

TRUSTWORTHINESS, RESPONSIBILITY AND VIRTUE

BY ALISON HILLS

In the current philosophical literature on trustworthiness, two claims are very widely accepted, first that trustworthiness is a kind of reliability and secondly, that trustworthiness is not a virtue. Both claims are made, for instance, in Hawley's recent highly influential account of trustworthiness. I argue that both are mistaken. I develop and defend a new account of trustworthiness as responsibility, contrasting it with reliability and obligation accounts of trustworthiness. I argue that trustworthiness as responsibility is very plausibly a kind of moral virtue.

Keywords: trustworthiness, trust, responsibility, virtue.

I. INTRODUCTION

What is trustworthiness?¹ Someone who is trustworthy is competent and dependable. Trustworthy people tell the truth, keep their promises, fulfil their contracts. The untrustworthy are sly, exploitative, manipulative; they trick or cheat you if they can, or they are weak, irresolute, incompetent and will let you down. Since the trustworthy are dependable and competent and the untrustworthy are not, it should come as a surprise to find in the philosophical literature a broad consensus that trustworthiness is not a virtue.²

Thus Jones: 'Trustworthiness does not rate a mention on classical lists of the virtues. . . I argue that the classical lists are correct in excluding trustworthiness' (2012: 78).

Baier: 'Women, proletarians, and ex-slaves cannot ignore the virtues . . . of judicious untrustworthiness' (1986: 253).

¹ This paper is a response to Katherine Hawley's work on trustworthiness. Katherine was one of my philosophy tutors in my first year at university; the first to explain how to argue about and how to write about philosophy. For me, as for many others, she was a role model and an inspiration. She is very much missed.

² An exception is Potter (2002).

‘It may sound odd to insist that trustworthiness is a virtue or, in other words, a moral disposition to be trustworthy’ (McLeod 2021).

Hawley: ‘this account of trustworthiness does not sit comfortably within a virtue ethical approach’ (2019: 76).

This should be puzzling. After all, calling someone untrustworthy is clearly a criticism; calling them trustworthy a compliment. Imagine a morally perfect agent: is that person untrustworthy? Or neither trustworthy nor untrustworthy? Of course not. An ideal moral agent is obviously trustworthy. So contra McLeod, it sounds much odder to deny that trustworthiness is a moral virtue than to insist that it is.

Of course appearances can be misleading. Those who deny that trustworthiness is a virtue have arguments that it does not fit the profile of a virtue. Nevertheless, I will argue that in this case, appearances are not misleading. Trustworthiness really is a virtue: the arguments against fail, sometimes because they mistake what is required for virtue; sometimes because they mistake what is required for trustworthiness.

In the first part of this paper, I will contrast the dominant account of trustworthiness in the literature, trustworthiness as reliability, drawing particularly on Hawley’s influential version of trustworthiness as reliability in carrying out commitments, with a different approach that I introduce and develop here, understanding trustworthiness as a kind of responsibility. In the second part of the paper, I will argue that trustworthiness, understood in this way, is a kind of moral virtue.

II. PART 1: WHAT IS TRUSTWORTHINESS

II.1 Hawley’s conception of trustworthiness: reliability in commitment

In much of the literature on trustworthiness, it is understood as a type of reliability (Baier 1986; Goldberg 2020; Hawley 2014, 2019; Jones 1996, 2012). Initially, this sort of account looks very appealing. In ordinary language reliability is a synonym for trustworthiness and it is very natural to think that trustworthy people are reliable.

Reliability accounts of trustworthiness often begin with and develop from an account of trust. For instance, though Hawley’s book (Hawley 2019) is called ‘How to be trustworthy’, chapter 1 is an account not of trustworthiness, but of trust and distrust.

What is it to trust someone? One obvious answer is: to trust someone is to rely on them. More specifically, S’s trusting X to φ is a matter of S’s relying on X to φ (Goldberg 2020: 99). Thus, if you trust someone’s testimony then you rely on what they say; if you trust them to fulfil their promise, then you rely on their doing what they said they would do.

So if the idea of trust is prior to trustworthiness, and if trust is a kind of reliance, then to be worthy of trust is, presumably, for it to be appropriate for someone to rely on you. So trustworthiness must be a kind of reliability. Again, more specifically, if X is trustworthy with respect to φ for S , then it is appropriate for S to rely on X to φ . Notice that reliability (and hence trustworthiness) is qualified here: it might not be appropriate for someone else to rely on X , and it might not be appropriate for S to rely on X for anything other than φ -ing.

If reliability were necessary and sufficient for trustworthiness, then I think it would really be implausible that trustworthiness were a moral virtue. Moral virtues are dispositions to act well—to do the right thing—on the basis of good motivations. Reliability does not have the right profile for a virtue: for one thing, you can be reliably morally bad as well as good. Suitably qualified reliability does not even require action: the comatose are, after all, very reliable in some respects. Nor does reliability even in morally good action, like telling the truth and keeping your promises—require good motivations: they may be good, bad or indifferent.

Of course, it is swiftly acknowledged by all that mere reliance is not sufficient for trust and mere reliability not sufficient for trustworthiness. For we rely on inanimate objects—shelves that hold precious objects, bicycles that we need for transport—but we do not trust them, at least in a ‘rich’ sense (Baier 1986; Hawley 2014, 2019; Jones 2012). A sign of rich trust and trustworthiness as opposed to mere reliance or reliability is a set of distinctive reactive attitudes. If you rely on a shelf and it collapses, you are angry, disappointed, frustrated. But if you trust me to keep a promise and I fail you, you feel betrayed, let down; you would demand redress, an apology, restitution. I would or at least should feel remorse, guilt, and ask for forgiveness. None of this would be appropriate with regard to a shelf.

What exactly is the difference between reliance and this richer kind of trust that explains this difference in reactive attitudes? According to Richard Holton’s influential account of reliance, to rely on someone to X is to act on the supposition that she will X (1994). Many claim that reliance of this kind is necessary but not sufficient for rich trust. Rich trust also requires an extra factor, typically a particular kind of motivation attributed to the person trusted. For instance, for Baier (1986), trust implies good will in the person trusted; for Hardin (2002), that I incorporate your interests into my own, for the sake of maintaining our relationship; for Jones (1996: 5–6), that I will be moved favourably by the thought that you are depending on me.

This ‘reliance plus motivation’ might seem like an attractive account of trust, since it successfully distinguishes trustworthy people from solidly built household objects. But there are some significant problems for it nonetheless.

In the first place, there is, I think a good question whether these motivation accounts make good sense of betrayal being the characteristic mark of trust

that has been failed. If you thought someone had good will or would take your interests into account or that they would treat as a reason that you depended on them, then it would be disappointing to find out that you were wrong. But it is far from obvious that this is a betrayal.

Secondly, these sorts of accounts do not make good sense of distrust (Hawley 2014, 2019). For distrust cannot simply mean failing to rely on someone, or failing to attribute the right motives to them, because we can do that to inanimate objects but we do not distrust them (in the rich sense). Is it then to fail to rely on them, because you attribute bad motives to them, for instance, ill will? But distrust does not require this—someone may be completely indifferent to others, bearing them neither ill will nor good will. It could be appropriate to distrust this person. So motivation accounts cannot make good sense of distrust. Hawley presses these problems against motivation-reliance accounts, but in the service of a conception of trust understood as reliability plus a different quality, reliability in commitment:

To trust someone to do something is to believe that she has a commitment to doing it, and to rely upon her to meet that commitment. To distrust someone to do something is to believe that she has a commitment to doing it, and yet not rely upon her to meet that commitment. (Hawley 2014: 10)

This explains the difference between trusting and relying. Since inanimate objects do not make commitments, we can rely on them but we cannot trust or distrust them. It neatly solves the difficulties we found with a good motivations account: it has a proper account of distrust, and so allows for a difference between distrust and no attitude of trust. It also explains very well why betrayal is the characteristic response to misplaced trust, as it is the appropriate response to a failure to fulfil a commitment.

Hawley's account of trustworthiness draws on this account of trust. To be trustworthy is to be the kind of person who does what she commits to do. Trustworthy people are reliable in fulfilling the commitments that they make and, Hawley emphasizes, judicious in taking on commitments, to make sure that they are not overcommitted. To the objection that people can demonstrate trustworthiness even in situations they have made no explicit commitment, Hawley appeals to implicit commitments: strangers are trustworthy insofar as they are disposed not to attack me, even if they have made no explicit commitment not to do so, because they have implicitly committed not to do so.

As with reliability accounts generally, Hawley's version of trustworthiness is not, as she notes, very plausibly a moral virtue (2019: 76). For one thing, Hawley puts no limits on what commitments count: a commitment to extort money or taking on a contract killing is just as much a commitment, on this view, as a commitment to help a student or look after a child. So trustworthy

people, according to Hawley, may do very bad actions as well as morally good or neutral ones.

Secondly, there is no requirement on the trustworthiness as commitment view that the trustworthy act for any particular motive. A trustworthy person might fulfil their commitments because they care about those commitments, or because they care about the effect on others if they do not fulfil them. But equally they might fulfil commitments for fear of sanctions or the sake of their reputation. A businessman who fulfils commitments to hire and promote women might do so because he thinks it is important, or because he takes his commitments seriously, or for fear of legal challenge, but as long as he does so consistently, he is trustworthy, according to Hawley.

Trustworthiness as fulfilling commitments does not have the right profile for a moral virtue, but it might well be a kind of executive virtue. An executive virtue is a trait of character that is useful whatever ends one has set. Instrumental rationality is a kind of executive virtue, for being good at taking means to your ends is useful whatever those ends are. Similarly, trustworthiness as commitment is plausibly an executive virtue. It is useful to you to have the resolution and determination to follow through on your commitments, whatever they are, and to choose them carefully so that you do not overcommit. Executive virtues are creditable in some respects, but not praiseworthy all things considered. Anything that makes bad people more effective agents cannot be considered good without (significant) qualification.

III. TRUSTWORTHINESS AS RESPONSIBILITY

Hawley's account of trustworthiness is an improvement on 'reliance plus motivation' accounts and I think there is no doubt that it does capture a centrally important feature of trustworthiness. But reliability is not all there is to trustworthiness, even a sophisticated reliability account is too impoverished to understand trustworthiness fully. A trustworthy person is not, or not merely stolidly predictable, difficult to distinguish from a solidly constructed shelf, even a committed shelf.

Suppose that I am looking for someone trustworthy to look after my children for a few hours after school. I need to know that this person will do what they agreed: turn up on time, pick my children up from school, stay with them until the agreed time and so on. So far, Hawley will agree: this is trustworthiness as reliability in commitment.

But someone might be reliable in these ways yet wholly unsuited to be trusted with my children. Although we may have discussed explicitly and implicitly agreed a set of commitments, looking after children is inherently unpredictable. Children run into the road, hide and refuse to come out, eat dirt or leaves or dishwasher tablets, attack each other, make drawings on the house or themselves, they refuse to eat vegetables and demand to eat jelly and

so on and on and on. These are also difficult to navigate because they require negotiating different values: to what extent should children be free to make their own decisions and face the consequences of them? What level of fighting is acceptable and what is too serious? And so on and on.

It is simply not feasible for someone to have a commitment, even an implicit commitment, as to what to do in each of these situations, for they are without end. When I leave my children with someone else, then, I need to find someone who has the judgement, motivation and competence to handle whatever transpires: someone who is not just reliable but also responsible.

Trustworthiness is not just about reliability but also about responsibility, and in particular taking responsibility: for a job, for a project, for an activity, for a person. Trustworthiness as responsibility is missing from accounts that are based on reliability, and so these miss out on some of its most significant features.

The world around us, the physical environment and the social world, is extremely complicated, changeable and surprising. This makes life interesting, but also presents significant challenges. In particular, setting out to achieve any project or major action is difficult, as we cannot confidently plan a route to our goal that will guarantee success, because we cannot be sure that nothing on which we depend will change in that time. One way of handling this challenge is to try to reduce our dependence on the outside world. In Aristotle, for instance, the best kind of life, the best happiness and flourishing, is self-sufficient (NE 1 6–9; Broadie and Rowe 2002).

But self-sufficiency is not really an ideal for humans who are limited in agency and dependent on one another. It is not an ideal of human life to withdraw from the world; we can live a much fuller and more exciting life if we instead engage with it. So it is not really possible or desirable for us to evade these challenges by turning inwards.

So what should we do instead? There are basically two approaches to the instability of human life. One is to try to make the world, at least the part around us in the scope of our action, more regular, and so more safe and predictable. We can do that by identifying reliability: both in inanimate objects that will do what we expect—the solidly constructed shelves—and in reliable people, with whom we can make agreements and who will do precisely what they agreed. We can rely on these people and these objects, in Holton sense's that we can act on the supposition that the people will fulfil their commitments and the shelves will stay secure.

But there is a second way of handling the complexity and instability of the world, which makes a greater use of the full agency of other people. We can ourselves accept that the world will never be completely reliable, and instead of trying to make it more predictable, we can instead attempt to become more flexible and responsive in return. We can make sure we have toolkits to repair objects that fail us, fall-back plans in case the bicycle or car breaks down or

won't start, or the pay does not come through on the day it was expected. We can do all this ourselves, but we have limited agency and to carry out more complex plans, we need to be able to make use of other people. Trustworthy people are those to whom we can entrust those plans, projects, objects or people, and who will do a good job of taking responsibility for them.

Thus we can define trustworthiness as responsibility:

X is trustworthy iff X is disposed successfully to take responsibility for valuable items entrusted to them.

A person may be trustworthy in a qualified sense, with regard to certain people, with respect to certain items.

X is trustworthy with regard to Y with respect to item I iff X is disposed successfully to take responsibility for item I with regard to person Y.

What does it mean to 'take responsibility' or to be 'given responsibility'? Suppose that I give you responsibility for a task, say, looking after my children. If you take on that responsibility, then you take control of that task and you manage it. What that means depends on the task. But in general, control or management may require a combination of theoretical and practical inquiry, and action, judgement and emotional responses. For instance, you might need first to understand the task itself: what exactly are you being asked to do? In what ways could the task be carried out and what would be the advantages and disadvantages of any of these methods? As we have seen, one of the points of taking responsibility as opposed to carry out a commitment is the open-ended nature of what you are taking on, so it will often be unclear exactly what the task even is and what would count as carrying it out well.

Once you have decided that, you need to choose and carry out a plan. Again, as this is typically open-ended and unpredictable, you need to manage the progress of the plan. In part this is a practical matter: you need to check how things are going and make changes where necessary to ensure success. But there is also likely to be an emotional component. Before the task is complete, you may well feel anxious as to whether it will succeed, worried that you have made a mistake, disappointed when things go wrong, and fearful as to whether you can put them right. Part of taking responsibility is managing these emotions yourself. You might be able to carry out all of these stages yourself, from the initial inquiry to the completion of the plan; but you might have to delegate some parts of the plan to others. Nevertheless, it is up to you to make sure that they do what is needed so that the task is finally done.

Often if you take on responsibility in the sense of taking control of a task, you will also be then be praised for success and blamed for failure, that is, you will also be held morally responsible for that task. But this is not always the case, and being held responsible is not the primary function of taking responsibility. Suppose that I gave responsibility to you and you failed. It is not always the case that all the blame should go to you. Perhaps I should at least share it, for having chosen badly, misplacing trust. Or perhaps most or even all the blame

should be mine. If I give responsibility for looking after my youngest child (3) to my oldest child (5), they are not to blame if things go wrong: all the blame is mine.

A responsible agent, in the sense required here has the right values, has good judgement and is competent, in such a way that she is able to respond flexibly to unforeseen and unforeseeable events and make generally good judgements about what to do in response to them and so act well. Because these circumstances are not predictable, and because often more than one response is reasonable, it typically won't be possible to rely on them doing anything in particular, other than the very general: doing a good job. Nonetheless, these people are trustworthy, and it is a kind of trustworthiness at least as valuable, and perhaps more so, than those who simply carry out their commitments.

Could a defender of the commitment account argue that this idea of trustworthiness as responsibility is not actually different from trustworthiness as commitment? For instance, in the childminder example above, could we not understand the trustworthy childminder as someone who has committed to looking after my children well and carries through that commitment? Of course carrying out any commitment will require judgement and competence, and carrying out a more open-ended commitment, like 'I commit to doing my job well' or 'I commit to looking after your children properly' will require all the qualities we identified for responsibility—the right values, judgement, flexibility and so forth.

It is true that to some extent the commitment account can incorporate responsibility as well as reliability through commitments that are open-ended rather than specific. But there are at least a few places where the two will pull in different directions.

In the first place, according to Hawley's account, you can be a trustworthy agent, that is an agent who is disposed to fulfil her commitments, by refusing to take on commitments in the first place. Thus according to Hawley, you could be trustworthy with respect to looking after my children by consistently refusing to look after them. This, she notes, is not how we normally think of trustworthiness. Here, the responsibility account does better. It is not possible to take responsibility for my children by refusing to look after them. Of course, responsible agents might turn down commitments, and very much for the reasons outlined by Hawley, to ensure that they are not overcommitted. But part of being suitable for taking responsibility for a job, project or person, is a willingness to take responsibility for it. And more generally, trustworthiness in the sense of taking responsibility is demonstrated in taking on commitments, rather than in the refusal to do so.

Secondly, responsibility can outrun commitments, in the sense that responsible agents will act well even if they do not have an antecedent commitment. We have already seen that a responsible childminder will make good decisions even if she has made no commitment regarding what to do if the child attempts

to glue herself to the carpet. Similarly, a trustworthy colleague at work is not someone who carries out only her contract and no more: in fact, doing so is a well-known form of industrial action, work to rule.

Responsibility can outrun reliability, and it can also come into conflict with it. Suppose that I am unavoidably late in picking up my children from the childminder. A responsible childminder will not note that she has fulfilled her commitment at 5 pm and throw my children onto the street to fend for themselves, rather she will look after them until I arrive, despite having no (antecedent) commitment to do so. In fact, she will do so, even if doing so means that she fails to fulfil another of her commitments: perhaps she has promised to meet her friend. It follows that reliability and responsibility can come into conflict and reliability is not always the most important factor. Someone trustworthy can therefore be unreliable, for she will break a commitment for the right reasons (the reasons acknowledged and responded to by a responsible agent), which are not necessarily further commitments that she has made.

A defender of the commitment account might respond: the childminder and work to rule examples are unfair to the commitment account of trustworthiness. It might be true that anyone working to rule is fulfilling the explicit commitments they made in the contract with their employer. But there are implicit commitments too, or 'meta-commitments' to take on future commitments as they arise. Childminders implicitly commit not to abandon children, and employees implicitly commit to doing more than their contracts (Hawley 2014: 11, 2019: 77).

This defence is correct that there are sometimes implicit commitments as well as explicit ones, and reliable agents will fulfil both. I may never have explicitly confirmed with my childminder that she should not poison my children (indeed it would be a bad sign if such a conversation were necessary) but it is reasonable to think that in agreeing to look after them she committed not to poison them. But as we have already noted, the world is complex, varied and unpredictable. We cannot foresee all the challenges that might arise at work or with children. It is not in the least plausible that in some sense, we were already implicitly committed to some particular response to any and all of these unforeseen events, which as trustworthy agents we must then carry out. In fact, one of the most valuable aspects of a trustworthy colleague or childminder is that you do not have to have an explicit or implicit agreement with them to cover any and every eventuality. No such commitments are necessary precisely because they are trustworthy: you can trust them to make their own decisions and act accordingly.

If trustworthiness as responsibility cannot be fully understood in terms of reliability in carrying out commitments, perhaps the opposite is possible, that is, perhaps reliability can be understood as a kind of responsibility.

Responsible agents will often be reliable. But as we have seen, carrying out commitments can be in tension with responsible agency: sometimes responsible agents will not fulfil their commitments, even though they could. One

possibility is to see trustworthiness as having two separate dimensions, a dimension concerned with reliability; and another with responsibility. However, I think there is a good case that responsibility is the fundamental quality of trustworthiness, as responsible agents will characteristically fulfil their commitments, tell the truth and keep their promises but they will also characteristically fail to do so when there are very good reasons not to. In both cases, their reliability (or unreliability, for good reasons) is explained by their taking responsibility. So although trustworthy people will typically be reliable, trustworthiness itself is not a kind of reliability but a kind of responsibility.

IV. THE DIFFERENCE BETWEEN RESPONSIBILITY, OBLIGATION AND GOOD MORAL CHARACTER

In the literature on trustworthiness, there is an approach that is different from the standard reliability account; instead trustworthiness is thought of as a disposition to carry out obligations (Simion and Kelp 2022). The advantage of this is that it explains why people who do not do what is expected of them at work or as friends are not trustworthy, without having to suppose, implausibly, that those people are secretly committed to being good friends and colleagues. It might seem that the obligation account of trustworthiness has a strong affinity with my account of trustworthiness as responsibility. Surely an agent who carries out her obligations is responsible, and a responsible agent will carry out her obligations. Are the two accounts equivalent?

They are not equivalent, because carrying out obligations is not the same as taking responsibility. An agent can show responsibility through supererogatory action, which is good but not required: helping others whom they are not required to help for instance, as part of taking responsibility for them. Trustworthy people will deal flexibly and sensitively to situations where moral reasons are fairly minor, less than moral obligations, good but not required. Or by contrast, where the stakes are very high—the moral reasons are pressing but so are countervailing reasons of self-interest. A trustworthy childminder would make good decisions when faced with a very dangerous situation, e.g. where a child has fallen into a river with a strong current, where attempting to swim to save the child would be risky, where action would be beyond what is required of anyone (note that this does not imply that she would risk her life since it might or might not be responsible to do so, rather the point is that she would make a good decision, even when not in a situation where she is obliged to act).

This sort of objection to an obligation account of trustworthiness invites the response that perhaps trustworthiness is not (merely) a matter of carrying out obligations but of responding well to all moral reasons, that is, it is a matter of general moral virtue. A trustworthy person has a good moral character.

There are some attractive features of the 'good character' account. It fits with the natural view that trustworthiness is a virtue. It can make sense of the kind of trust where you trust the people around you—strangers to you—not to attack you, or to treat you with kindness. But the disadvantage of this account is that trustworthiness is no longer a distinctive quality. It is a matter of moral virtue in general, rather than itself being a specific moral virtue.

The responsibility account is not equivalent to general moral virtue. Someone may demonstrate responsibility in non-moral situations. For instance, in the workplace, a responsible agent does a professional job, using their values and judgement to make good decisions about whatever arises at work, without the need for constant supervision or guidance. Most jobs are not morally required and there are not normally moral reasons to do a good job: it is typically morally permissible to resign, to do something else, or (in most circumstances) to work to rule. If people are not doing a good job for moral reasons, then, do they have a different sort of reason to do so, a professional obligation? At least with regard to some jobs, it is plausible that people do think of themselves as professionally obliged to do a good job: they must, whether they would prefer to or not, and they would feel guilty if they fell short. For these jobs, being a trustworthy colleague might require you to carry out your professional obligations. But again, responsibility can be shown in professional behaviour that is good but not required: working longer hours, more diligently, treating students or patients exceptionally sensitively and so on. And it is not obvious that all domains in which one can be a responsible agent are domains where there are obligations at all: perhaps friendships, for instance, do not come with obligations, and it is questionable whether reasons relating to friendship are moral or non-moral. Still, one can be a trustworthy, responsible friend—or not. So the responsibility account is better than the moral obligation account and the general moral virtue account, because it allows for trustworthiness in non-moral contexts (at work, amongst friends).

Does it share the problem that the moral virtue account had, of being too general, and not identifying a specific kind of quality of character? If anything it is even broader than the moral virtue account, as it extends to moral and non-moral contexts.

I suggest that the difference between trustworthiness as responsibility and general good character (moral and non-moral) is that trustworthiness is especially relevant when something—a person, project, activity—is entrusted to another person. Trustworthy work colleagues can be appropriately trusted with their job. A trustworthy childminder can be entrusted with my children. Trustworthy people are people to whom in general, it is appropriate to entrust things. It is appropriate to entrust things to trustworthy people, because those people will take responsibility for them. To take responsibility is to ensure, as far as possible, that a satisfactory outcome is achieved, for instance by monitoring the progress towards the outcome, by taking appropriate action, or by ensuring

that someone else takes appropriate action, and so on. It therefore involves the values, judgement, flexibility and so on that we already outlined as part of responsibility. And it may involve taking on commitments, but also, as we have seen, to act responsibly even when they have made no prior commitment to do so.

V. THE NATURE OF TRUST

One of the advantages of understanding trustworthiness as reliability is that it fitted well with, and in fact was frequently presented as an implication of, a highly plausible view about trust. Hawley, for instance, began with an account of trust (and distrust) as a kind of reliance (Hawley 2014: 10), specifically, a reliance on an agent fulfilling her commitments (or failing to fulfil them), from which was derived her account of trustworthiness as reliability in fulfilling commitments.

The alternative account of trustworthiness sketched here, trustworthiness as responsibility, then has a problem. Trust as reliance is not a natural fit with trustworthiness as responsibility. To rely on someone to X is to act on the supposition that she will X. But as we have seen, there is typically no particular action that someone who is responsible will do. They have been entrusted with something, and they will take care of it appropriately and handle whatever unexpected thing arises, but precisely because the situation is unpredictable, though you can act on the supposition that they will make sure things go well, you cannot act on the supposition that they will do X rather than Y: you may have no way of foreseeing what X and Y even are.

So if responsibility is an essential part of trustworthiness, what is the right way to think of trust? What is it to trust someone, if not to rely on her in some way or other? I suggest that the right way of thinking about trust in a person as responsible, is to think of it as confidence in her, either confidence in entrusting something in particular to her—a job, a child—or confidence in her in general, as a person, to take responsibility for whatever is entrusted to her.

What is it to have confidence in someone? It is not one single attitude, but a group of cognitive and non-cognitive attitudes and dispositions to act. It is to feel confident in entrusting children to her care; not to worry, to feel a relaxed sense of freedom when you drop them off at her door. It is to believe confidently, with no significant doubts, that she will successfully meet any challenge that looking after children raises. If you have confidence in her as a childminder, you might give some general instructions as to what to do, but you would not feel it necessary to give highly detailed and specific instructions for her to follow. In fact, doing so might even be counterproductive: it would be better for her to rely on her own judgement in whatever situations arise

than to try to follow a set of rules that might or might not be well-fitted to the circumstances.³ Nor is it necessary for you to monitor what she is doing, to keep checking up or requiring her to report back constantly. There is no need for constant surveillance, or indeed, virtually any surveillance at all.

If this is trust, what is distrust? And is there a difference between distrust and having no attitude of trust or distrust towards a person? Distrust is when you lose confidence in a person and doubt her. If you doubt your childminder, you question whether she will meet the challenges she might face successfully. You feel anxious leaving children with her and worry about them in her care. You do not have a strong credence that she will do a good job. You are inclined to monitor her, to check in regularly, perhaps unexpectedly, to see whether everything is all right. You are inclined to ‘micromanage’, to try to think of situations she might face and instruct her what to do in them all.

Note that distrusting her in this sense is compatible with trusting her in the Hawley sense of relying on her to meet her commitments. In distrusting her general ability to adapt appropriately to whatever arises, you have to as far as possible anticipate them and turn them in to explicit commitments. Of course, someone can be untrustworthy in both senses. Even if it were possible successfully to set out a complete and final set of rules for being a good childminder, and your candidate agreed to them, she might fail through incompetence or malice (or simply exhaustion from trying to remember all the rules) and so be unreliable as well as irresponsible.

So you can have an attitude of confidence in a person as a childminder, a colleague, a friend, or in general, as someone to whom things can safely be trusted. Or you can have an attitude of doubt. Or you can entirely neutral, with no particular belief about how they would fare if entrusted with your children, no attitudes of anxiety or calm, no view on whether or not micromanaging would be appropriate, nor even whether they would stick to a set of agreed rules or not.

Can you have confidence in an inanimate object? It may seem that you obviously can. For instance, you can be confident that the shelf won’t collapse. But thought the difference is subtle, I take it that this is a different attitude to having confidence in a person. It is not, for instance, completely natural to say that you have confidence in the shelf, rather than that you are confident that it will not collapse. Your confidence in someone as a person depends on what you take to be their abilities to take responsibility (or refuse to, or fail to) when entrusted with something; neither confidence nor doubt in this sense is appropriate to a shelf. So this account of trust meets Hawley’s criterion of having an adequate account of distrust, and adequately distinguishing trust from attitudes which we can have towards inanimate objects.

³ This is a theme of Onora O’Neill’s work on trust, e.g. O’Neill (2020).

Finally, one of the markers of trust, according to Hawley and many others in the literature, is that there are distinctive reactive attitudes when one trusts a person and they turn out not to be trustworthy. One does not merely feel disappointed or frustrated, but betrayed. Trust as confidence can make good sense of these feelings of betrayal. If I have confidence in you and you fail me, you have betrayed my confidence. A betrayal of confidence in ordinary language means that I entrusted you with a secret, and you did not keep it. I suggest that we can understand a betrayal of confidence in a broader sense too, whereby I entrust you with anything important to me, and you fail to look after it properly. Note that you betray my confidence in the everyday sense, if you tell my secret, even if you did not commit implicitly or explicitly to keep it. So this is different from Hawley's account of a betrayal of trust, which always involves the failure to carry out a commitment. Both seem to me plausible as a kind of betrayal. So it is legitimate to try to explicate a kind of trustworthiness by means of either kind of betrayal, the betrayal of a commitment, or the betrayal of confidence.

VI. TRUSTWORTHINESS AS RESPONSIBILITY AND WRONGDOING

So far we have considered trustworthiness in contexts which are morally neutral (a trustworthy colleague at work) or morally good (a trustworthy childminder). But is it possible for morally bad people, entrusted with some morally wrong job, to be trustworthy?

In everyday language, it might seem as if they can. In fact, we might think that trustworthiness is a very important quality in any society, including a criminal gang. Doesn't the success and longevity of a criminal gang depend on the trustworthiness of the gang members, thriving when they are trustworthy and falling out when they are untrustworthy? A valuable gang member can be entrusted with criminal activities (burglary, extortion) without constant supervision, allowed to use their own judgement when and how to carry them out, since they have the 'right' values and will devise a reasonable plan whatever circumstances they face. Of course the 'right' values here are morally wrong: valuing profit over harm, accepting or even enjoying cheating, manipulation perhaps even assault or murder. Nevertheless gang members might well value what they take to be responsibility in a potential gang member. To anyone outside the gang, however, the same person could not be entrusted with anything important, and so would not be trustworthy.

What should we make of a 'responsible', 'trustworthy' gang member? One option is to distinguish trustworthiness in a respect (X is trustworthy towards Y with regard to Z) from full or general trustworthiness (X is trustworthy).

The gang members are trustworthy in a qualified way, because they can be entrusted with criminal activities by each other. They are not trustworthy in other respects, to other people, and so are not fully or perfectly trustworthy.

Another option is to deny that gang-trustworthiness is a proper sense of trustworthiness, since responsible agency requires you to make good judgements, and gang members who carry out criminal activity are not making good judgements.

The advantage of the first option is that it captures an intuitive sense in which even criminal societies rely on trust and correspondingly, on responsible agency. This is very important.

But the disadvantage is that it threatens not to distinguish morally good and morally bad responsible agents adequately. For just as the gang member is trustworthy in some sense (to members of the gang) and not in others (to outsiders) and is not trustworthy without qualification, so outsiders will be trustworthy (to others) but not to the gang (as they will refuse to carry out criminal activity) and so will not be trustworthy without qualification.

If we do choose to distinguish these in terms of (full) trustworthiness, it might be best to deny that criminal gangs really are trustworthy, even when they are loyal to each other.

In support of this, I suggest that by the very nature of the values they hold and the kind of judgements they make, it will be unusual for criminal gangs to be trustworthy even to each other. Since gang members tend to value profit over harm to others, they are likely to choose to cheat each other if the opportunity arises, and so require at the least close monitoring and checks of loyalty. It would therefore be rare for criminal gangs to have genuine confidence in each other, and to be the appropriate recipient of such confidence.

Therefore, even if we allow that in principle, agents who are engaged in immoral enterprises can be responsible and so trustworthy, at least to their conspirators, in practice any confidence that they inspire is likely to be temporary and by its very nature fragile.

Reflection on these sorts of examples illustrates a final important point: that just because someone has entrusted something valuable to you, it does not follow that you should take responsibility for it. You would not overall be more trustworthy by being prepared to do so: that is, entering into a criminal conspiracy where you are completely loyal to the gang might make you trustworthy from their perspective, but you would fail to be trustworthy with regard to others, for instance, you would not be a suitable person with whom to entrust their life savings. Since there are many more people outside the gang, and many more valuable activities, people and objects outside it, to be trustworthy within the gang but not outside it would overall reduce your trustworthiness.

VII. PART 2: IS TRUSTWORTHINESS AS RESPONSIBILITY A MORAL VIRTUE?

Moral virtues are excellences of character. A trait of character is a multi-track disposition, a set of dispositions for action, choice and non-cognitive attitudes. The virtues are sets of dispositions that are a moral ideal: a credit to the person who has them, praiseworthy and admirable. They are traits of character that, other things being equal, generally speaking improve your life. This is not to say that there are no occasions when you might overall benefit from an unjust act, for instance, even though justice is a virtue. Rather that, in general, overall in a life, being just tends to make it go better. As a consequence, we hope that our own children to be just and we try to bring them up to be just, not only for the sake of ourselves or other people they will live with, but for their own sake.

In addition, moral virtues are connected to the challenges, temptations and obstacles distinctive of human life. We are limited creatures, dependent on each other, and vulnerable to harm both from the natural world and from each other. The virtues are qualities that help us do well in this sort of life, living well together and, as far as we can, facing and overcoming these challenges. Specific virtues are related to acting well in a particular domain or contrary to particular kinds of temptation. Thus beneficence, for instance, is the virtue connected to treating other people well who need help; justice is the virtue connected to treating people fairly, fulfilling promises and contracts and so on. Both of these virtues are very plausibly essential to any good human life, as we will always need to help each other and always need to make agreements with each other.

Trustworthiness as responsibility clearly meets these conditions. It is a disposition to take responsibility for whatever is (appropriately) entrusted to you. It requires the right kind of values, motivations and judgements. And as we have seen, it is an indispensable quality for people like us whose agency has limits, and whose surroundings and society are changeable and unpredictable. Since we are not self-sufficient and cannot carry out all our plans and projects by ourselves, we need to entrust parts of those plans, or objects or people that are valuable to us, to others. One way of doing so is to trust people who are reliable at fulfilling their commitments. This allows us to act on the supposition that they will do what they agreed, and so increases the predictability of our social world. This is trustworthiness in Hawley's sense. But another response is to entrust people who are responsible, who will act well whatever arises. This does not make the world more predictable, except in the sense that we predict that they will act well. But it does mean that we don't have to monitor ourselves the changing world and the right response for the valued project or person. We can leave it up to them, and devote our limited time and attention somewhere else.

Obviously it is enormously beneficial to live with people who are responsible in this sense. And equally obviously, it is good for each of us to be responsible in this way, as we are able to deal flexibly and sensibly with whatever challenges we face in pursuit of our goals. It is very plausible, then that trustworthiness as responsibility is a virtue.⁴

Why then have so many denied that trustworthiness is a virtue? One problem is, as we have seen, a focus on reliability instead of responsibility means that the kind of trustworthiness that has the profile of a virtue has been ignored in favour of something that appears to be at most an executive virtue.

But there have also been some arguments against trustworthiness as a virtue. For instance, Baier remarks that 'Women, proletarians, and ex-slaves cannot ignore the virtues of . . . judicious untrustworthiness' (1986: 253). How could trustworthiness be a virtue if its opposite (untrustworthiness) is sometimes a virtue? Proper vices, like injustice, cowardice and cruelty are never virtues. But Baier is not suggesting, and nor would it be very plausible, that the character trait of untrustworthiness is sometimes a virtue. Being, quite generally, unwilling or unable to take responsibility (or to be reliable) is not good for you, or others and is not admirable, even amongst the socially disadvantaged. It is much more plausible that particular acts that would be characteristic of the untrustworthy might be morally good, even prudentially or morally required. If you are asked to take on much more than your fair share of burdens, as women for instance frequently are, it is reasonable to refuse. And you might even be morally required to refuse if you are asked to take on something morally wrong. But in this respect, trustworthiness is no different from other virtues. You might be morally required to do what the cruel would do (refuse a benefit) or what the cowardly would do (refuse to face danger), in an individual instance, for very good reasons. This does not undermine the claim that courage or beneficence is a virtue, and the same is true for trustworthiness.

Jones acknowledges that you can sometimes be required to act in a way that is not virtuous, but presses a related, but she thinks much more serious problem for trustworthiness. She argues that trustworthiness can be inwardly 'riven' and self-contradictory, so you can be required by trustworthiness itself to be trustworthy (2012). For example, she has become involved in a community group that wants to preserve the local heritage, and another which promotes social inclusion and diversity. One group wants her to block a proposal for a high-rise development, the other to support it. 'Whatever I do, I must let down one group; whatever I do, one group will rightly feel betrayed' (2012: 83).

But now we have the responsibility account of trustworthiness, it is not so clear that she must let down one group and that whatever she does, one group will feel betrayed. For unlike the reliability account, whereby she commits

⁴ The ability to make flexible, sensible judgements is a kind of understanding, and this understanding is a component of moral virtue (as I have argued elsewhere, e.g. Hills 2009, 2016).

herself to doing some specific action—to support or to crush the proposal or to support it—a responsible agent could be entrusted with local heritage and social inclusion by finding a route through this difficult situation, for instance by modifying the proposal, or by finding another place for development, or similar. A responsible agent, like Hawley's careful promisors, would take care not to commit herself specifically both to support and not support the same proposal.

Nevertheless, even if you take great care, you can find yourself having taken on commitments or taken responsibility for things that you cannot carry out. So though trustworthy agents will do their best to avoid it, it is possible for trustworthiness to be 'riven'. But although Jones suggests that this problem is unusual perhaps even unique to trustworthiness, the same issue can arise for other virtues. Since there are many more needy people than I have resources to help, acting beneficently towards some forces me to ignore the needs of others. If two people have an equally just claim to the same object, and it cannot be divided between them, acting justly towards one may require me to act unjustly towards another. So even if there are some situations in which trustworthiness towards some would require me to act in an untrustworthy (irresponsible) way towards others, this is no reason to deny that trustworthiness is a virtue any more than we should reject the claims of justice and beneficence.

VIII. CONCLUSION

Trustworthiness is typically understood as a kind of reliability, i.e. it is thought that reliability is a necessary but not sufficient element of trustworthiness, as in Hawley's account of trustworthiness as doing what one has committed to do. This conception of trustworthiness captures the importance to us of making our unpredictable world easier to navigate by making it more predictable.

But I have argued that there is a different and more fundamental conception of trustworthiness as responsibility, where the central example is of entrusting something—a person, object, project or activity—to someone and having confidence that they will use good judgement to act well with regard to it. This captures a very different way of dealing with an unpredictable world, making use of the agency of other people who are themselves good at navigating it.

Trustworthiness as reliability is plausibly an executive virtue, but it does not have the profile of a moral virtue: it does not require any particular motivation and agents can be reliable with respect to any ends, good, bad or indifferent.

Trustworthiness as responsibility does have the profile of a moral virtue. It requires the right kind of motivation. It requires you to take responsibility only when doing so is morally right or morally neutral. It contributes to a good life and a well-functioning society. It is admirable and praiseworthy. And ideal

moral agent would be responsible, and we would have confidence in entrusting anything of importance to her.⁵

REFERENCES

- Broadie, S. and Rowe, C., eds. (2002) *Aristotle: Nicomachean Ethics: Translation, Introduction, Commentary*. Oxford: Oxford University Press.
- Baier, A. (1986) 'Trust and Antitrust', *Ethics*, 96: 231–60.
- Goldberg, S. C. (2020). 'Trust and Reliance', in Simon Judith (ed) *The Routledge Handbook of Trust and Philosophy*, 97–108. New York: Taylor and Francis.
- Hardin, R. (2002) *Trust and Trustworthiness*. New York: Russell Sage Foundation.
- Hawley, K. (2014) 'Trust, Distrust and Commitment', *Noûs*, 48: 1–20.
- . (2019) *How to Be Trustworthy*. Oxford: Oxford University Press.
- Hills, A. (2009) 'Moral Testimony and Moral Epistemology', *Ethics*, 120: 94–127.
- . (2016) 'Understanding Why', *Noûs*, 49: 661–88.
- Holton, R. (1994) 'Deciding to Trust, Coming To Believe', *Australasian Journal of Philosophy*, 72: 63–76.
- Jones, K. (1996) 'Trust as an Affective Attitude', *Ethics*, 107: 4–25.
- . (2012) 'Trustworthiness', *Ethics*, 123: 61–85.
- McLeod, C. (2021) 'Trust', The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/fall2021/entries/trust/>.
- O'Neill, O. (2020) 'Questioning Trust', in Judith Simon (ed) *The Routledge Handbook of Trust and Philosophy*, 97–108. New York: Taylor and Francis.
- Potter, N. (2002) *How Can I be Trusted? A Virtue Theory of Trustworthiness*. Lanham, MD: Rowman & Littlefield.
- Simon, M. and Kelp, C. (2022) 'What Is Trustworthiness?', *Nous*.

University of Oxford, UK

⁵ I would like to thank Hallvard Lillehammer, Rachel Fraser and John Tasioulas for very helpful comments on an earlier draft.