for example, the confidence we must have in the behaviour of others if we are to share public spaces with complete strangers. If I have no confidence at all in others' respect for social norms of minimally acceptable conduct then I am likely to avoid, say, sharing a train carriage. In such cases I must expect that those around me will be responsive to obligations, either moral (do not harm others on the train) or conventional (norms about personal space), and I depend on my confidence in others' obligation-responsiveness in order to use public transport. But, arguably, this general social attitude is not an instance of trust.[41] Trust's expectations must be more specific, and

the category of normative commitments – which generate, for instance, promissory obligations, but are not interchangeable with obligations *per se* – could offer the required specificity.

However even with this greater specificity there is reason to be wary of including normative commitments in the commitments ascribed to the trusted. Promising and contracting are paradigmatic examples of normative commitments; they are the most familiar ways of undertaking a commitment and thereby generating an associated obligation. But such mechanisms are sometimes used precisely where trust is lacking. If I do not trust a person I may seek from them an explicit undertaking of a normative commitment as reassurance. And such an undertaking would be no comfort to me if I did not expect the person to be responsive to normative commitments. Thus sometimes we expect a person to be motivated by normative commitments without trusting them. Indeed, this expectation can help us make up for the lack of trust.

I believe there is room for diverging intuitions here, both regarding the nature of confidence in the behaviour of strangers – whether we trust when we share public spaces – and regarding whether willingness to depend on a promise can count as trust.[42] I choose to remain agnostic on both counts. Moreover, I propose it is a virtue of the commitments account I have outlined that it is flexible enough to accommodate both more or less inclusive intuitions about what counts as interpersonal trust. This flexibility allows for two versions of the commitments account. Both maintain that when we trust a person to Φ we depend on them to Φ, and we are confident that they will be motivated to Φ by a commitment we ascribe to them. The restrictive version of the commitment account maintains that when we trust we expect the trusted will be motivated to Φ by a psychological commitment that gives them reason to Φ. The non-restrictive version maintains that when we trust we expect the trusted will be motivated to Φ by either a psychological or normative commitment that gives them reason to Φ.

6 Further details

My account needs further detail in order to head off 4 potential objections: that my commitments account is a risk-assessment account; that it fails to exclude the confidence trickster; that it introduces one thought too many; and that it fails to account for the vulnerability generated by trust. I will take each in turn.[43]

First, it may seem that without appeal to moral-motivations, my account relapses into a risk-assessment account of trust. Risk-assessment accounts maintain that trust is willingness to depend on another person secured by an attractive ratio of risk-to-potential-benefit, usually in virtue of a person's relatively high predictability. My account may be accused of treating commitments in the same way that risk-assessment accounts treat evidence used to predict future behaviour; at best, it might be said, I have simply added detail to the folk-psychology supporting judgements of predictability. Perhaps, for instance, I rely on Kant's regularity because I believe he has a commitment that increases the likelihood he will keep a regular schedule. But if my account is a risk-assessment account, it is vulnerable to the objection that prompts moral-motivation theories in the first place: risk-assessment accounts fail to distinguish trust from reliance.

However there is an important difference between my account and risk-assessment accounts. Risk-assessment accounts are indiscriminate with regards to motivations. For a risk-assessment account, a truster's confidence in the predictability of another can be supported by any psychological disposition that secures high probability of the desired behaviour: fear of social sanction, addiction, stubbornness, etc. By contrast, the ascription of a commitment requires that the truster takes the trusted to be responsive to practical reasons. Thus confidence grounded in another's commitments rules out, for instance, the kind of reliance we have towards inanimate objects, which have no commitments. It also rules out treating the person I trust as a non-rational conduit for psychological dispositions that they may or may not be aware of. Consider some of the examples raised earlier. If optimism about my conference co-organiser, whom I expect to go beyond the call of duty, is based on my judgement that they are a workaholic, then on my account I rely on this addiction, but I do not trust them. Similarly, if I rely on Kant's regularity because I

believe him to have a pathological compulsion to an orderly schedule, I rely on his compulsion but do not trust him. If, however, I rely on his regularity because I think that he values routine, and that he is capable of acting on the internal reasons that are generated by that value, then I trust him.

This requirement of minimal confidence in the trusted's rationality applies regardless of whether the truster ascribes a normative or psychological commitment. If I trust another because I think they will be motivated by a normative commitment, I take them to be responsive to the obligations generated by that commitment.[44] If I trust because I think the trusted will act on a psychological commitment, I take them to be responsive to the internal reasons comprising their commitment. Andrea would not trust Ben if she thinks he is unable to recognise his fondness for her, *a fortiori* if she thinks that he lacks the rational capacity to do what fondness for her gives him reason to do. Relying on the predictability of persons – the kind of "trust" that features in risk-assessment theories – requires no such faith in basic rationality.

Trust's confidence in rational responsiveness also allows me to explain how it is possible to bootstrap trust by generating, through trusting, the same conditions that secure my confidence in the trusted. Recall the lawless frontier example, cited earlier to explain the difference between obligations and commitments. Say that in this scenario I put down my gun because I trust that the stranger will reciprocate this conciliatory gesture. For this to be trust and not simply a gamble, I must be confident that the stranger will reciprocate. But it could be that the stranger has no plausible motivation to disarm until I have done so. I thus create the stranger's motivation to disarm – the same motivation that supports my trust – by communicating that I trust them, through my own disarming gesture.

This bootstrapping is made possible by the fact that I can alter the rational landscape of the stranger, generating reasons to disarm and eliminating reasons to stay armed. This non-verbal gesture is in this respect a particular form of rational address, much like other forms of address I might use to engage the practical reason of the trusted and attempt to show them what I think

they have reason to do. But the lawless frontier example also brings into relief how this bootstrapping requires that I think that the trusted has some prior commitment that allows me to generate reasons for them in this way. Unless I am confident that the values of the stranger will incline them to see my white flag as a reason for them to disarm, my conciliatory gesture gives me no new confidence that they will be motivated to refrain from attack, and my surrender is just a gamble.

A potential problem re-emerges, however, when we consider other ways in which we can manipulate the reasons that are generated by a person's commitments. The confidence trickster (see section 2) may rely on their victims' commitments and their rational responsiveness for the con to work, in which case it seems that my commitments account, like other accounts considered earlier, has also failed to exclude the con-artist. For my account to solve this problem, we must further stipulate that when we trust we share the commitment that we ascribe to the trusted. That is, when I trust another to do something I judge that they have values, goals, and projects that are sufficiently similar to my own for me to expect them to act as I trust them to. A confidence trickster might share the commitment of their victim on one level of description – say, a commitment to financial security – but the commitments ascribed by trust are more specific, and the con-artist certainly does not share their victim's specific commitment to the financial security *of the victim*.

But we must go further still in order to exclude other cases of manipulation. Consider two political activists, both genuinely committed to their cause, one of whom exploits the commitment of the other in pursuit of their own selfish ends. To avoid mislabelling this as trust, the commitments account must say something about the relation between the commitment that a truster ascribes and the action that they expect: the act on which the truster depends is specifically an act that the truster believes serves the commitment they ascribe to the trusted. Thus a parent trusts a babysitter to act for the good of the child, a researcher trusts a colleague to act in order to organise a good conference, and even adversaries might trust one another to act in ways that

uphold a commitment to civility. Conversely, exploitation of fellow activists expects acts that serve oneself, not the cause. And when exploitation instead depends on threats against that which the victim is committed to – in which case the act coerced from the victim does serve their commitment, but this time by preventing the threatened harm – the exploiter's threat against the cause indicates that, like the con-artist, they do not share the commitments of the other.[45]

The shared-commitments specification might seem to require that the truster holds another very strong presumption, this time of similarity of character between themselves and the trusted. But note that sharing a commitment can be narrowly domain specific, such that I might e.g. share a commitment to high standards in teaching with a colleague, and on this basis trust them professionally, while we nonetheless disagree on most political issues and have wholly incompatible tastes and interests. Note also that my analysis of psychological commitments allows for the possibility that two people may be committed to the same thing though they perform different actions, for different internal reasons that depend on different psychological features. A person has a psychological commitment to C when they have an internal reason to A, where A is an action that sustains or contributes to the achievement of C. But two people can be committed to the same C without having the same internal reasons, or having the same subjective motivational set that supports their internal reasons. Say that I am committed to high standards in teaching insofar as my benevolence towards students gives me reason to work hard for their benefit. I might also trust my colleague to do right by their students because they too are motivated by a commitment to high standards in teaching, though I believe this commitment consists in their valuing a well-educated citizenry and the practical reasons generated by this value. My trust is thus compatible with a judgement that we are committed to the same thing for different reasons. A manipulator need not judge they share a commitment even in this thin, undemanding sense.

The introduction of shared commitments makes my account similar to Mullin's commitments account. According to Mullin, when we trust a person we assume that they share our own commitment to a relevant social norm.[46] Thus Mullin maintains that when I trust a friend

I am confident that their behaviour will be motivated by their commitment to a social norm of friendship, and that this friend interprets this norm as I do. Though I am sympathetic to Mullin's account, there are two problems with it. First, Mullin's stipulation that the norm in question is a social norm is incorrect, because trust can sometimes expect commitments to standards that are in conflict with social norms. Thus e.g. if I have an unorthodox understanding of good teaching, then I might reserve a particular kind of trust only for those colleagues who share this standard of good teaching, despite its divergence from prevalent norms.

Second, even where we do expect the commitments of the trusted to abide by a social norm – our standards are orthodox – we do not always expect them to be directly committed to that social norm. Sometimes I trust my friends to make time for me not because I am confident that they aspire to live up to the standard of being a good friend, but because I am confident that they are committed to our friendship, that is, that they value the time we spend together and care about my happiness, and that these motivations give them reasons to act in ways that sustain our friendship. Mullin's account conflates being motivated by a commitment to a norm with being motivated by commitments that accord with a norm. As a consequence, her commitments account results in a "one thought too many" error in its characterisation of trust between friends, as if friends expect each other to first consider what it would be to be a good friend, and then apply this understanding to their friendship.

It might be objected that my own account also introduces one thought too many, just a different thought. I have maintained that trusting friends have confidence in one another's commitment to their friendship. My superfluous thought, we might say, is introduced by my claim that we trust friends to be motivated by their commitment to a particular friendship. It seems that on my account a trusted friend is expected to acknowledge both the friendship itself and their commitment to the friendship, and to act on the basis of this double acknowledgement. But if Ben had fulfilled Andrea's trust, is it not possible that he could have been motivated only by the friendship, and not a superfluous thought about his commitment to their friendship?

I suggest that one of the advantages of my analysis of psychological commitments is that it prevents the need for this unnecessary extra thought. On this analysis, Ben remains committed to his friendship with Andrea so long as he still acts on internal reasons to do things that allow the friendship to flourish (return her calls, spend time with her, support her through difficulty, celebrate her happiness etc.) But this commitment to their friendship need not involve reasons that directly refer to that commitment. For Ben to be committed to the friendship, and for his behaviour to be motivated by this commitment, it is enough that Ben values Andrea's happiness and takes pleasure in spending time with her, that these motivations provide the reasons that determine his behaviour, and that this behaviour supports the flourishing of their friendship. Ben need not intend the flourishing of the friendship as his end in order for his motivations and actions to show his commitment to that friendship. Accordingly on my version of a commitments account Ben can satisfy Andrea's trust without having a superfluous thought about their friendship.

The final refinement to the commitments account follows from my earlier observation that in trust we are vulnerable to changes in circumstance that do not retrospectively invalidate trust but instead disappoint it. Trust is, in short, vulnerable to change over time. This is possible only because trust is most often a future-oriented attitude; in many cases time passes between the moment when one first trusts, and the moment when the trusted satisfies or disappoints our trust. When trust is future-oriented it involves an expectation about another person's future behaviour on which we depend, and hence according to the commitment account trust sometimes involves the expectation that a person will be motivated by a commitment that they will have at the relevant future moment. This may but need not be because we expect them to continue to be committed to something to which they are presently committed. In Andrea's case, she has confidence in Ben's current commitment to their friendship and in the continuation of this commitment in future. But we could also have reason to think that someone will develop a relevant commitment in future, when we need them to, and that they will act on this commitment in a way that is favourable to us. Less common, but nonetheless possible, is trust that is past-oriented. I can trust that my

housemate has taken care of the cleaning while I have been away, perhaps on the basis of the same judgement about their commitments to cleanliness that would also warrant my trust that they will do the cleaning when expected in future.

The future-orientation of trust in personal and intimate relationships renders it particularly risky because it involves an expectation that the trusted's present psychological commitments will persist in future. This is why this kind of trust is especially hard won, valuable when established, and vulnerable to failure. It is also why judgement of another's trustworthiness in the particular case of future-oriented intimate trust will likely include a judgement about their constancy, that is, how prone they are to fluctuations in interests, desires, and personality. If Andrea thought Ben was too changeable she would not trust that in future he would be there for her to help her through difficulty and share her celebrations.[47] But while expectations about constancy may well count as another form of moral expectations, not all decline of intimate relationships is due to inconstancy. Consider how Austen describes the change in relation between Anne Elliot and Captain Wentworth in *Persuasion*:

> They had no conversation together, no intercourse but what the commonest civility required. Once so much to each other! Now nothing! There *had* been a time, when of all the large party now filling the drawing room at Uppercross, they would have found it most difficult to cease to speak to one another. With exception, perhaps, of Admiral and Mrs. Croft, who seemed particularly attached and happy, (Anne could allow no other exception even among the married couples) there could have been no two hearts so open, no tastes so similar, no feeling so in unison, no countenances so beloved. Now they were as strangers; nay, worse than strangers, for they could never become acquainted. It was a perpetual estrangement.[48]

Once close, almost betrothed, Anne and Wentworth see each other after many years apart, now without the prior attachment. Anne mourns this new estrangement, and is uncomfortably

uncertain about whether it is caused by asymmetrical lack of feeling (has Wentworth's view of her changed, or is he hiding his abiding affections?). But though the prospect of Wentworth's change of heart is hurtful for Anne, she does not think less of Wentworth for it. Even Austen, usually so disparaging of the inconstancy of her less mature, more whimsical characters, sees fit not to present Wentworth's apparent change as evidence of vice.

Notes

[*] For conversations about trust and feedback on drafts I am very grateful to David Batho, Karamvir Chadha, Jacopo Domenicucci, Matteo Falomi, John Filling, Richard Holton, Nick Joll, Rae Langton, Gaby Silva Rivero, anonymous readers for *Ethics*, and students in my moral psychology lectures at the University of Cambridge. Thanks also to Jason D'Cruz for sharing his chapter with me before it was available in print.

[1] E.g. JM Bernstein 'Trust: on the real but almost always unnoticed, ever-changing foundation of ethical life' *Metaphilosophy* 42.4 (2011), 395-416; Pamela Hieronymi 'The Reasons of Trust' *Australasian Journal of Philosophy* 86.2 (2008), 213-236; Richard Holton 'Deciding to trust, coming to believe' *Australasian Journal of Philosophy* 72.1 (1994), 63-76; Andrew Kirton (2020) 'Matters of Trust as Matters of Attachment Security' *International Journal of Philosophical Studies*; Thomas Nys, 'Autonomy, Trust, and Respect' *Journal of Medicine and Philosophy* 41.1 (2016), 10-24. Note that in later work Domenicucci and Holton demur on Holton's earlier emphasis on betrayal (see fn.7 in Jacopo Domenicucci and Richard Holton 'Trust as a Two-Place Relation' in Paul Faulkner and Thomas Simpson eds. *The Philosophy of Trust* (Oxford: Oxford University Press, 2017)).

[2] Zac Cogley, 'Trust and the Trickster Problem' *Analytic Philosophy* 53.1 (2012), 30-47; Karen Jones, 'Trust as an affective attitude' *Ethics* 107 (1996) 4-25; Karen Jones 'Second-Hand Moral Knowledge,' *Journal of Philosophy*, 96.2 (1999), 55–78; Carolyn McLeod, *Self-Trust and Reproductive Autonomy*, (Cambridge, MA: MIT Press, 2002); Philip J Nickel, 'Trust and Obligation-Ascription' *Ethical Theory and Moral Practice* 10.3 (2007) 309-319.

[3] For trust in institutions see Onora O'Neill, *A Question of Trust: the BBC Reith Lectures* (Cambridge: Cambridge University Press, 2002); for self-trust, McLeod, *Self-Trust and Reproductive Autonomy*; for therapeutic trust, Jones, 'Trust as an affective attitude', 5.

[4] Russell Hardin, *Trust and Trustworthiness*, (New York, NY: Russell Sage Foundation, 2002), 6-7.

[5] I also follow most of the literature in treating trust as a three-place relation ("I trust her to Φ"), though I remain agnostic on whether trust is ever a two-place relation ("I trust her *simpliciter*"). For more see Domenicucci and Holton, 'Trust as a Two-Place Relation'.

[6] See e.g Nickel, 'Trust and Obligation-Ascription', pp.313-314; Amy Mullin, 'Trust, Social Norms, and Motherhood,' *Journal of Social Philosophy*, 36.3 (2005), 316–330.

[7] Jones, 'Second-Hand Moral Knowledge', 68.

[8] Risk-assessment accounts include: Diego Gambetta, 'Can We Trust Trust?' in Diego Gambetta (ed.) *Trust: Making and Breaking Cooperative Relations*, (New York: Basil Blackwell, 1988); and Philip Pettit, "The Cunning of Trust," *Philosophy and Public Affairs*, 24 (1995) 202–225.

[9] The Kant example is from Annette Baier 'Trust and Antitrust' *Ethics* 96.2 (1986) 231-260

[10] See e.g. Hardin, *Trust and Trustworthiness*, 12.

[11] Baier, 'Trust and Antitrust', 234; Jones, 'Trust as an Affective Attitude'.

[12] Jones' early theory of trust is a moral-motivation theory; Baier's is not. For Baier, goodwill secured by moral qualities is just one of many motivations we might ascribe to someone we trust (see e.g. Baier , 'Trust and Antitrust, 243). Jones' later work on trust is also pluralistic about the motivations ascribed to the trusted (see e.g. Karen Jones 'Trustworthiness,' *Ethics*, 123.1 (2012), 61–85).

[13] Jones, 'Second-Hand Moral Knowledge', p.68.

[14] McLeod, *Self-Trust and Reproductive Autonomy*, 21.

[15] See e.g. Jones, 'Trust as an affective atttiude', 10, and Jones, 'Second-Hand Moral Knowledge' 68.

[16] Jones, 'Trust as an affective attitude', 21.

[17] Ibid., 7.

[18] McLeod, *Self-Trust and Reproductive Autonomy*, 21.

[19] Ibid. 22

[20] Ibid. 27; see also McLeod, 'Our Attitude Towards the Motivation of Those We Trust' *Southern Journal of Philosophy* 38.3 (2000) 465-479.

[21] See e.g Nickel, 'Trust and Obligation-Ascription', 313-314; Mullin, 'Trust, Social Norms, and Motherhood', 316.

[22] Hieronymi, 'The Reasons of Trust'; Holton, 'Deciding to trust, coming to believe'.

[23] Holton, 'Deciding to trust, coming to believe'; Cogley, 'Trust and the Trickster Problem'.

[24] Hieronymi, 'The Reasons of Trust'; Holton, 'Deciding to trust, coming to believe'.

[25] Cogley, 'Trust and the Trickster Problem'; Kirton, 'Matters of Trust as Matters of Attachment Security'.

[26] McLeod, 'Our Attitude Towards the Motivation of Those We Trust', 474.

[27] Nickel argues that the obligation ascribed by trust is a moral obligation because it imposes a requirement, and failure to meet that requirement would warrant blame ('Trust and Obligation-Ascription', 312). Whether this is enough to label an obligation moral is debatable. But my objections in the next section hold even if obligation-responsiveness theories abandon the "moral" qualifier, because trust does not always expect fulfilment of blame-warranting (moral or non-moral) obligations.

[28] Cogley 'Trust and the Trickster Problem'; Marc A Cohen and John Dienhart, 'Moral and Amoral Conceptions of Trust, with an Application to Organizational Ethics', *Journal of Business Ethics*, 112 (2013), 1–13.

[29] Do obligation-responsiveness theories thereby exclude the possibility of immoral trust between immoral people? Nickel argues that they need not. On his view, we can distinguish between the moral character of the attitude trusters have to trustees and the grounds trusters have for having that attitude (Nickel, 'Trust and Obligation-Ascription', 313-314). Thus trust among thieves could involve amoral people ascribing obligations to each other, thereby feeling betrayed when they think they are entitled to, even though they do not think these obligations have moral grounds. Perhaps Nickel is wrong about this, in which case obligation-

responsiveness theories are open to both extant objections to moralising theories and my own objections.

[30] Mullin, 'Trust, Social Norms, and Motherhood', 318

[31] E.g. Katherine Hawley, 'Trust, Distrust, and Commitment' *Noûs* 48 (2014) 1-20

[32] Some of the possibilities outlined here can be accommodated by moral-motivation theories. Moral-integrity theories, for instance, can maintain that my trust in my housemate to clean can be innocently disappointed due to changes in external circumstances, because moral integrity should lead my housemate not to clean when another's more urgent need for help arises. But no moral-motivation theory accommodates innocent disappointment due to a change of heart. The important change for Ben is not what morality demands of him, nor what he thinks morality demands of him, but what he remains committed to. Neither Ben nor Andrea can find solace in the fact that Ben has disappointed Andrea in pursuit of a competing moral cause; their friendship simply faded.

[33] In her 2012 paper on trustworthiness Jones rescinds her earlier inclusion of goodwill in her theory of trust ('Trustworthiness', p.67) and has more recently argued for the trust-responsiveness account (see Karen Jones "'But I was counting on you!'" in Paul Faulkner and Thomas Simpson eds. *The Philosophy of Trust* (Oxford: Oxford University Press, 2017) and Karen Jones 'Trust, Distrust, and Affective Looping' *Philosophical Studies* 176.4 (2019), 955-968). Trust-responsiveness also features in accounts proposed by: Victoria McGeer and Philip Pettit 'The Empowering Theory of Trust' in Paul Faulkner and Thomas Simpson eds. *The Philosophy of Trust* (Oxford: Oxford University Press, 2017) and Paul Faulkner 'The Problem of Trust' in Paul Faulkner and Thomas Simpson eds. *The Philosophy of Trust* (Oxford: Oxford University Press, 2017).

[34] Hawley, 'Trust, Distrust, and Commitment', 10.

[35] Ibid. 5-6.

[36] Hardin's encapsulated-interests theory offers another non-moralised motivation-based account, but it is less likely to successfully characterize trust in personal relationships. As my argument is focused against moral-motivation theories I will omit a full defence of this claim, and state only the following. Hardin's account, when extended to trust between friends, renders it an attitude in which we expect our friendship will be in the interest of our friend, and we trust them to act accordingly because of that interest. But this mischaracterises friendship. For instance, it introduces one thought too many into what I hope for from a friend: when I trust a friend to give me the benefit of the doubt I do not expect him to do so because it sustains the friendship, and because our friendship is in his interest; I expect him to be charitable because he is optimistic about my character in the way that friends are.

[37] Hawley, 'Trust, Distrust, and Commitment', 19.

[38] Bernard Williams, 'Internal and External Reasons', reprinted in *Moral Luck* (Cambridge: Cambridge University Press, 1981), 101-113.

[39] Trust ascribes not only a commitment to Φ but also competence to Φ. Given the competence claim relatively uncontroversial, I will leave further analysis of trust's competence-ascription for another paper.

[40] Andrew Kirton ('Matters of Trust as Matters of Attachment Security') objects to Hawley's commitments account on the grounds that it fails to distinguish the legitimacy of a trustor's feeling betrayed from the culpability of the trusted. Kirton cites unattached sexual relations as an example: a person might legitimately feel betrayed by a sexual partner who sleeps with other people even if that partner has undertaken no commitment – explicit or implicit – to monogamy. The inclusion of psychological commitments in my account distinguishes my position from Hawley's in a way that forecloses Kirton's objection. The feeling of disappointment in e.g. the trusted sexual partner (Kirton calls this betrayal, I call it innocent disappointment) can be explained by the trustor ascribing a psychological commitment to an exclusive relationship. The disappointment arises either because that psychological commitment changes over time, or

because the initial ascription of the psychological commitment turns out to have been mistaken. The important point is that on my commitments account, trust's distinctive vulnerability can be generated by a form of trust that ascribes only psychological commitment, and does not require that the truster thinks the trusted is normatively bound.

[41] The difference is well understood by reactive-attitude accounts, despite my objections to those which moralise trust. If a stranger steals my laptop this could legitimately elicit resentment, but feelings of betrayal would be misplaced. But if the stranger has first promised to watch over my laptop while I step out to take a phone call, I would be right to feel betrayed.

[42] Contrast e.g. McGeer and Pettit, 'The Empowering Theory of Trust', 16, with e.g. Eric Uslaner *The Moral Foundations of Trust* (Cambridge: Cambridge University Press, 2002).

[43] Even with this further detail, my account is still incomplete. More could be said, for instance, about how my version of a commitments theory would account for distrust. For more on distrust in commitments theories, see Jason D'Cruz 'Trust and Distrust' in Judith Simon (ed.) *The Routledge Handbook of Trust and Philosophy* (Routledge, 2020), 43-44. Note that D'Cruz raises unresolved questions about distrust for commitment-based theories, but does not give reasons for thinking that commitments theories cannot answer these questions.

[44] Is my account thus an obligation-responsiveness account? No, because (a) obligation-responsiveness accounts do not specify that the relevant obligations are generated by commitments and (b) my theory allows for obligation-responsiveness as one possible motivation ascribed to the trusted, but does not maintain that trust always involves this.

[45] I thank an anonymous reader at *Ethics* for pressing the point about manipulation cases.

[46] Mullin, 'Trust, Social Norms, and Motherhood', 316.

[47] Hardin observes that the English word trust has its origins in the middle English "tryst", once used to refer specifically to holding one's place as part of a team effort, particularly in hunting, while game is chased one's way. Thus we might once have said "it is my trust to hold my place" (*Trust and Trustworthiness*, 76).

[48] Jane Austen, *Persuasion* (Penguin Random House, [1818] 2015), 60.