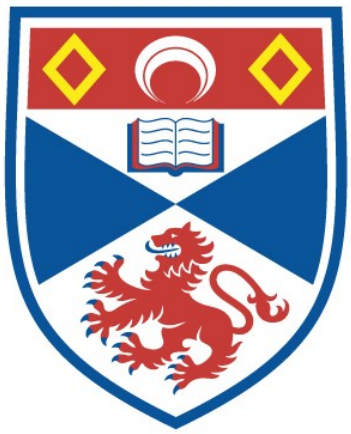
THE NATURE AND RATIONALITY OF TRUST AND TRUSTWORTHINESS

Thomas James Arthur Mitchell

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2023

The Nature and Rationality of Trust and Trustworthiness

Thomas James Arthur Mitchell

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy (PhD)

at the University of St Andrews

October 2022

Abstract

Trust and trustworthiness are highly important concepts; Trusting well enables us to achieve things we cannot do alone and to discover things that we cannot check for ourselves. Issues of trust sit in the intersection of ethics and epistemology; it is commonly thought of as a moral good to be trustworthy and trusting the testimony of others can be a valuable source of knowledge.

By trusting too easily, however, we become hopelessly naïve and risk falling prey to scams and lies. Knowing when to trust is therefore of significant practical value.

The aim of this thesis is to come to a greater understanding of these concepts. The first three chapters deal with what trust and trustworthiness are. In the first, I argue that trust is a kind of reliance, rather than a kind of belief. In the second, I argue that trust is specifically reliance on another to be trustworthy. The third chapter presents a theory of trustworthiness according to which being trustworthy is a matter of keeping commitments.

The following two chapters are about the rationality of trust. The first considers various ideas proposed for what justifies trusting someone and shows why they are not successful. The second lays out my own view, on which trust is justified by both evidence and practical reasons addressing another's trustworthiness.

Finally, I consider the relationship between commitments and responsibility. It is argued that, by accepting a commitment to do something from another, one is absolved of responsibility for ensuring that it is done. This is also found to help explain why holding victims responsible is a mistake.

Ultimately, a holistic theory of trust and trustworthiness is proposed, which I believe has significant advantages over those already in the literature and will be useful in future philosophical research.

Candidate's declaration

I, Thomas James Arthur Mitchell, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 80,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in October 2018.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 23/10/2022 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 24/10/2022 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Thomas James Arthur Mitchell, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 23/10/2022 Signature of candidate

Date 24/10/2022 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Thomas James Arthur Mitchell, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 23/10/2022 Signature of candidate

General acknowledgements

There are many people to whom I owe a debt of gratitude for support, inspiration, and encouragement throughout my time as a PhD student and before. I consider myself extremely fortunate to know those mentioned here, as fellow philosophers, as mentors, as family, and as friends.

First and foremost, my thanks go to Professor Katherine Hawley, my original supervisor. She is the reason for my choosing to study at St Andrews, as I knew her to be one of the best philosophers working in the field of trust. More than that, she was a patient and kind teacher, generous with her time and her knowledge. She encouraged; she challenged; she pushed my half-baked ideas to be the best they could be. I often left our meetings knowing that I needed to improve, but never feeling downcast. I hope one day to follow her example as I make my way into the academic world. We miss you, Katherine.

I would also like to thank Justin Snedegar, who supervised much of this thesis. Reading through my often overlong and overdue chapters, bearing with my numerous tangents, and highlighting various key points that I had missed, his help has been invaluable. Thanks also to my second supervisor, Philip Ebert, who gave me the benefit of his wisdom where I was out of my depth. I have relied heavily on the forbearance of my supervisors while writing this thesis, and they have done more than I had a right to expect.

Writing a PhD thesis can be a solitary experience, and so thanks are due to those who have kept me company in times that might otherwise have been rather lonely. My fellow philosophers, who have shown solidarity and understanding, and especially Katherina Bernhard and Matt Jope, for stimulating discussions about trust. The organisers of the Thesis Boot Camp in April this year, at which a significant portion of the thesis was written up and which provided its participants with company and motivation. The community at St Andrews Baptist Church, in particular Tricia Tooman, who understands the PhD experience only too well.

Special thanks go to my girlfriend, Rachel Elsey, who has been a source of mutual support over the last four years. She is always keen to be there for me, whether it be helping with house moves, long evenings of writing and re-writing, or just keeping me company over interminable cups of tea and coffee (and maybe the odd whisky). Thanks for putting up with me, Razza.

Thanks also to the rest of the Elsey family. To Caroline, for always welcoming my frequent visits; to James, for good company and friendship; to Fern the cat, for occasionally showing affection; and to Helen, the embodiment of love and kindness.

My family have constantly supported me, through this PhD and long before. I cannot express enough gratitude for all they have done for me, but here are some special mentions. To my brothers, Luke Wallace and Joel Mitchell, who first sparked my interest in philosophy years ago and have helped maintain it ever since. To my parents, Andy and Fiona, who have always encouraged me and to whom I can always turn for advice. To Archie, our beloved dog, whom we greatly miss.

Finally, I would like to thank Grandma Lawrie, Grandma Lydia, and Great Aunt Debs, for always showing an interest in whatever I've been up to. Thanks in advance also to Phoebe Wallace, my 21-month-old niece, who will help me push the 'submit' button.

Funding

This work was supported by a postgraduate scholarship funded by the AD Links Foundation.

Contents

Introduction

This thesis is a philosophical inquiry into the nature of trust and trustworthiness. These are important concepts whose exploration is of great interest to anyone who would understand how we ought, and how we in fact, behave towards one another. Well-placed trust comes with numerous advantages; cooperative endeavours, information sharing, and personal relationships all work better with trust. However, it also comes with certain risks. In trusting, one makes oneself vulnerable to being taken advantage of. Trusting poorly can result in being betrayed, tricked, and believing false information. We therefore all have an interest in others being trustworthy. Trustworthiness can be seen as the counterpart concept to trust, that which makes the risk of trusting someone pay off. We cannot expect everyone to be worthy of our trust, but we want there to be enough trustworthy individuals that we can usually trust others in our daily lives. Trust and trustworthiness generally underpin cooperative and personal relationships. Understanding these ideas is therefore of interest to us all.

Issues of trust also have implications for various topics that have long intrigued philosophers. The ramifications of making a promise, the nature of assertions and testimony, and the moral and rational rectitude of cooperative behaviour are all connected to trust. As I will argue in due course, trust is also connected to the highly important matter of responsibility, albeit less directly. The subject that this thesis deals with therefore has a prominent, though often unacknowledged, place within philosophy, situated in the intersection of ethics and epistemology.

In what follows, I will consider both what it means to trust someone and what constitutes trustworthiness. These are closely connected questions, since trust and trustworthiness are closely interrelated concepts. I see them not only as counterparts to one another, but as two parts to a conceptual whole. One cannot fully understand one without also grasping the other.

The construction of this thesis is therefore somewhat analogous to putting together a jigsaw puzzle. The positioning of each piece is determined by its relation to the others. A given point is justified partly by how well it fits with the theory as a whole. Furthermore, it can only be fully appreciated when seen in the context of the completed picture – only when it is finished does either the whole or the part become fully coherent. For this reason, it will sometimes be necessary to anticipate certain points, assuming them in advance of the place at which they are more fully justified and explained, although I have tried to minimise this.

This notwithstanding, the thesis has been written as a compromise between a collection of papers and a single monograph. Each chapter builds on those that have gone before, but I have also tried to make each understandable on its own. A reader can therefore read only the chapters that take their interest, or the thesis in its entirety. This necessitates a certain amount of repetition; points important for multiple chapters are explained again as required. I have tried

to do this in such a way that it will not be tedious to the reader, and hope that it will be helpful to be reminded of various ideas as they become relevant in different contexts.

Most philosophers addressing questions of trust – we are a small but growing group – assume that we can trust other people, but not objects. This is a view that I embrace. Of course, if someone talks of trusting their car or some other machine, we will usually understand them and not consider them to be talking complete nonsense. But we will take them to be talking in a metaphorical sense. Consider the attitude of blame, for instance. We can blame someone for behaving badly. We might also talk of blaming excess salt for high blood pressure. But we do not mean the same thing. If, in the latter case, someone offered to prosecute salt on our behalf, or asked us what kinds of punishment should be inflicted on salt crystals, or whether we thought sodium chloride might eventually be rehabilitated, we would be unlikely to take them seriously. The 'blame' we place on salt is an anthropomorphism; what we mean is that it is the salient cause of an undesirable effect.

Something similar is occurring when we 'trust' non-persons. Although we can make sense of the locution, we do not take someone who says that they trust their car to mean the same things as when they say that they trust their spouse, their friends, or their doctor. Similarly, only persons are assessable in terms of trustworthiness. This will be one of the basic assumptions of the thesis.

Another common idea that I shall be adopting is that trust is a three-place predicate. We do not merely *trust*; we do not merely trust *someone*; we trust someone to *do something*. It therefore takes the form 'A trusts B to ϕ', where A and B are persons and ϕ is a behaviour. It is this predicate that I shall be analysing in the following chapters.

The structure of the thesis is as follows. The first three chapters examine what trust and trustworthiness are. I have mentioned that I do not think these questions are fully separable, but I have tried to draw them apart as neatly as possible for ease of exposition. The next two chapters are about the norms of trust. Can trust be rationally justified? What is the role of evidence in doing so? Is it reasonable to take practical reasons into consideration? The final chapter moves on to a different ethical application of the central concept of commitment, discussing its implications for responsibility in risk-taking. The aim is to give a holistic theory of trust and trustworthiness and in so doing demonstrate the philosophical importance of the concepts.

I begin with a fairly general question about trust: is belief essential to it? Though it is not often explicated, there is substantial disagreement among philosophers on this matter. In proposing a theory of trust, some will point out that it seems nonsensical to trust someone without believing what they say and proceed to give a theory on which trust is a kind of belief, along with various other conditions. Others will point out that it seems possible to choose to trust and that we often trust others because we consider trust to be valuable, or we think they deserve a second chance, or some other reason unconnected to what the evidence tells us. Since these are not the reasons of belief and we cannot directly choose to believe, they proceed with the view that trust is some other kind of attitude, usually reliance. This first chapter tackles the controversy, weighing the arguments proposed by various philosophers against one another. All are found to have problems that make them unsuitable as accounts of trust, although it will be found that Richard Holton's approach comes closest. I believe that his take on the problem is on the right track, but does not fully succeed. Having identified its problems, I present an

improved version. It will be argued that trust is a kind of reliance and that reliance entails belief in the specific case of relying on another's testimony.

Chapter 2 takes up the argument by asking what distinguishes trust from ordinary reliance. This is a question that all reliance-based accounts of trust need to address and so several such accounts are considered. These are discussed in what I take to be ascending order of plausibility, but none are found to be completely adequate. Building especially on Katherine Hawley's work, it is argued that trust is reliance on another to be trustworthy. Ordinarily, the object of reliance is merely something happening or someone doing something; to count as trust, they must not only be relied on to do it, but to be trustworthy in doing so. This brings out the intuitive idea that trust and trustworthiness are intimately connected; one is built into the other. It also demonstrates the fact that trust is only appropriate for persons, not objects, since only persons are capable of being trustworthy. It is something that requires not only action, but agency – responsiveness to reasons. We can also use this to distinguish between genuinely trusting relations and relations characterised by manipulation, trickery, or even grudging cooperation. Trust intuitively requires something more and this view can explain it. Finally, the theory can be naturally extended to cover distrust. This is a concept that often receives less attention than trust, but a good theory of trust should also be able to account for its contrary.

The natural next question is that of what it means to be trustworthy, for this is required to fully understand the theory of trust presented in Chapter 2. It is this matter that is addressed in Chapter 3. Here, I again draw on the work of Katherine Hawley, whose writing on trustworthiness inspires much of this crucial part of the thesis. Like her, I take the view that being trustworthy is a matter of fulfilling commitments and, by extension, avoiding commitments that one will be unable to fulfil. I do this by starting with the assumption that trustworthiness is a specific kind of reliability that can only be had by persons, then analysing this concept of person-specific reliability. Something is reliable insofar as it does what it is meant to do; a person, being autonomous, can choose what they are meant to do. To do so is to make a commitment. If I promise to do something, then I am exercising my power as an autonomous agent to create an obligation on myself to do it. Both my approach and my conclusions are somewhat different from Hawley's. In starting with the concepts of reliability and personhood, I work upwards from a more basic level, not only presenting a theory but a philosophical foundation thereof. I also disagree with her on the status of those who lack commitments, and on that of those who try but fail to fulfil a commitment due to bad luck. This involves delving into what it means to have and fulfil a commitment, as well as what it means to be untrustworthy. I argue that if one lacks a commitment to do something, then one is neither trustworthy nor untrustworthy with respect to that thing. Furthermore, people do not always succeed in making the commitments that they take themselves to have made; to be valid, a commitment must be voluntary and informed. The upshot is that being neither trustworthy nor untrustworthy is more common than might initially be thought.

The first three chapters, then, develop a theory of trust and trustworthiness. To be trustworthy is to fulfil one's commitments; to trust is to rely on another to fulfil a commitment. In the rest of the thesis, this theory is assumed, although for the sake of keeping the chapters to some extent independent of one another, I avoid drawing too heavily on what has gone before where possible.

Chapters 4 and 5 are about the norms of trust. In Chapter 4, I consider various ideas that have been offered by other philosophers of trust, dividing them into Evidentialist and Pragmatist views. The former take trust to be concerned with truth; as such, only the evidence should count when we are considering whether to trust someone or not. Trust is rational only insofar as it is likely that the trusted party will perform. The latter take trust to be justified by practical considerations as well; the extent to which we can expect to benefit from cooperation, or the value that we place on living and working within an atmosphere of trust, for instance, may be taken into consideration. Each kind is further subdivided and analysed in turn. I argue against all versions that are discussed, showing that they do not manage to grasp what justifies trust. Evidentialists are unable to satisfactorily explain away why it seems possible to trust for practical reasons, like the fact that cooperating with someone will be beneficial. Pragmatists tend to appeal to reasons that are concerned with the attitude of trust rather than its content, akin to justifying a belief by its being valuable or convenient rather than backed by evidence. I end this chapter by sketching out the features that a successful account of the reasons of trust ought to have.

Chapter 5 is more positive, proposing an account of the norms of trust that I think improves on those discussed in the previous chapter. Given the view for which I argue in Chapter 1, that trust is not a kind of belief, it will not be surprising that I reject Evidentialism. In this chapter, I present a nuanced form of Pragmatism, on which both practical and epistemic reasons can support trust, but not just any of each kind will do. For this, I draw on the distinction between object-given and state-given reasons. The key to this is to recall that the object of trust is not merely that the other will do the thing in question, but, as argued in Chapter 2, that they are trustworthy. Therefore, evidence that someone will act in a certain way does not favour trusting them to do so unless it is evidence that they are trustworthy with respect to that behaviour. Similarly, a practical reason that appeals to the value or positive consequences of trust itself does not actually count in favour of trusting someone, any more than being offered a reward for believing something would really justify that belief. Unlike belief, however, some practical reasons do count for trust; namely, those which address the other's trustworthiness. It is not rational for me to trust you because I think that trust is valuable and there should be more of it in the world; it is rational to trust you (in part) because your proving trustworthy would benefit me. This account therefore avoids the main problems faced by the theories of both kinds considered in Chapter 4.

Chapter 6 takes a different direction. Having established that commitments play a central role in trust and trustworthiness, I apply the concept in a different ethical arena. I address what I call the 'Responsibility Problem'. Usually, if someone takes an unnecessary risk and comes to harm as a result, then they are responsible for that harm. However, if the harm that they risked and which was manifested involved being victimised by someone else, then they do not bear responsibility. What explains this? Why does the involvement of another absolve one of responsibility for the harm that one voluntarily and knowingly risked? My answer is in terms of commitments. Part of the ethical import of a commitment is that it has the power to transfer responsibility. For instance, if I need something done and you promise to do it for me, I would not be responsible for it not being done if I accept your promise and you fail to follow through. Furthermore, I argue, we all have standing commitments to not harm one another, not just ordinary moral obligations to refrain from harming others. If this is so, then in the special case of risked harm being from another person, then the victim bears no responsibility; only the

perpetrator does, since they broke their commitment to non-harm. Thus, the idea of commitments has broader philosophical application than trust and trustworthiness. If the argument in Chapter 6 is correct, then it can also explain what is wrong with victim-blaming.

I hope that what follows will be considered interesting and thought-provoking. It draws on numerous ideas presented by various philosophers who have worked on trust and trustworthiness before. Although the theory that I present is, to my knowledge, original, it builds upon others' work in a way that makes this a continuation of the still-young tradition of the philosophy of trust.

1

Is Belief Essential to Trust?

Introduction

Trust is an important attitude. Well-placed trust enables us to learn from others what we could not discover for ourselves and facilitates co-operation that allows us to achieve what cannot be done alone. But it also leaves us vulnerable. Poorly placed trust will let others take advantage of us and make us generally naïve. But it is almost impossible to get by without it; we need the benefits that it can provide in order to live in communities. The questions of what trust is and when it is warranted are therefore of great philosophical interest.

This chapter is meant as a step towards a greater understanding of trust. It deals with one puzzling question that arises about its nature: does trust essentially involve belief? Clearly, belief is an attitude that frequently accompanies trust, but there do seem to be exceptions.

Suppose that you are a shopkeeper and find out that one of your employees was once convicted of petty theft (Holton, 1994, 63). Should you trust them with the till? There are a number of reasons why you might: it will improve your working relationship, facilitate their rehabilitation and simply contribute to the smooth running of your business. But none of these reasons have anything to do with the truth of the matter; they do not help you to answer the question of whether the employee would steal from you, given the opportunity. Furthermore, it seems that whether you trust the former thief is a choice, not an attitude you are compelled to (not) have.

This is the first part of our puzzle. Belief is something that we cannot directly choose and the reasons that justify it are epistemic reasons – considerations that bear on whether the proposition in question is true. To be sure, it is possible for us to make decisions that impact our beliefs. We may choose to examine certain pieces of evidence rather than others, for instance. But this is not the same as choosing to believe through an act of will, as one might choose to perform certain actions. It is also possible for us to believe for reasons that have little to do with the truth of the matter. Our beliefs can be the products of wishful thinking, or we might be persuaded by social indoctrination rather than by reason. But such beliefs are not justified. Yet it seems that we can directly choose to trust and that trust can be justified, at least in part, by non-epistemic reasons. You decide whether to trust your employee and make your decision for reasons that do not all bear on the truth. How is this to be accounted for if belief is necessary for trust?

Perhaps, then, trust does not entail belief but just often comes with it. But this would be too hasty, for there is a second part to the puzzle. Suppose that a friend tells you something that you are not in a position to verify yourself. This is a situation in which trust would seem

appropriate. But would it be reasonable to respond, 'I trust you in what you say, but I do not believe you'? Intuitively not; it seems nonsensical. So in cases like this, belief would appear to be necessary for trust.

It has often been acknowledged that there are at least two ways of trusting someone: we can trust them to do something and we can trust what they say.[1] Call the former 'act trust' and the latter 'testimony trust'. Of the two examples just given, the first is one of act trust; it is a matter of trusting someone to act in a certain way, in that case, to not steal from the till. The second is an example of testimony trust. Something to note is that, as the examples illustrate, cases of act trust and testimony trust suggest different answers to our question. With act trust, trusting seems to be a matter of choice and can be justified by non-epistemic reasons. It therefore indicates that belief is not essential to trust. With testimony trust, however, trusting seems to require belief.

This, then, is the puzzle. For either kind of theory of trust – those which involve belief and those that do not – there are some counterintuitive consequences. They may not represent insurmountable problems, but some explanation of them is owed by any theory's proponent. To get to the bottom of this issue, we will examine four philosophical views on trust in Part 1. For each, we will consider how they try to solve the problems associated with their kind of theory, the extent to which they succeed, and what can be learned from their attempts. Then, in Part 2, I will bring these lessons together to suggest my own view on what kind of attitude trust is. I will argue that it is a kind of reliance, rather than belief, but that this entails belief in the special case of trusting testimony.

Part 1: Theories of Trust

In this part, I will examine four theories of trust, focusing on how each deals with the problem of belief outlined above. I consider them not in the order in which they have been proposed, but in what I consider to be their order of accuracy. Thus, we will begin with a theory that I take to be straightforwardly incorrect, since it takes the side of belief being essential to trust without, to my mind, seriously engaging with the problems associated with that view. We end with a more nuanced account which, although it is not successful, is on the right track. The theories are also arranged so as to alternate between belief-based and non-belief-based views, with each improving on the one before.

Although none fully succeeds, there are lessons to be learned from them which will indicate what a correct theory should be like. This enables me to give my own view in Part 2.

Trust as Believing Beyond the Evidence

[1] See, for instance, Hawley (2014, 16), Hieronymi (2008, 219), and Holton (1994, 73).

In her 1987 paper 'Trust and Rationality', Judith Baker espouses the view that trust is a kind of belief. Her main supporting example is of a friend who, having been accused of a crime, protests her innocence and asks to be trusted.

> If I trust her in such a situation, I do not merely stand by her, acting in ways that support her, either materially or emotionally. I believe she is innocent. I do not, however, come to believe she is innocent, despite the evidence, by weighing or balancing present evidence against her past record. … I believe that there is an explanation for the alleged evidence, for the accusation, which will clear it all up.

(Baker 1987, 3)

This is a compelling case. If one is merely supportive of the friend, standing by her without believing her, she is likely to be disappointed. What really mattered to her was being believed. (Baker 1987, 6)

Baker goes on to claim that trust is a commitment to believe beyond the evidence. This is not limitless; it would be irrational and perhaps impossible to believe against hard and irrefutable evidence. But even so, trusting beliefs tend to be stronger than the evidence warrants and display a certain resistance to counterevidence. (Baker 1987, 3-5, 10)

As her view is that trust is a kind of belief, Baker must deal with the problems of choosing to trust and trust being justified by non-epistemic reasons. On both of these, she bites the bullet. According to Baker (1987), we can, in the case of trust, choose to believe and such beliefs can be partly justified by practical reasons. To some extent, therefore, believing against the evidence can be rational.

There are a number of objections that can be raised to this theory. First, it is hard to form and sustain beliefs for reasons one acknowledges to be non-epistemic. Having a certain belief might strengthen your relationship with a friend, giving you a practical reason for holding that belief. But if you realise that the evidence does not support it, it would be psychologically difficult to convince yourself of it – and even if you did manage it, the belief would likely be unstable, since you would know that it is not properly supported (Hieronymi 2008, 221). Second and relatedly, it is implausible that Baker has found the correct norms for trusting beliefs. It would not be correct to claim that a belief formed for very strong practical reasons but with little epistemic support is a well-justified belief. What lies behind both of these objections is the important point that belief is fundamentally concerned with truth.[2] This is something that Baker seems to lose sight of.

Thirdly, even on Baker's own terms, it is unlikely that she has the correct norms. She admits that the reasons for believing one's friend will be partly epistemic; trusting beliefs can be formed only after spending time with someone and observing their generally decent and trustworthy character (Baker 1987, 4). However, suppose that, having spent time with and

[2] Some philosophers, such as Stroud (2006), Reisner (2018, 705-28) and Rinard (2018), have argued that belief need not be about truth; that it can be rational to believe for non-epistemic reasons. I do not find such arguments very convincing; I take belief to simply be the attitude of taking a proposition to be true, so it is misguided to suggest that it might be justified for reasons unconnected to truth. This is something I will assume for present purposes. However, the view that I ultimately argue for is compatible both with practical reasons counting for belief and with such reasons not counting. Even if it can be rational to believe for practical reasons, this would not imply that my view is false, therefore, although it would harm some of my arguments.

gotten to know one's friend, one has discovered that she is not particularly trustworthy or decent. She is just the sort of person to commit this crime and then play innocent, lying even to her friends. The epistemic reasons would then lie heavily in favour of believing that she is guilty. If one's friendship has yielded evidence that the other person is untrustworthy, then it is doubtful that the friendship is a reason in favour of trusting them. Even if one treated her as innocent, giving her the benefit of the doubt, one would not (rationally) believe her. Once again, belief is about truth.

Finally, belief that is motivated by friendship is likely to be ineffective. Even if one can bring oneself to believe, the friend is likely to be disappointed if they discover that the only reason for the belief is the friendship itself. In wanting to be believed, she wants reassurance that what she has said is being taken seriously; that others consider it to be true. Instead, she is believed only to make her feel better. Believing her only for the sake of friendship will therefore be as disappointing as merely standing by her and supporting her.

What we can learn from Baker is that whatever trust is, it is not just belief. Belief is concerned with truth, so is supported by evidence. But if trust is supported by non-evidential reasons as well or even instead, then it cannot be (just) a matter of believing. Alternatively, if trust does require belief in cases like the one Baker describes, then we should not trust in those cases, since the evidence does not support belief. Even if belief is an essential part of an account of trust, there must be more to it, that will explain why we seem able to trust beyond and even against the evidence.

Trust as Either Belief or Acceptance

We turn now to a theory that does involve belief, but does not require every case of trust to be a kind of belief. Karen Frost-Arnold holds that belief is not necessary for trust, even if it frequently accompanies it. She identifies the central problems mentioned above with requiring belief to be essential to trust: trust is sometimes voluntary and is justified for non-epistemic reasons (Frost-Arnold 2014, 1959-60). Accordingly, she presents an account of trust on which belief is not necessary, which is motivated by what Frost-Arnold takes to be three genuine kinds of trust which do not involve believing that the other will do what they are trusted to do. These are *therapeutic*, *coping* and *corrective* trust. Therapeutic trust, a concept introduced by Karen Jones (2004, 5), is trusting someone with the aim of encouraging them to become more trustworthy, as one might trust a child. One does not believe that they will actually perform, but that is not the point; one is trying to inculcate an important value. (Frost-Arnold 2014, 1960) This is also a factor in the case of the former thief mentioned above; in trusting him, one might be attempting to 'draw him back into the moral community' (Holton 1994, 63).

Coping trust is trust for the purpose of simplifying matters and avoiding anxiety. This is unlikely to be appropriate when there is a great deal at stake, but, Frost-Arnold argues, we can sometimes choose to trust to prevent us from worrying about the possible negative outcomes (Frost-Arnold 2014, 1960-1). If we are concerned that an associate will let us down in some matter, we can choose to trust them for the purposes of avoiding the stress and anxiety that accompanies uncertainty. This will put our mind at rest, lightening our mental load and enabling us to focus on other things.

Finally, corrective trust is an attitude that we adopt in order to avoid committing testimonial injustices.[3] It involves deliberately and consciously giving greater weight to what another has to say so as to counter the effects of prejudice that one might have against them. (Frost-Arnold 2014, 1961-2) For instance, suppose that someone realises that they do not always take what certain people – those of a certain gender or ethnicity, perhaps – say as seriously as they ought to. It is hard to completely rid oneself of a prejudice, even after becoming aware of it. Therefore, in an attempt to become both a better knower and a more just person, they might decide to place greater weight on what those people say than they think is warranted. They recognise that their judgement is skewed and correct for it by choosing to trust.

One might wonder whether these count as good reasons to trust, or even count as cases of trust at all, as opposed to merely pretending to trust. Corrective trust in particular may seem odd, since it seems to require belief, or at least increased epistemic credence, yet is supposed to be responsive to non-epistemic reasons and can be chosen. Frost-Arnold does discuss these issues (2014, 1962-3; 1967-71), but it would take us too far afield to evaluate them here. They shall be addressed in chapter 4, which is devoted to the norms of trust. For now, I will accept for the sake of argument that one can trust in these cases and for these reasons.

None of these kinds of trust is grounded on evidence, and nor do they directly aim at truth.[4] Therefore, it seems that they do not require belief. Nevertheless, Frost-Arnold acknowledges that trust is sometimes involuntary, evidence-based, and does involve belief (2014, 1963). She thus finds herself with the problem that is the subject of this chapter. Her solution is to suggest the following disjunctive account of trust:

> A trusts B to ϕ iff A either believes or accepts that B will ϕ, and this belief or acceptance is the basis of A's practical reasoning.

(Frost-Arnold 2014, 1964)

To accept a proposition is to use it as a premise in one's practical reasoning. This means planning and acting as though it were true, which does not require believing it. To demonstrate this, she borrows an example from Michael Bratman (1992, 7): when planning a building project, one might use the highest estimates for costs, rather than a realistic estimate for what the various expenses are likely to be. When getting quotes from subcontractors, one assumes for the sake of planning that the highest price in the given range is what one will ultimately be charged. This is not the same as believing that the costs will be so high, but is a useful assumption to avoid going over budget. It is Frost-Arnold's view that, in the cases of therapeutic, coping and corrective trust, one accepts rather than believes that the other will perform (2014, 1965-66).

This does seem to be an improvement on Baker's view. Frost-Arnold is not committed to trust always being a kind of belief, so will not run into the problems which plagued the previous theory. Whatever we think of Frost-Arnold's own examples of non-believing trust, her account of trust enables us to say that we can trust without believing. We can therefore choose to trust

[3] See Fricker (2007, 17-29).

[4] Corrective trust arguably aims at truth, since its purpose is to correct for one's prejudiced beliefs. However, it is indirect, since its primary aim is to avoid doing the speaker an injustice, rather than to get to the truth. Not every instance must include belief, and it is not directly responsive to (what one takes to be) evidence, since one's perception of the evidence is supposed to be prejudicially skewed.

and do so for non-epistemic reasons. At the same time, there is room for trust sometimes being constituted by belief. This appears to mean that the problem of testimony can be avoided; in testimony cases, we can say that trust takes the form of belief.

But let us take a closer look at this. Since Frost-Arnold thinks that belief is not essential to trust, there being examples of trust that do not involve belief, the salient problem for her is that testimony trust seems to require belief. On the face of it, this is the advantage of a disjunctive account. Trust is something that seems to sometimes involve belief (in testimony trust) and sometimes not (in act trust). In having trust sometimes constituted by belief and sometimes mere acceptance, Frost-Arnold's view fits with this. Cases of testimony trust are a subset of those cases of trust which involve belief. For act trust, either belief or acceptance will do.

Unfortunately, having a disjunctive account does not entirely solve the problem. The puzzle is not simply that trust sometimes involves belief and sometimes does not. It is that trust sometimes *requires* belief and sometimes does not. In particular, we do not trust someone's testimony unless we believe it. Therefore, a successful solution must not only show that trust is sometimes constituted by belief and sometimes by some other non-believing attitude; it must show that cases of testimony trust are always constituted by belief. Frost-Arnold's view allows testimony trust to be belief, but it does not have the resources to non-arbitrarily require that it be belief. According to it, one might trust someone's testimony by merely accepting that what they say is true. That is, one may plan and act as if it were true without believing it, which is intuitively insufficient for testimony trust. Certainly the friend in Baker's example would be disappointed by such 'trust'; she wants to be believed, not merely have someone stand by her supportively, acting and planning as if they believe her.

So, this view falls into the problem of belief being required for testimony trust and does not manage to escape it. Nonetheless, progress has been made. It was mentioned at the end of the last section that trust cannot be just a matter of belief and Frost-Arnold's account fulfils that criterion by adding the idea of acceptance, a 'planning on the basis of' attitude that is voluntary and responsive to non-epistemic reasons. What is needed next is some way of excluding mere acceptance from testimony cases. If that can be done, we can embrace without contradiction the intuitive ideas that belief is not essential in act trust, but is essential in testimony trust.

Full-Fledged Trust and Mere Entrusting

One view that tries to achieve this is that of Pamela Hieronymi. She holds that trust is a kind of belief, on the basis of an example of an accused friend similar to that given by Baker (Hieronymi 2008, 219). Specifically, trust is belief based on the trustee's supposed trustworthiness (Hieronymi 2008, 224).

This is a fairly intuitive idea, but, like Baker, Hieronymi must answer to the problems associated with belief-based views. How is it that we seem able to choose to trust? How can trust be justified by non-epistemic reasons?

The first point to note is that, unlike Baker, Hieronymi takes seriously the idea that belief is fundamentally concerned with truth. According to her, it is irrational and difficult to believe against the evidence (2008, 221). Trust is therefore involuntary and justified epistemically. If one has good epistemic reasons for thinking another person trustworthy, then, in the absence

of relevant counterevidence, one will trust them. Similarly, if one lacks such reasons, or has reasons for thinking another untrustworthy, then one cannot trust them.

In order to account for the occasions when we seem to have a choice, or when trust appears to be justified by practical considerations, Hieronymi makes a distinction between full-fledged trust and mere entrusting. Full-fledged trust is the primary sort of trust that involves belief. Entrusting is not really trust, though it is sometimes colloquially called trust. It is a matter of acting and planning on a certain assumption, similar to Frost-Arnold's acceptance. (2008, 217-8)

Entrusting can be used in place of trust where there is insufficient reason to believe. It is the product of a two-stage process. Suppose you are invited to trust someone to do something – Hieronymi takes Richard Holton's (1994, 69) example of a trust exercise in which one must fall backwards with eyes closed to be caught by the others. In this case, you are invited to trust the person behind you to catch you. The first stage is your consideration of whether you believe they will catch you. If you do, then you will fall without further deliberation. But if you conclude that this is something you do not believe (though you do not specifically believe that they will not catch you), then there is the second stage: deciding whether to fall without the relevant belief. It is at this point that the apparent non-epistemic reasons for trust come into play. But if you choose to fall, on the basis of those reasons and without belief, then you are not really trusting, but merely entrusting. (Hieronymi, 2008, 216-9)

This two-stage process will not work when it comes to trusting someone's speech. Suppose that you are deliberating about your accused friend. You consider first whether you believe that they are telling the truth, as you considered whether your fellows in the trust exercise would really catch you. As in that case, you are uncertain; you neither believe nor disbelieve your friend. Can you now decide to trust her for non-epistemic reasons? According to Hieronymi, you cannot – or at least, it would not be rational to do so. The reason is the difference between act trust and testimony trust. In the trust circle, what is required is that you fall and this can be done with or without belief. What your friend wants, though, is to be believed. This is what trusting her would amount to. But you cannot rationally believe her once you have concluded the first stage and found that you do not believe her. We can entrust our bodies, but not our beliefs. (Hieronymi 2008, 219-21)

Since entrusting is not here counted as a kind of trust, the problems that were discussed for Frost-Arnold's view do not arise. We can act and plan on the basis of what others tell us, but this is not trusting them unless we believe them. But this represents a significant departure from the notion of trust as it is normally used. Even if we harbour doubts about another's performance, it appears coherent to talk of trusting them. If they let us down, it is still a breach of trust, no matter how little faith we had in them.

Hieronymi is aware of the counterintuitive consequences of her view. To defend it, she considers two rival ideas of trust, a more liberal notion, according to which both entrusting and full-fledged trust count as trust, and a purist's notion, according to which only full-fledged trust counts, so belief is required. The rationale for each of these competing ideas is the measure by which the extent of trust is judged. On the liberal notion, one trusts to the extent that one incurs vulnerability. The greater the potential loss associated with betrayal, the more one is trusting. On the purist's notion, one trusts to the extent that one trustingly believes. The greater one's degree of belief that the other will perform, the more one trusts. (Hieronymi 2008, 228-9) One

can see the appeal of both; we do sometimes use 'trust' in both ways, saying that we trust someone a lot because we greatly value what they have been trusted with, or that we trust them a great deal because we are very sure that they will perform. Hieronymi then presents a series of arguments for thinking that the purist's notion is the better one. Taking vulnerability to betrayal as a 'touchstone' for trust (2008, 215), Hieronymi argues that the purist's notion accounts better for the experiences of being trusted and being betrayed.[5]

Firstly, a betrayal is likely to be felt more keenly if one actually believes that the other will perform. Suppose you tell someone a secret, impressing upon them that they are not to tell anyone else. They agree, but later breach your confidence. Now, if you did not really believe that they would keep the secret, you will not be happy with the outcome, but the fact that this was not unexpected is some consolation. You told them the secret because you wanted to give them a chance, or because you were trying to become more open and trusting of others, or for some other practical reason. You took this decision having factored in the risk of their telling others. What they have done is therefore disappointing to you, but not shocking. You can still own that decision, knowing that you were not taken in by anything the other person said or did. In contrast, if you fully believed the other to be trustworthy and you never doubted that they would keep your secret, their breach of confidence will affect you more deeply. As Hieronymi puts it, 'Not only was [your] secret told, but [your] faith in [them] has also been broken.' (2008, 230) This will be a more shocking experience. Both might count as cases of betrayal, but the latter seems more seriously so. Hieronymi concludes that this is because there is more trust to betray. One is more vulnerable in the case of actual belief, so that is the more trusting case.

Secondly, one is more likely to feel distrusted if the other person does not believe that one will perform. Consider a variation on the same case, but this time from the other side. You have been entrusted with a secret and have agreed to keep it. This time, you faithfully keep the confidence, telling no one. Now, suppose you discover that the other person did not believe that you would keep the secret, but told you for one of the supposed practical reasons of trust. If you have not given them reason to doubt you, it seems that you might justifiably feel resentful at the lack of trust being displayed.[6] Insofar as this is a reasonable response, it seems that not believing entails not trusting – or at least, not fully trusting. (Hieronymi 2008, 230-1)

Finally, Hieronymi points out that a common strategy for avoiding the pain of being betrayed is to try to avoid belief (2008, 231). If we are keen to shield ourselves from the emotional agony of a betrayal of trust, we might tell ourselves repeatedly that the other person is not likely to perform. If successful, this will leave us unsurprised at a betrayal, so it will not be such a devastating experience. Of course, we may not succeed in convincing ourselves, but if we do, the pain of betrayal will be mitigated. Since avoiding belief in another's performance thus reduces our vulnerability to betrayal and vulnerability to betrayal is necessary for trust, it seems that we must believe that another will perform in order to be really trusting them.

[5] Hieronymi is not explicit about what she means by a 'touchstone', but her discussion in various places indicates that she takes vulnerability to betrayal to be a necessary condition for trusting someone (2008, 215; 219; 222). She is also clear that she considers it possible to be betrayed by someone without trusting them (2008, 229), so it is not a sufficient condition.

[6] What if you have given them reason to doubt, perhaps having given away secrets before? On Hieronymi's view, the lack of belief is still a lack of trust. We might colloquially call it 'trust' if you are entrusted with the secret, since it is the closest one can reasonably expect to get to being genuinely trusted, and one does not have the same basis for feeling resentful. But you are not really being trusted. (Hieronymi 2008, 218-9)

Hieronymi tries to show by these arguments that, all else being equal, one trusts only to the extent that one believes in the other's performance. However, I do not think that these are compelling arguments. The sense of shock and disappointment that Hieronymi considers characteristic of realising that one has been betrayed is at least partially explicable without appealing to trust. It is shocking to realise that something one had previously believed, especially if one did not even question it and it was a comforting or important belief, is false. So, disappointed trust that was combined with belief will be more shocking than disappointed trust that was not combined with belief. This does not mean that trust must always require belief.

At best, Hieronymi's observations show only that one's trust is stronger if one believes. There is a sense in which one does not fully trust the other person without belief. Thus, we might feel more betrayed if we believed; we might resent the fact that the other does not trust us as much as they might have done if they do not believe us; and we might shield ourselves from the worst emotional effects of betrayal by trying to withhold belief. This much does seem plausible. But all this is compatible with the view that we can have trust without belief and that trust with belief is merely a stronger kind of trust, which is also more in keeping with how we normally talk of trust. Below, when I present my own solution to the problem of trust and belief, I hope to clarify the sense in which trust with belief is stronger and more complete. For now, though, I do not think that Hieronymi is successful in showing that belief is necessary for trust. We can still complain of a breach of trust when we are let down by someone we doubted. They are not let off the hook by the mere fact of our unbelief, which may be well-founded. It is just that we trusted to a lesser extent, or our trust was of a less complete form.

Hieronymi and Frost-Arnold have very similar ideas. Both hold that 'trust' often refers to a kind of belief, but is sometimes also used for an attitude of acting and planning as if something were the case, without necessarily believing it. Frost-Arnold calls this acceptance; Hieronymi calls it entrusting. The main difference between their views is that Hieronymi thinks that this acting/planning-as-if attitude is not strictly a kind of trust, whereas Frost-Arnold thinks that it is.

This difference gives rise to a dilemma. If acting/planning-as-if does not count as trusting, then we face the problem just outlined for Hieronymi: it is hard to reconcile this with our ordinary use of the concept, which is more in line with trusting belief being a stronger or more complete kind of trust, but there still being genuine cases of trust without belief. If it does count as trusting, then we have Frost-Arnold's problem of acting/planning-as-if sufficing for testimony trust. The ideal, then, seems to be that something akin to acceptance or entrusting is real trust, but that it entails belief when applied to testimony. This would avoid both horns of the dilemma. The next section will consider a view that attempts to do just this.

Reliance from the Participant Stance

The view espoused by Richard Holton is that trust need not involve belief, but is a specific kind of reliance. Reliance is characterised as acting and planning as though something were true (Holton, 1994, 72), so seems no different to Frost-Arnold's acceptance or Hieronymi's entrusting. In particular, trust is reliance from the participant stance (Holton, 1994, 67), a concept borrowed from Peter Strawson. To take the participant stance towards someone is to

view them as a participant in the world and not a mere feature of it. Certain attitudes are appropriate only for fellow participants. (Strawson, 1974, 9-10) For instance, we should not feel grateful or resentful towards the weather if it is or is not how we hoped, though we may be pleased or disappointed. To take the participant stance, then, is to be ready to adopt certain reactive attitudes towards someone, such as gratitude or resentment. Holton suggests that betrayal is one such attitude; mere features of the world can disappoint us, but not betray us. When we rely, we are ready to feel disappointed if what we hope for fails to occur; when we rely from the participant stance, or trust, we are ready to feel betrayed.

On Holton's view, reliance is not governed only by epistemic reasons and can be chosen. He supports this view with the examples of the trust circle (1994, 69) and the former thief working in the shop (1994, 63). In both cases, there does seem to be a genuine choice about whether to trust the other person and factors other than truth play a role in justifying that decision. The participant stance marks the difference between trust and mere reliance. One can rely on the other drama students to catch one without the participant stance, feeling 'no sense of betrayal; just a grim confirmation of [the classmates'] alien ways' (1994, 69) if one is not caught. But the participant stance does not change the fact that there is choice involved in trusting, nor that the decision can be made for non-epistemic reasons (1994, 69-70).

So how does Holton, with his view that trust does not require belief, respond to the problem of trusting testimony? As indicated above, he suggests that, in testimony cases, belief is entailed by reliance. When we trust someone's testimony, we are trusting them to do something specific: to speak knowledgeably and sincerely. That is, we are relying on them to speak knowledgeably and sincerely from the participant stance. Holton reasons that if we do not believe someone, then we cannot be thus relying on them. In this way, belief follows from reliance in the case of testimony. (1994, 74)

This at first appears reasonable. 'I am relying on you to speak knowledgeably and sincerely, but I do not believe you' seems about as nonsensical as 'I trust you in what you say, but I do not believe you'. The problem, though, is that, even in testimony cases, Holton thinks that we can choose to rely and for non-epistemic reasons. If in such cases reliance entails belief, it seems that we can, albeit indirectly, choose our beliefs and justify them with the wrong kinds of reasons.

Holton is aware of this problem and tries to mitigate it by referring to another aspect of his view: though we may rely on something without believing that it will happen, we cannot rely on something if we positively believe that it will not happen (1994, 71-2). In testimony cases, then, we cannot choose to rely on someone to speak truly if we already believe that what they are saying is false. Thus, we will not be led to have a contradictory set of beliefs. He also highlights that this process of coming to believe through reliance is indirect. We do not simply choose our beliefs; we choose to rely on a particular person's testimony and in so doing come to have certain beliefs. (1994, 76)

However, these mitigations do not fully solve the problems. Even if we could not uproot our existing beliefs through this method, we could, as Hieronymi (2008, 221) points out, gain new beliefs that we know to be based on non-epistemic reasons. We believe because we chose to trust someone for practical reasons. Even if we do not directly choose to believe, but only choose who to trust, the choice still seems to be too direct. If we want to acquire a belief, we can simply choose to trust someone who espouses that belief. Somewhat ironically, in trying

to avoid the problem associated with his own type of view, Holton stumbles into both the problems associated with belief-based views: this would enable us to choose at least some of our beliefs and do so for non-epistemic reasons.

Separately, an objection can be raised to Holton's view that reliance from the participant stance on another to speak knowledgeably and sincerely entails belief. It was mentioned that it seems nonsensical to have such an attitude without belief, but this is only an initial appearance. When we consider what Holton means by reliance, it becomes clear that it is quite possible to rely on testimony without believing it. One can act and plan as though another has spoken knowledgeably and sincerely without believing that they have, so testimony trust, on the current theory, need not involve belief. Adding that the acting/planning-as-if must be done from the participant stance does not change this. It therefore seems that Holton is not successful even in his attempt to avoid the problem associated with his kind of view. He falls into the same problem as Frost-Arnold, and which Baker was trying to avoid: testimony trust can still merely be acting and planning on certain assumptions, so does not guarantee belief.

Nevertheless, there are lessons to be learned here. First, Holton gives a key insight in telling us that trusting testimony is trusting someone to speak knowledgeably and sincerely. This shows that testimony trust is a kind of act trust – it is trusting the other person to behave in a certain way. Without this insight, we may have been tempted, upon noticing the intuitive differences in the nature and norms of act trust and testimony trust, to think of them as entirely separate concepts with no call for a unifying theory.

Second, Holton's unsuccessful attempt shows that we need a more specific and detailed account of how reliance may entail belief. His theory failed because reliance from the participant stance does not entail belief and even if it did, it runs into the problems of choosing beliefs and justifying them with non-epistemic reasons. But if these problems could be solved, something similar to this view may well be correct; although it does not avoid the problems sketched above, it seems to be on the right track insofar as it aims to give an account that does not require belief for act trust but which entails belief in the case of testimony trust.

If, then, we can give an account of reliance that genuinely entails belief when applied to testimony and which, in such cases, also could not be chosen or justified by non-epistemic reasons, we would have the beginnings of a suitable account of trust. What follows will not be a fully fleshed-out theory of trust, but will hopefully suffice to show that some theory that avoids the problems so far identified is possible. I will present my preferred theory of trust in Chapter 2, and will expand on it in Chapter 3. For now, I turn my attention to developing an account of reliance that will be a suitable foundation for it.

Part 2: Reliance, Belief and Trust

I will now present my own way of solving the puzzle. I will argue that, with a better account of reliance, it can be shown that reliance on testimony does entail belief. What is more, while we can usually rely by choice and for non-epistemic reasons, reliance on testimony is a non-arbitrary exception to this. We cannot simply choose to rely on testimony, and when we do so,

we can rationally rely only for epistemic reasons. We cannot, in contrast to Holton's view, choose to rely and by doing so come to believe. If my argument is successful, then we can say that trust is a kind of reliance that entails belief in the case of testimony without falling foul of the problems that we have so far encountered. How trust is distinguished from mere reliance is a question that I leave aside until the next chapter.

A Better Account of Reliance and Testimony

What does it mean to rely on someone or something? Holton suggests that it is a matter of acting or planning on a supposition, but we have found that such an account is not suitable for a theory of trust based on reliance. Acting/planning-as-if, as shown with Bratman's acceptance adopted by Frost-Arnold, with Hieronymi's entrusting, and with Holton's reliance, does not entail believing another's testimony. Moreover, treating reliance in this way does not fit with ordinary usage. When we rely on something, we stand to suffer some loss or harm if it does not do what we rely on it to do. Reliance is thus a matter of having something at stake.

Take the example of the drama class trust exercise. Part of the reason why talk of reliance is appropriate is the fact that, if one is dropped by one's classmates, one will come to harm. One's bodily comfort and safety is at stake and one relies on them to prevent the potential harm. In contrast, consider the example given to demonstrate the idea of acceptance. We assume the maximum possible costs of materials in calculating the budget for a building project. We act and plan as if they will cost that much, but we do not thereby *rely* on their costing that much. This is because, if they are cheaper (by hypothesis, they will not be more expensive), we will not be disappointed; we have not suffered a harm or loss. On the contrary, we will be glad of the savings made. This implies that reliance cannot just be acting and planning as if something is the case, in contrast to Holton's view.[7] Rather, it is putting something at hazard, such that one will suffer some harm or loss if what is relied upon fails to occur. This will typically involve acting as if it will, as the drama student acts as if their peers will catch them, but, I will argue, sometimes requires belief rather than action. I thus do not take reliance to be a specific type of mental state, but a way of engaging with the world that may involve a variety of actions and attitudes, depending on the situation and what is relied upon.

This idea of reliance is somewhat akin to Hieronymi's 'liberal' notion of trust discussed above. On that idea, we trust not to the extent that we trustingly believe, but to the extent that we are dependent on the other. Here, one similarly relies to the extent that one depends on a particular event happening. One relies more just in case more is at stake. This is to be expected, since her arguments for rejecting the liberal notion were not found to be convincing and we are accordingly now taking a reliance-based approach to trust.

What reasons justify reliance on this view? They will be a mixture of practical and epistemic. Rational reliance is like a rational bet: one must take into account the potential loss, the potential gain (practical reasons), and the probability of the event's occurrence (epistemic reasons). Weighing these factors against each other, one determines whether the risk is worth taking. Incidentally, this accounts for the rationale behind the purist's notion of trust that

[7] For the same reason, acceptance is not an appropriate basis for trust. Trust requires a certain vulnerability to harms or costs and, as shown by the building project example, not all cases of acceptance entail vulnerability. Indeed, that is an example of *avoiding* vulnerability.

Hieronymi favours. If we take trust to be a kind of reliance, we will often trust more when we believe more strongly; all else being equal, it will be more rational to trust when we have sufficient evidence for belief than when we lack it. But that is not the same as trust requiring belief.

Armed with this idea of reliance, let us turn to the distinction between act trust and testimony trust. The former presents no great difficulty. If trust is a kind of reliance and reliance is how I have characterised it, then we find no problem with the intuitive ideas that act trust is voluntary and responsive to practical reasons. We can choose to risk various harms and losses by our actions, as we might choose to place bets. Although rational reliance is subject to epistemic norms, there are also the practical considerations of what might be lost set against what could be gained. For instance, when choosing whether to rely on the former thief employed in our shop, we will think about what might be gained – greater efficiency in the business, better relations with the employee, a step towards reformation for a former criminal – what might be lost – the money in the till – and what we judge the chances are of the money being stolen. Ultimately, whether it is rational to do so or not, we can choose to leave them alone with the till. To do so would be an act of reliance, since it is putting something at stake. This is not to say that there is no more to act trust than reliance; it is possible to rely on the employee without trusting them, and not all the reasons that support reliance need support trust. The point is that taking trust to be a kind of reliance does not present problems with the view that act trust is voluntary and justified by practical, as well as epistemic, considerations, in the way that taking trust to be a kind of belief does.

The more complex issue, of course, is that of testimony trust. I have said that, properly understood, reliance on testimony entails belief. But how can reliance, which is voluntary and responsive to practical reasons, entail belief in such cases, if it does not do so when relying on someone to do something? Reliance is a matter of putting something at stake, so we must consider what is at stake when we rely on another's testimony. This could include a great variety of things, depending on the situation. To take the examples of Baker and Hieronymi of the accused friend, one thing that is at stake is the maintenance of the friendship. Another, if we are considered a good character witness, may be the conviction or acquittal of someone who may or may not be innocent, or at least their social ostracism or acceptance. Do we rely on her testimony and say that they are certainly not the kind of person to do this, or do we say that we are unsure? Or again, suppose that in the trust exercise case, you have been specifically told by your drama classmates that they will catch you. In either case, there is a sense in which there can be reliance on testimony without belief – by one's actions, one may stake something on the testimony being true. But not all cases of relying on – or trusting – testimony are like this. Suppose that there are no practical consequences at all. They are offering their testimony, hoping to be believed, but there is nothing you can do that would stake anything on the truth of their belief. If you rely on them, must this amount to belief? Is it even possible to rely on them if there are no practical consequences?

Let us refine the question. When we rely on another's testimony, we are, to adopt Holton's phrase, relying on them to speak knowledgeably and sincerely. This implies that we are relying on them to have told us the truth. So, if we rely on their testimony that p, we rely on its being the case that p. The idea of reliance advocated above entails that this means there is something staked on p being the case; we are vulnerable to some cost if their testimony turns out to be false. That cost, however, must be an *epistemic* cost. What is at stake is, by hypothesis, not

practical, so is purely epistemic. If we rely on testimony and that testimony turns out to be false, what have we lost? What is meant by an epistemic cost?

The most obvious answer is that an epistemic cost is believing something that is false. If we rely on someone's testimony, we are making ourselves vulnerable to acquiring a false belief. This entails that reliance on testimony does indeed change our beliefs – we are not at risk of acquiring a false belief unless we acquire a belief. Therefore, relying on testimony must entail forming new beliefs.

So, if I rely on someone's testimony and that testimony turns out to be false, then I will gain a false belief. This much follows from what it means to rely on testimony. What is the content of the beliefs thusly gained? The most straightforward answer is that, in relying on testimony, one gains the belief that the asserted proposition is true. We therefore suffer the epistemic cost of gaining a false belief if our reliance on another to speak truthfully is disappointed.

This, then, is what I suggest is on the line when we rely on others' testimony: the truth of our beliefs. This in turn indicates that reliance on testimony is possible and that it does indeed require belief. If we rely on someone to be truthful in their assertions and they are not, then we suffer the epistemic cost of believing something false. This avoids the problem with seeing reliance merely as acting-as-if. It is entirely possible to act as if someone is speaking truthfully without believing them; it is not possible to stake the truth of one's beliefs on their speaking truthfully without believing them.

Avoiding Holton's Problems

Might it be, then, as Holton argues, that we can choose to rely on someone's testimony and thus come to believe what they say? He holds that reliance on testimony entails belief and I have now argued for the same conclusion. Since we can choose to rely and do so for non-epistemic reasons, was Holton correct after all? No. When it comes to relying on testimony trust, I will argue, only epistemic reasons count. Since it requires belief, which is not a voluntary attitude, reliance on testimony is likewise not under our direct voluntary control. I will now further unpack my approach to reliance on testimony, showing how it can avoid the problems encountered when we discussed Holton's idea.

To explain why my view does not have the same implications as Holton's, it is important to appreciate the difference between our views regarding the relation that holds between reliance on testimony and belief. We agree that the former entails the latter, but not all entailment relations are the same. As a reminder, here is how Holton explains his view:

> When I trust my friend I rely on her to speak knowledgeably and sincerely. As a result of that reliance, I believe what she says. The belief *follows* from the reliance in the sense that I would be failing to act on the supposition that she speaks knowledgeably and sincerely if I did not believe what she said. As a result of deciding to rely, I come to form new beliefs … I suggest then that sometimes we can trust a friend to speak knowledgeably and sincerely, without believing that they will. And as a result of this we will believe what they say.

(Holton 1994, 74-5 [emphasis Holton's])

What is being claimed here is that there is a close *causal* relation between reliance and belief. One chooses to rely without believing, but this causes one to believe. It is unlike many common causal relations, such as an open oven causing a fire, in that the cause must always lead to its effect; it is entirely possible to leave an oven open and no fire to result. This is a kind of necessary causal link; it is impossible to genuinely rely on testimony without the corresponding belief following. Reliance on testimony is thus a sufficient causal condition for belief on Holton's view.

Mine, however, is different. I do not hold that the belief is *caused by* reliance; rather, I hold that the belief *constitutes* reliance. Just as reliance on one's fellows does not cause one to fall backwards in the drama class trust exercise, but rather one's falling is itself the act of reliance, so believing someone's testimony just is reliance on it. I can therefore agree with Holton that '[t]he belief *follows* from the reliance in the sense that I would be failing to [rely] if I did not believe what she said', but I have a different explanation of the entailment. With this in mind, we can turn to the reasons and voluntariness of reliance on testimony.

Only epistemic, not practical, reasons count when it comes to relying on testimony. This is because reliance on testimony does not cause belief, but is itself a kind of belief. It is therefore subject to the norms of belief, rather than pragmatic norms. That is, it is rationally justified only by epistemic reasons. Practical reasons do not count simply because they are the wrong kind of reasons for belief; anything that constitutes belief, such as reliance on testimony, can therefore not be justified by them. To say that only epistemic reasons count is, of course, not to say that all epistemic reasons count. If I believe what you say because I consider you to be a reliable source of information, then I am relying on your testimony. But if my reasons for believing have nothing to do with you, then I believe your testimony without relying on it. My independent reasons for belief do not support relying on you.[8] Again, I am here discussing reliance rather than trust. The reasons of testimony trust, as opposed to mere reliance, will be an even more restricted subset of the epistemic reasons for belief. More precise discussion of this matter I defer to Chapters 4 and 5.

Could Holton make a similar point, arguing that, since reliance on testimony entails belief on his view too, it is justified only by epistemic reasons? I think not. As mentioned in discussing his theory above, his idea of reliance does not actually entail belief at all, since one can act on the supposition that another is speaking knowledgeably and sincerely without believing them. If reliance is acting on a supposition, it would be arbitrary to claim that a specific kind of reliance is immune to practical reasons, just as it would be arbitrary for Frost-Arnold to claim that all cases of testimony trust involve belief rather than mere acceptance. Furthermore, even if we accept Holton's argument that reliance on testimony necessarily causes belief, it still seems arbitrary to claim that such reliance is only rationally responsive to epistemic reasons. If we can choose to do something that will lead us to have a belief, there is no reason why we should not take into account practical reasons. If we realise that, true or not, having a certain belief will benefit us, and we are able to bring about that belief in ourselves, say by taking a certain belief-inducing drug, then we would have practical reason for doing so. The drug and the practical reasons we have for taking it would cause the belief, but would not justify it. On the other hand, it is not arbitrary to claim that reliance on testimony is rationally sensitive to

[8] Hieronymi (2008, 222-4) makes a similar point, saying both that one can rely on testimony without trusting and that one can believe testimony without either trust or reliance.

only epistemic reasons where reliance is seen as putting something at stake. In testimony, the stakes are the truth of one's beliefs, so the reliance is constituted by (rather than causing) belief, which is only justified by epistemic reasons. Since, in such cases, the reliance just is a belief, it is subject to the reasons of belief. The crucial difference is that, on Holton's view, reliance on testimony is a way of bringing about a belief, from which practical reasons need not be excluded; on mine, it is itself a kind of belief, which does exclude practical reasons.

For that same reason, reliance on testimony is also not voluntary on my view. Reliance is usually something that we can choose, since it is normally constituted by performing some action that puts something at stake. However, in the special case of testimony, when reliance is constituted by belief, it cannot be chosen. As a belief, it is non-voluntary. This stands in contrast to Holton's view. If reliance on testimony is not itself belief, but causes belief, then there is no reason why it should not be as voluntary as other cases of reliance.

Holton was aiming for an account of reliance that does not entail belief in act cases, but which does entail belief in testimony cases. This, I believe, is a good aim. If, as mentioned earlier, it can be accomplished, then we would have the basis for an account of trust that avoids the problems associated with both the belief-based views and the non-belief-based views. However, he does not achieve his objective. I rejected his view for three reasons: it allows us to choose our beliefs too directly by deciding who to rely on; it thereby allows those beliefs to be rationally responsive to non-epistemic reasons; his idea of reliance does not actually entail belief in testimony cases. However, the problems that affected Holton's view do not affect mine. On my view, reliance on testimony is a type of belief, rather than a cause of belief, whereas other kinds of reliance are not. This view of reliance thus fulfils Holton's aim and so has strong potential to form the basis of an account of trust.

Before moving on, it is worth mentioning how this conception of reliance can also encompass some of the ideas highlighted by one of the other philosophers we have discussed. Although we take differing approaches to trust regarding the reliance or belief question, there is some agreement between the view here advocated and Hieronymi's. She points out that sometimes all that is required in trusting is believing; there is no action to be taken at all (2008, 219; 222). She takes this to indicate that trust must be fundamentally a kind of belief (2008, 221), but we can now see how it is compatible with a reliance-based account. Reliance entails believing and nothing more in testimony cases. She also tells us, in discussing Holton's drama class example, that we cannot entrust our beliefs in the same way as we can entrust our bodies (2008, 220). We might choose to fall backwards, entrusting our bodies to our classmates, without believing that we will be caught; it is contradictory to suggest that we might choose to believe someone – so entrusting our beliefs to them – without believing. This, too, is explicable on my view. Since reliance on testimony consists in belief in that testimony, it does not make sense for us to rely on another's testimony without believing it. To properly 'entrust' (to borrow Hieronymi's term) our beliefs to someone else, we have to believe them. We can, however, rely without belief in practical cases, like the drama class trust exercise, because reliance does not then consist in belief.

To summarise the argument concerning reliance and belief: To rely is to put something on the line, to make it the case that one will lose something if what is relied upon does not obtain. For testimony cases with no practical stakes, the potential loss must be epistemic; the truth of one's beliefs are staked on the other person asserting accurately. Therefore, reliance consists in

believing the asserted proposition. Since it is a kind of belief, we can only rely on testimony for epistemic reasons and we may not directly choose to rely. But for non-testimony cases, reliance does not entail belief, so it can still be chosen and for practical reasons. This fulfils the aim stated at the start of this part: to give an account of reliance that does not usually require belief, but does in the case of relying on testimony.

An account of trust as a kind of reliance could therefore, on this view, avoid all of the problems of the other views, not just Holton's. Since belief is not essential to trust, the fact that act trust is subject to non-epistemic reasons and can be chosen is not troubling. Neither is the fact that testimony trust requires belief, since reliance on testimony involves belief. We do not bite the bullet, as Baker does, claiming that belief can be chosen and motivated by reasons of friendship. Nor do we find that acting and planning as if someone's testimony is true counts as testimony trust, as is implied by Frost Arnold's view. Finally, unlike Hieronymi, we can say that there can be genuine trust without belief – and therefore that the complaint of betrayal if someone lets us down is just as legitimate when we did not fully expect them to follow through.

Mixed Cases

But what about cases in which someone makes an assertion and there are practical considerations depending on its truth? We have said that relying on someone in their testimony requires belief, for it is truth that is at stake. Relying in practical cases does not, since the stakes are practical. But often, as already noted, there might be practical stakes attached to believing what someone says, and we can still believe in non-testimony cases. How do these mixed cases fit into my account? The latter is easily explained. Reliance is subject to both practical and epistemic reasons. Sometimes, the epistemic reasons are sufficient for belief. Therefore, it is not uncommon for reliance – and by extension, trust – to be accompanied by belief. Indeed, this commonality in the conditions conducive to trust and belief may be one source of the confusion in the question of whether belief is essential to trust; even though it is not essential, it is frequently present at the same time.

What of the practical stakes attached to testimony? To use again the adapted example of the trust circle, suppose that you know in advance which classmate will be responsible for catching you. Perhaps knowing that you are nervous, they reassure you in advance, explicitly telling you that they will catch you. Now, as you stand with your eyes closed, deciding whether to fall backwards, the situation is a little different to that in Holton's original case. As before, you can decide to fall backwards, but you cannot decide to believe that you will be caught. This time, however, there is the added consideration of the classmate's testimony. If you fall backwards, are you trusting them?

To understand this scenario, it must first be pointed out that there are two separate but related instances of potential reliance or trust. The first is whether you trust that they will catch you. This is no different to in the original version. You can rely – and trust – without believing, since in falling backwards, you are putting your bodily comfort and safety literally in their hands. The second is whether you trust your classmate in their saying that they will catch you. This may seem like just another way of describing the same thing, but the fact that they are non-identical is shown by the fact that one can exist without the other; it is possible to trust that one will be caught without receiving the reassurance, so never having the chance to trust your

classmate's testimony. Now, if you fall backwards, are you trusting their testimony? Let us start with the case in which you do believe that you will be caught. If your belief is based at least partly on the fact that you have been told that you will be caught, then it seems that you do indeed trust your classmate's testimony. However, it may be that your belief has nothing to do with what you have been told. Maybe you did not require this reassurance and would believe without it. You may not be thinking about your classmate's testimony at all, or even have forgotten it, and have formed a belief based on other reasons. In this case, although you trust them to catch you, you do not trust their testimony that they will catch you. There is act trust, but no testimony trust. You are not relying on your classmate's testimony for your belief.

Now consider the case in which you do not believe that you will be caught. Perhaps the reassurance has made a difference and you now think it more likely than you would otherwise have done, but you still do not believe. This seems to present a problem for my view. Relying on testimony, I have said, requires belief. This was based on the idea that reliance requires putting something at stake. In purely epistemic cases, there are only epistemic stakes, so reliance requires belief. But here, there are practical stakes, so it seems that the motivation for thinking that reliance on testimony involves belief is absent. However, we can again appeal to the two instances of reliance: you can rely on the classmate to have spoken truthfully; and you can rely on them to now catch you. In the case of the former, what you are relying on them for is truth, not bodily comfort and safety. Therefore, what has been said of purely epistemic cases applies and such reliance requires belief. In the case of the latter, you are relying on them for comfort and safety, not truth. Being sure of the truth of what they have said will be a reason for relying on them to catch you, but you are not relying on them to have told you the truth. So, such reliance does not require belief and is responsive to both practical and epistemic reasons.

When criticising Hieronymi's arguments for her purist's notion of trust over the more liberal notion, I argued that what she said in support of her view was consistent with trust involving belief being a stronger and more complete form of trust than that without belief, rather than, as she claimed, trust without belief being no trust at all. We can now explain why this is so, applying what has just been said of reliance to trust. If you agree to do something for me, such as keep my secret or catch me when I fall, then I can simply trust you to do the thing in question. But, as we have seen, there may be an additional layer. I might also trust your testimony – trust you to have spoken knowledgeably and sincerely – when you agree to do it. This is a separate, though closely related, instance of trust. Therefore, Hieronymi is right in thinking that trust involving belief is, in a sense, a more trusting kind of trust than trust not involving belief. In such a case, I am trusting you with two things, not just one. We can therefore say that trust without belief is genuine trust while acknowledging that it is more trusting to believe.

So, in a mixed case, in which one may rely on or trust someone's testimony with more than epistemic stakes, we can still say that testimony trust entails belief and act trust does not. Once we understand that there are two connected but different things with which we may trust the other, the problem posed to my view of reliance on testimony evaporates.

Conclusion

In this chapter, I have made the argument that belief is not essential to trust; that trust is a kind of reliance and that this entails belief in the case of testimony. This solves the puzzle with which we began, avoiding the problems associated with both types of view. However, I have not laid out a full account of what it means to trust someone.

It might be wondered, for instance, what distinction is to be made between reliance and trust. If trust is fundamentally a kind of reliance, rather than belief, what is added to mere reliance to make it into trust? There must be some difference, since it is plainly possible to rely without trusting. We routinely rely on objects like cars, shelves, and chairs without trusting them. As Holton points out, it is possible to rely on one's classmates without trusting them (1994, 69). Extending the point to testimony, what is it that distinguishes a trusting belief from an ordinary belief in another's testimony? As Hieronymi says, it is possible to believe what another says without trusting them – we might believe the content of their speech for reasons that have nothing to do with their testimony, or we might take what they have to say as reliable evidence, treating them 'as a good thermometer' (2008, 222). We have shown how a proper understanding of reliance can avoid the problems discussed above, so solves the puzzle of the relationship between trust and belief. But we have yet to show what trust actually is.

This is a task that I defer to the following chapter. This chapter has merely been the groundwork to the larger question of what it means to trust. As things stand, I have only given an account of reliance that is fit to serve as a basis for the concept of trust; relying on someone on the view that I have presented is just treating them 'as a good thermometer', and we can of course rely on objects as well as persons.

In the next chapter, I hope to provide a satisfying account of trust that distinguishes it from mere reliance and explains why trust is appropriate only for persons, not objects. Later, in Chapters 4 and 5, I will return to the matter of the kinds of reasons that justify trust, which have here received only a cursory treatment. On this subject, there is one further lesson that can be drawn from the above attempts to give an account of trust. Hieronymi argues that any epistemic reasons which justify trust must make reference to the other's trustworthiness (2008, 224). This is what distinguishes an ordinary belief from trust. Although my view is based on reliance rather than belief, I do think that this is an important insight and shall make use of it in the following chapters, as I develop a theory of trust.

We now have the foundation for an account of trust by which it is a kind of reliance, not belief, but on which belief is necessary in the case of testimony. On this view, reliance on testimony without belief is not possible, so neither is trust in testimony without belief. Correspondingly, there are no practical reasons for trusting testimony and it is not under our direct control. However, we also have that trust can generally be chosen and justified by non-epistemic reasons. Testimony trust is a non-arbitrary exception to this.

If a more complete view of trust can be built from this, we will have an account that avoids all the problems raised above. It would therefore have a significant advantage over many existing theories.

2

What is Trust?

Introduction

It was argued previously that trust is a kind of reliance rather than a kind of belief, but that it entails belief in the case of testimony. I turn now to the task of getting more specific. Given that trust is a kind of reliance, what more can we say about what it is?

There are two main questions to here. What distinguishes trust from mere reliance? What exactly are we relying on when we trust? I shall adopt two necessary conditions for an account of trust which will guide the discussion. First, trust typically – though not necessarily always – entails vulnerability to betrayal (Baier 1986, 235). If an account of trust does not allow for a plausible sense in which one has been betrayed when one's trust is deliberately frustrated, then it needs amending. Second, trust involves the recruitment of another's agency (Jones 2012, 65). It is a part of the purpose of trust that it facilitates cooperation between agents, so if an account of trust does not imply that the trustor is in some way making use of another person's agential capacities, then it is wide of the mark.

Once again, we will consider a series of existent views on the matter. In Part 1, we shall discuss and ultimately reject each, but we will learn from all of them. In Part 2, I will set out and defend my own view.

We will examine, in turn, four influential theories of trust. All are theories according to which trust is a kind of reliance, since that much we have established already. The first is that proposed by Annette Baier (1986, 234), which takes trust to be reliance on the goodwill of another. The distinction between trust and reliance is that the former involves imputing goodwill to the other person; they are relied on to act out of some positive disposition towards us. The second is reliance from the participant stance, which is proposed by Richard Holton (1994, 67). We have already encountered Holton's view in the previous chapter, but we shall this time examine it at greater length. Third is the idea that to trust someone is to rely on them to be moved by the very fact of our reliance (Jones 2017, 99). When we trust someone, we impute to them a disposition to respond to our dependency. Finally, we will consider the view put forward by Katherine Hawley (2014, 10), that trust is belief that another has a commitment, combined with reliance on them to fulfil it.

My own view, which shall build on what can be learned from these others, is that to trust is to rely on another to be trustworthy in the given instance. This is a simple account and has much intuitive appeal, but also requires unpacking. What exactly does it mean to be trustworthy?

That question will be answered in only a cursory way in this chapter; a more detailed answer is given in Chapter 3, which is dedicated to the topic.

Part 1: Reliance-Based Accounts of Trust

This section examines, in turn, four accounts of trust. These all agree, as I argued in the previous chapter, that trust is a form of reliance. However, they differ substantially in their details. Each will be found to be subject to certain objections, which will enable us to craft a more accurate theory of trust.

Goodwill

The goodwill account of trust is an initial attempt to distinguish trust from reliance. The main advocate of this view is Annette Baier, who makes the marking of this distinction her explicit aim:

> What is the difference between trusting others and merely relying on them? It seems to be reliance on their good will toward one, as distinct from their dependable habits, or only on their dependably exhibited fear, anger, or other motives compatible with ill will toward one, or on motives not directed on one at all.

(Baier 1986, 234)

The thought here is that we can plainly rely on others in various ways without trusting them. Comedians, advertisers, blackmailers, extortioners and terrorists all rely on others to act in various ways in response to what they do, but are not trusting those upon whom they rely. Kant's neighbours are said to have relied on his regular walking habits to know the correct time, but they never *trusted* him to come by at a given time. (Baier 1986, 234-5) What seems to be missing in these cases of reliance without trust is goodwill being imputed to the one relied upon. That does not entail that there is ill will in all such examples; there may be indifference, or there may be general goodwill that is not (thought to be) influencing the motives of the relied-on person on the given occasion. Trusting someone, according to Baier, involves relying on them not only to act in a certain way, but to do so out of the goodwill that they bear towards oneself.

This does seem to fit the guiding conditions that we have adopted. There is a clear sense in which trust, seen in this way, is vulnerable to betrayal. If we rely on someone to do something out of goodwill towards us, but, out of ill will, they refuse to do it, then they have plausibly betrayed us, such that this betrayal is quite distinct from the disappointment we might experience had we merely relied on them without trusting. It also seems to involve recruitment of another's agency in a way that mere reliance does not. In a certain sense, Kant's neighbours were recruiting his agency, in that taking a walk at that time each day was a decision that he took for his own reasons. But his agency does not substantially matter in this reliance; it makes

no difference to the neighbours which reasons he is responsive to, or even if he walks for no reason. They are treating him like a reliable clock, rather than like an agent. On the goodwill account of trust, the other person is supposed to be responding to us, because they are favourably disposed towards us. This is something that agents can do, but non-agents cannot, so trust as reliance on goodwill can encapsulate the condition of recruiting another's agency.

Although it is initially plausible, this account is not free from objections. One example that Baier herself gives to demonstrate the sheer breadth of types of trusting relations is of trusting a declared enemy to not fire upon us if we hold up a white flag (Baier 1986, 234). True, we cannot always trust enemies in this way, but it does seem possible that one can, at least sometimes. Richard Holton (1994, 65) picks up on this instance, pointing out that it does not seem compatible with the idea that trust involves imputing goodwill to the one trusted. Surely there is no goodwill between enemies ordinarily prepared to kill each other and there is little reason to suspect that it is mistakenly imputed, that we think a declared enemy bears us goodwill when they do not. So, when we hold up a white flag and step out into no-man's-land, with confidence that we will not be shot at, there must be something else going on.

There are reasonable responses that can be made to this objection. One is to row back and say that there is, after all, no trust involved in these circumstances. The surrendering soldier does not trust the enemy, but only relies on them. They rely on the enemy to have basic moral standards, or on their fear of the consequences of committing war crimes. It is therefore quite reasonable, where this reliance is justified, for them to confidently expect to not be shot, but this does not amount to trusting the enemy troops.

Another response is to claim that there in fact is goodwill, but only of a very minimal kind. Baier makes it clear that her theory only requires a low level of goodwill, as there may be between complete strangers. We often trust people we don't know simply to not harm us. (Baier 1986, 234) If this is so, then perhaps, with some enemies, we do genuinely consider them to have this very minimal kind of goodwill towards us. They may be firing on us (prior to the white flag), but they are not doing so out of pure malice, but because that is what they are ordered to do. Indeed, as the famous example of the Christmas Truce shows, soldiers on opposite sides can find that they have much in common with one another and even have a certain camaraderie.

A combination of these responses is more likely to be accurate than just one alone. Sometimes one might surrender with reasonable reliance that does not amount to trust; sometimes one might think the troops on the other side do bear one some minimal level of goodwill. Either way, Holton's argument that apparent trust in the enemy is a clear counterexample to Baier's theory does not seem as strong as it may at first appear. He thinks that '[t]o impute goodwill here would be to deprive the notion of all content' (1994, 65), but upon reflection, this seems too bold a claim. Nevertheless, it is worth acknowledging that imputing goodwill to a declared enemy is at least a contentious issue. It might reasonably be thought that it stretches the ordinary notion of goodwill somewhat, even if not quite to breaking point.

We could go on to analyse the concept of goodwill and discuss its limits, but I do not think this necessary. There is another objection to the goodwill account which is rather more decisive. Another example of Holton's (1994, 65), this is the point that a con artist will often rely on the goodwill of their victim. Having invented a plausible-sounding story of their own desperate need, perhaps backed up with forged evidence, they ask for money or valuable information,

banking on the supposition that their victim will want to help. They rely on the goodwill of their victim. Yet such cases of fraud and manipulation should not count as cases of genuine trust.

Baier does point out that not all trust is good. There are immoral trust relations. When people generally trust one another, they are vulnerable to being taken advantage of: 'Exploitation and conspiracy, as much as justice and fellowship, thrive better in an atmosphere of trust.' (Baier 1986, 231-2) However, this would be misguided as a response to the con artist objection. That exploitative relations flourish in an atmosphere of trust does not mean that con artists trust their victims, but that potential victims are more likely to trust con artists and so become actual victims in such an atmosphere. Other nefarious trusting relations, such as those among members of the mafia, might also flourish, but their attitude towards those whom they exploit is not a trusting one.

The con artist does not trust their victim, even in an immoral way. They do not recruit the victim's agency to help them, but manipulate them into getting what they want. Manipulation, on any plausible view, is surely antithetical to full agency. They also fail to be vulnerable to betrayal. If a victim does not do what the con artist wants, they may be disappointed and even consider themselves to have suffered a loss, if they have used time and resources to try to persuade their victim that what they have said is true. But it clearly would not be a betrayal.

This objection from Holton indicates that, even if expecting the other person to act out of goodwill is a necessary part of distinguishing full trust from mere reliance, it is not sufficient. Goodwill is too broad a motive to differentiate between trusting and manipulating. We turn next to the theory favoured by Holton himself, to see if that fares any better.

The Participant Stance

The idea of the participant stance is developed from Peter Strawson (1974, 6-13). It is a readiness to feel certain reactive attitudes, given rise to by seeing another as an agent, as a participant in the world, as opposed to a mere feature of it. When we take the participant stance towards someone, we do not feel merely glad or disappointed by their beneficial or detrimental actions towards us; we feel grateful or resentful. Such responses are appropriate only to other people, not to features such as the weather or other natural forces. It has been argued that the participant stance is what distinguishes trust from mere reliance.

Although the main advocate of this view is Richard Holton (1994), versions of it are also put forward by Pamela Hieronymi (2008, 215-6) and Berislav Marušić (2015, 191-205). Here, we will focus on Holton's original version.

Holton's argument runs roughly as follows. Trust requires a readiness to feel betrayed, not merely disappointed, if we are let down. The feeling of betrayal is a reactive attitude of the kind only appropriate to other agents; unless we are anthropomorphising, we will not feel betrayed by objects or anything that lacks agency of its own. Since the readiness to feel such a reactive attitude entails that we are taking the participant stance towards something, Holton concludes that trusting someone must involve taking the participant stance towards them. (Holton 1994, 66-7) Since trust is a kind of reliance, we can see it as reliance from the participant stance: reliance with a readiness to feel betrayed, not merely disappointed.

This approach does seem to do better than the goodwill account. It can, for instance, avoid the con artist objection. The con artist does not take the participant stance towards their victim. They see their victim as something to be manipulated and managed to produce certain results; the agency of the victim is not of any particular importance.

Moreover, it is quite explicit about the need to see the other person as an agent, so can plausibly encapsulate the condition that trust requires the recruitment of agency. It also seems able to capture the necessary vulnerability to betrayal, since the readiness to feel betrayed is an in-built feature of the participant stance.

It might be wondered, however, why I am examining Holton's view here when I have already considered it in the previous chapter. There, I argued that it was not able to do what Holton thinks it can do, namely, tread the line between being a kind of reliance in act trust yet entailing belief for testimony cases. Surely, if it was wrong then, then it is still wrong now. So why am I still discussing it?

The answer is that the problem we encountered before was with Holton's method of applying trust as a kind of reliance to the matter of testimony trust, rather than with his view of trust as reliance from the participant stance itself. There is no reason, in principle, why an advocate of such a view could not apply the argument that I used in the last chapter to the participant stance view. It does not rule out the general idea that reliance involves putting something at stake, or the specific idea that reliance on testimony entails putting the truth of one's beliefs on the line; it is just that Holton does not appeal to these considerations. The participant stance account should therefore still be considered a live option.

Nonetheless, it is an option that I do not think is correct, for three main reasons. The first is that it seems possible to rely from the participant stance without trusting. A readiness to feel betrayed may imply that one has adopted the participant stance, but the implication does not run the other way. There are many other reactive attitudes, besides betrayal, which would indicate the presence of the participant stance. If we rely with a readiness to feel one of those, then we rely from the participant stance, but we do not necessarily trust. As Baier tells us, we need to distinguish trust from the reliance of comedians and advertisers, among others (1986, 234-5). They may rely on others and do something from the participant stance without trusting them. The comedian may rely on their audience for a warm reception. If it is not given, they may feel a crushing sense of having been judged unworthy by their fellow human beings and may come to resent their audience for not appreciating them. These are reactive attitudes only appropriate for people, not mere features of the world, so imply that the participant stance has been taken. Yet this kind of reliance from the participant stance is not trust; the comedian has not been betrayed and their attitude does not make them vulnerable to betrayal. Similar considerations apply to the advertiser. They rely on prospective customers for their success. Like the comedian, they may come to feel resentful of those prospective customers for their poor taste if they do not buy the advertised products. They take the participant stance towards those on whom they rely, but do not trust them. A further example comes from Katherine Hawley, whose view we will discuss below. She describes a case of regularly cooking for one's partner and their coming to rely on this. Such reliance is from the participant stance, since it is appropriate for the partner to feel grateful that one is looking after their nutritional needs. This, however, should not be a matter of trust. It should not be considered even a minor betrayal if one does not cook one evening; nor should it impugn one's trustworthiness. (Hawley 2014, 7-

8) Thus, one expects to be relied on from the participant stance, but being trusted would be unwelcome. Reliance from the participant stance is therefore insufficient for trust.

Perhaps it will be answered that these are the wrong kinds of participant stance, that the kind of stance required for trust is that which comes with a readiness to feel betrayal, not just any reactive attitude. This brings us to the second reason for thinking the current theory incorrect, which stems from the lack of clarity surrounding the participant stance itself. As it stands, it seems like circular reasoning to distinguish between trust and mere reliance by appealing to the participant stance, if betrayal is the only permitted reactive attitude. The thinking seems to be like this: trust is a kind of reliance, but unlike ordinary reliance, it must be vulnerable to betrayal. Therefore, trust will be defined as reliance with a readiness to feel betrayed if let down. Such an account does not tell us anything about betrayal, except that it is a particular reactive attitude. If we are to avoid circularity, then a more precise account of the (right kind of) participant stance is required, which rules out examples like those of the comedian and the advertiser just given, while ruling in cases of readiness to feel a sense of betrayal. Such precision does not seem to be Holton's aim – he admits that his account is 'nothing like a reductive analysis of the stance of trust' (1994, 67) – but it is needed to avoid the first objection and to clarify what it means to trust.

It is not clear what such an account would be like, or if it could avoid being simply arbitrary. But even if such an explanation can be given, there is a final reason for thinking that the participant stance account is taking the wrong approach, which I believe is quite decisive. That is that it takes betrayal to be a reactive attitude, a feeling. It is an affective response to an external event, not an external event itself. Yet this is surely not correct. There may well be characteristic feeling associated with betrayal – a sense of shock, coupled with anger and confusion – but this is not, or at least not always, what betrayal fundamentally is. It is not what we make ourselves vulnerable to when trusting – or rather, it is not the only thing. When we are betrayed, we typically suffer some kind of loss or harm quite apart from how we might feel. Whatever we trusted the other person with – some important task or precious object – is lost to us. Appealing to the participant stance cannot account for the material aspect of betrayal.[9]

It was mentioned above that this theory seems able to capture the required vulnerability to betrayal. We can now see that this is an appearance only. To make the point clearer, consider the fact that it is possible to be double-crossed without ever realising it. One will not *feel* betrayed, but will have *been* betrayed. In a similar way, it is possible to feel betrayed without having been betrayed; you might think that somebody has deliberately frustrated your trust, but in fact it was other factors entirely that has led to your current predicament. Since the feeling of betrayal can come apart from the act of betrayal – which is primarily what we make ourselves vulnerable to in trusting – the two cannot be identical. So the participant stance does not fulfil that necessary condition for trust; it is still insufficient as a theory.

We turn next to a theory that follows on from the goodwill account, developing it into something more sophisticated than reliance on a broad positive disposition. Since it moves

[9] There need not always be a material aspect. The harm of betrayal may be (non-exhaustively) psychological, emotional, practical, epistemic, or any combination of these, corresponding to what one trusts the other to do. The point here is that the participant stance only encapsulates a small part of what being betrayed entails.

away from the idea of trust differing from reliance in a purely affective way, it may be a step in the right direction.

Double-Layered Reliance

A view developed by Karen Jones (2012; 2017) is that trust involves a kind of higher-order reliance. Like the goodwill account, it is based on imputing to the one trusted an attitude about the trustor which moves them to act in the desired fashion. Specifically, it involves relying on them to be moved by our reliance on them to do something.

More formally, this idea can be stated as follows:

> A trusts B in domain of interaction D if and only if, A has an attitude of optimism that B's competence and responsiveness to the fact that she is being counted on will extend to cover that domain.

(Jones 2017, 99)

I take it that there is no salient difference between counting on someone and relying on them. Jones goes on to tell us that '[t]his dual structure of dependency – counting on the other, in a domain, and counting on them to respond to the fact that we [are] counting on them – is the heart of trust.' (2017, 100)[10]

Something to note about this account is that it is explicitly three-place; trust is not just an attitude that one person has towards another, but an attitude that they have to another about something else. For Jones, this 'something else' is a domain of interaction, a very broad notion that reflects the breadth of things with which we can trust others. I take the understanding of trust as a three-place predicate to be correct.

To turn to the details of this account, let us consider first whether it stands up to the two necessary conditions laid out at the start. First, it does seem to allow for betrayal, as distinct from mere disappointment. If B not only fails to do what she is counted on to do, but fails even to be moved by A's reliance, then this seems to qualify as an act of betrayal. B does not even try, so wilfully lets A down. Furthermore, unlike the participant stance account, this does not depend on betrayal being a reactive attitude, but captures the material element of betrayal as well. Second, there is a clear sense in which this allows for trust to involve the recruitment of another's agency to cooperatively fulfil one's goals. The supposed reasoning of the other person forms part of the content of trust, on this view. Their agency, understood as requiring responsiveness to reasons,[11] is thus explicitly encapsulated in the theory.

Another hurdle that it clears is that which felled the original goodwill account: it can avoid the implication that con artists trust their victims. At first glance, it seems that the double-layered reliance theory must have this unwelcome implication, for a con artist can surely rely on and

[10] Presumably, the attitude of optimism, on Jones' view, entails reliance. Otherwise, it is hard to see how optimism about another's responsiveness to being counted on could entail the dual structure of dependency – reliance – that she takes to be the heart of trust. It is this reliance, rather than the optimism, that I shall focus on. See also her (1996, 1).

[11] Being responsive to reasons is commonly thought of a necessary condition for agency, though I will not argue the point here. See, for instance, Fischer and Ravizza (1993, 338-9; 1998, 37) and Marušić (2015, 196).

be optimistic about their victim responding positively to their reliance. Presenting themselves as an honest trustor, they claim to be in need of assistance or perhaps spin some plausible story about an investment opportunity. They tell their prospective victim that they are depending on them for it, all the while relying on the victim's responsiveness to that perceived dependence to manipulate them into doing what they want. But this displays a too-simplistic understanding of the theory. This is where the importance of having trust as a three-place predicate is shown. The con artist relies on their victim in a particular domain. Call this D_1. D_1 pertains to their nefarious purposes and may involve anything from talking the victim out of a small sum of money to tricking them into giving up the entirety of their savings. However, the victim, who sees the trickster as an honest person, does not take them to be relying on them in D_1. Depending on the nature of the con, they may not consider the other person to be relying on them at all. If they do think that they are being relied on, it will be in some other domain, D_2. D_2 will pertain to the way in which the con artist has presented themselves; it may involve being in great need, or wanting to cooperate for mutual benefit. Either way, the domain of interaction in which the con artist relies on their victim is not the same as the domain to which they hope the victim will respond. By Jones' theory, then, the con artist does not trust their victim. [12] Genuine trust, on this view, requires transparency about what the trustor is hoping for. This is one advantage of taking this more precise approach over the much vaguer and more general idea of goodwill.

A simpler but related objection goes as follows. Suppose that I want you to do something for me, but I am not within my rights to demand it. You never said that you would and it's not in any way your responsibility. Nevertheless, I ask you to do it and make it clear that I am relying on you for it. My reliance is unwanted and uninvited. It may be, for instance, that I want you to write a favourable reference for me to help with my job application, but, since I have left it very late, you do not think that you will have the time. I am very clear on what I need from you and why; there is no mismatch of domain. Nor is there any trickery involved. We both know that you have no particular duty towards me in this regard; you don't owe me anything. But still, the very fact that I am relying on you might move you to make time in your schedule to write the reference. Perhaps you do it grudgingly, all the while resenting me for putting this extra pressure on you. You do not do it primarily because you have my interests at heart, but because you do not want to be made to feel guilty about it later. My genuine reliance moves you to act in the way I am relying on you to act. I employ this technique to get what I want, relying on you to be so moved.

On Jones' theory, it looks as if I am trusting you. I am relying on you to be moved by my reliance. But surely this is not right. Although there is no deception on my part, this is still a kind of manipulation. I am trying to guilt-trip you into doing something that you should not have to do. We should be able to mark the difference between trusting and manipulating, just as we should be able to distinguish between a trustworthy person and someone who is easily manipulated (although doubtless plenty of people fall into both categories). If the double-layered reliance theory fails to do so, that is a serious problem. It shows that, while the theory is meant to encapsulate the other person being concerned for the needs of the trustor, something that we will often look for and expect when placing our trust, by that very same idea, it has

[12] Jones does not emphasise the domains having to match, but it does seem to be entailed by her view. If it is not something she believes, then it is a natural amendment to her theory.

some unpalatable consequences. The positive motivation of responding to others' needs and dependency can be too easily exploited by others.

One way of answering this objection may go as follows. Being relied on can be a powerful motivator and Jones' theory does indeed imply that relying on another to be so moved counts as trust, even in a manipulative context. It is hard to see how this implication is to be avoided without adding an 'except for manipulative reliance' clause into the account, which would be arbitrary and require further explanation. However, the fact that this counts as trust does not in itself mean that it is a bad theory. Trust can be unwelcome and uninvited. The fact that one should not trust in a given situation does not mean that one is not really trusting at all. So, this response goes, the person who manipulatively relies on another's responsiveness to reliance is trusting them. We just need to acknowledge that trust can sometimes be manipulative.

Now, it is certainly true that trust can be immoral – members of mafia organisations may trust one another. It may also be inappropriate if it is not invited or wanted. This does not prevent the trust from being real. Whether it can be manipulative is contentious; manipulation seems to preclude the cooperative recruitment of agency that characterises trust. Maybe certain kinds of manipulation do recruit agency in some ways – they rely on the other's responsiveness to reasons – but this intuitively does not seem to be the right kind of agential interaction. It is also hard to see how, in the above example, it could count as a betrayal if you did not write the reference for me, no matter your motives and no matter how clear my reliance.[13] This, I think, brings us to the heart of the problem with the double-layered reliance theory. Let us grant – and I think it plausible if trust is voluntary[14] – that trust can sometimes be manipulative, that as misplaced and inappropriate as it might be, there are certain situations in which we really can adopt the attitude of trust in order to manipulate someone. Jones' account does not help us differentiate between the kind of trust that is vulnerable to betrayal and the kind of trust that is not so vulnerable, because it is uninvited and manipulative. There are at least some cases of double-layered reliance that are not vulnerable to betrayal. Perhaps that means that they are not cases of trust. But even if they are, a good theory ought to be able to separate them from cases of more appropriate trust, which this theory does not. Furthermore, at least some such cases would be blatant and deliberate manipulation, in which nothing like a trusting attitude is adopted, yet there would be this double-layered reliance.

What seems to be missing is reference to some undertaking on the part of the trustee. In the above example, trust would be more appropriate had you agreed to write the reference. If trust can be invited or uninvited, then, what constitutes an invitation? Below, we turn to a theory which aims to answer this question.

Commitment

Katherine Hawley gives an account of trust based on keeping commitments. As we shall see, it deals well with the problems that have been raised so far. The theory is stated as follows:

[13] Further illustrating the difference between *being* betrayed and *feeling* betrayed.

[14] As was argued in the previous chapter. Jones (1996, 15) does not agree that trust is voluntary, although I am unsure if she continues to hold this view for her later ideas about trust.

> 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.

(Hawley 2014, 10)

A commitment, in this context, is a kind of voluntarily incurred duty. A paradigm case is a promise, but a commitment can also be made by other means. It need not even be explicit; one might sometimes commit by remaining silent, or by behaving in a certain way. What exactly constitutes the making of a commitment is largely an empirical matter, depending on the background societal conditions, assumptions, and expectations placed on various people (Hawley 2014, 11; 2019, 72). It may sometimes be vague, so that it is unclear whether someone has made a commitment or not. We will not explore the question of what conditions must have been met for one to have made a commitment. For us, it suffices that there are such things, for instance promises, with which we can undertake duties owed to specific others. It is worth being clear about what a commitment is in this context, since it is an ambiguous term. It can also refer, for instance, to a firm intention, but this is not a duty owed to anyone else.[15] We also talk of being committed to a cause, or to a person, or to an idea, but these are not the commitments relevant here.

To turn to the problems that have affected the previous theories we have considered, the commitment view does seem able to deal with the con artist case. The con artist does not trust their victim, because they do not believe that the victim has a commitment. Now, in some cases, it may be thought that the con artist has extracted a commitment from their victim, so Hawley's view implies that they can trust them. For instance, they have persuaded someone to promise to hand over a large sum of money at a later date because they do not have it currently. However, promises extracted through trickery are not binding. Although the victim does not realise it, they have not really made a commitment at all.[16] The con artist, on the other hand, does realise that the commitment is not genuine because they know that it was made under a misapprehension. Therefore, they do not trust their victim. (Hawley 2014, 12) So, a distinction can be drawn between trust and manipulation.

The same holds true of manipulation through uninvited trust. On this view, there is a clear sense in which trust might be invited or uninvited: the trustee has or has not made a commitment, respectively. So, even if manipulative trust sometimes counts as genuine trust, this theory has the resources to distinguish between 'good' (welcome, invited, wanted) trust and 'bad' (unwelcome, uninvited, unwanted) trust. The latter cannot be betrayed, since betrayal requires breaking a commitment, which cannot happen if there is no commitment (see below). Sometimes, there might be genuine trust without commitment, if we honestly but mistakenly believe that the other person has made a commitment to us. But in cases like the one in which I relied on you to write me a reference and relied on your being moved by said reliance, there is no real trust, since I knew that you were not committed to doing so. Had I believed that you had committed when you had not, this theory would entail that I did trust you, but that my trust was misplaced, so not the sort of trust vulnerable to betrayal.

[15] Russell Hardin (1996) uses the term in this fashion in his discussion of trustworthiness, rather than using it to refer to anything promise-like.

[16] It is curious, but I think correct, that one does not always succeed in making the commitments that one takes oneself to be making. The matters of uninformed and involuntary 'commitments' are discussed at greater length in the next chapter.

There is also a clear account of distrust that readily fits with this view of trust, which Hawley states it as follows:

> 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)

Thus, there is clear space for an attitude that is neutral with respect to trust: not believing that the other person has a commitment. This does fit with ordinary experience. One defence against an accusation of betraying trust is, 'I never said that I would do that!' If one never committed, the trust was mistaken and inappropriate. Likewise, the involvement of a commitment draws a firm distinction between trust and reliance, as well as between distrust and non-reliance. Consider, for instance, Hawley's example of cooking for one's partner mentioned earlier. This should not be considered a matter of trust or distrust because there is no commitment involved. If, on the other hand, one *had* committed to cooking for one's partner, then it would become a matter of trust, and not doing so would be a case of betrayal and impugn one's trustworthiness. When we merely rely (or not rely), we do not take there to have been a commitment on the other's part to perform the task in question. Only once a commitment has been made does either trust or distrust become appropriate.

The commitment view also encapsulates the necessary conditions mentioned at the start. There is a clear sense in which it recruits another's agency, for only an agent can commit. There is no restriction on the reasons for which they might make the commitment, or on their supposed motives for fulfilling (or not fulfilling) it, but a commitment must be made of one's own volition. Therefore, trusting someone, on this theory, is a way of recruiting agency. This account also allows for the required vulnerability to betrayal. If someone deliberately breaches a relied-on commitment, this would be an act of betrayal, fitting neatly with common intuitions on the matter.

Nevertheless, there is a potential problem with this approach. Although it can explain why the con artist who tricks a victim into making a commitment does not trust the victim, there are similar cases which cause trouble.

To see this, consider a case in which a commitment is made voluntarily, but then fulfilled by force, coercion, or manipulation. Suppose that someone promises that they will do something and their interlocutor, sceptical about whether the promise will be kept, takes steps to ensure that they are well motivated. This hardly seems like trusting behaviour. If, say, I blackmail someone into doing what they have committed to doing, either because they are unlikely to do it on their own, or because I do not want to even give them the chance of going back on their word, then I believe that they have a commitment and rely on them to fulfil it. Yet I do not trust them. Note that, unlike with the con artist, the commitment was not extracted by trickery, so we cannot simply say that it was invalidated by the circumstances under which it was made.

This indicates that we need one of two things if a commitments-based account is to be successful. We could stipulate further conditions that would rule out reliance based on coercion and other techniques that bypass the other's agency. For instance, it could be required that there are certain motives that are to be imputed to the trustee, as in the goodwill and double-layered reliance accounts. Alternatively, we need a more sophisticated account of commitments, on which a commitment is voided not only if it is made involuntarily, but also if it is fulfilled

involuntarily. I favour the second option, since I believe that the concept of commitment, properly understood, has the resources to avoid the problem. Adding further conditions would therefore be an unnecessary complication.

As we will see in the next part, Hawley's commitments-based approach is very close to the mark. With some adjustments, I believe that this type of theory can be successful. I turn now to the theory of trust which I favour.

Part 2: Trust as Reliance on Trustworthiness

In this second part, I will present and motivate a theory of trust according to which trusting someone is relying on them to be trustworthy. This in itself should be a relatively uncontroversial idea; trust and trustworthiness are intertwined concepts, so we should expect them to be related and that our view of one will affect our view of the other. Understanding one ought to further our understanding of the other. Despite this, explicit discussion of trustworthiness does seem to be missing from certain accounts of trust.[17] Putting the other's trustworthiness at the heart of what it means to trust will, I believe, remedy the problems that the above accounts face.

Although the idea of trust as reliance on trustworthiness is unlikely to be controversial, there is no shortage of disagreement among philosophers over what it means to be trustworthy, as we shall see. For the sake of unpacking my account of trust, I will present my preferred theory of trustworthiness in this chapter without substantial justification. For detailed discussion and defence of that theory, turn to the next chapter.

Motivating the Approach

Let us begin by acknowledging that imputing trustworthiness to the other person in some fashion is a very plausible part of what it means to trust them. To compare it to other ethically-charged attitudes that we sometimes adopt towards others, consider praise and blame. When we praise someone, we tend to see them as praiseworthy. Similarly, when we blame someone, we tend to see them as blameworthy. What exactly this 'seeing as' amounts to is no doubt debatable, but these attitudes do seem to involve distinctive ways of representing their targets. That the other has this 'worthiness' of an attitude is part of the content of that attitude.

We have argued previously that trust is not a kind of belief. Accordingly, this account of trust involves a reliance on another's trustworthiness, rather than a belief that they are trustworthy.

[17] I am by no means the first to realise this, however. Hardin (2002, 22-32) and Jones (2012, 61-2) both think it important to discuss trustworthiness in the context of trust; Jones also brings it up in her (2017, 94-9). Hawley (2019, 74) relates her theory of trustworthiness to her (2014, 10) theory of trust.

But even if it is not a belief, it would be very odd to think of trusting someone without some kind of representation of them as trustworthy.

Trust as reliance on trustworthiness also makes for a clear distinction between trusting and merely relying. Although trust is a type of reliance, it is a matter of relying on another to behave in a very particular way: not only to do what they are relied on to do, but to be trustworthy in so doing. This is an important distinction. The difference between merely doing something and being trustworthy in doing it should become clearer in due course.

This account also meets the conditions set out at the start. Trust involves vulnerability to betrayal because if we rely on someone to be trustworthy and they prove to be untrustworthy in the matter, we can reasonably consider them to have betrayed us. It involves the recruitment of another's agency because being trustworthy involves the exercise of agency.

Now, to get a little more precise, I take trustworthiness to be a three-place predicate like trust. Thus, we do not merely say that some agent B is trustworthy, but that B is trustworthy for A with respect to a behaviour ϕ. This is because it is possible to be trustworthy for some people and not others and with respect to some behaviours and not others. Thus, my account of trust can be spelled out more fully as follows:

> A trusts B to ϕ if and only if A relies on B to be trustworthy for A with respect to ϕ.

Below, I will give an account of trustworthiness that will illuminate this view of trust. For now, it suffices to mention that, at minimum, trustworthiness requires acting of one's own volition and in such a way that one can be held responsible (Hieronymi 2008, 224).

This way of viewing trust deals well with the standard problems encountered so far, so compares favourably with rival accounts. Firstly, the con artist does not trust their victim because they do not rely on their victim to be trustworthy. If we take seriously the requirements just stated, it does not seem plausible that being trustworthy is compatible with being tricked or manipulated. Those who are manipulated into doing things are not genuinely doing them of their own volition such that they can be held responsible; they may consider themselves to have no choice, or have been misinformed. Secondly, it captures what was intuitive about the goodwill and participant stance accounts without running into the same problems. Rather than reliance on generalised goodwill, it is reliance specifically on trustworthiness. This is quite different from not only the reliance of the con artist, but also that of the comedian who relies on a warm reception, the advertiser who relies on sales, and the partner who relies on a hot meal. None of them display particular reliance on another's trustworthiness. At the same time, being trustworthy for someone does seem to involve a kind of positive attitude or response to them; something like goodwill, but more specific. This also rules out various other instances of force, coercion, or manipulation. Finally, it also provides a natural way to accommodate distrust. Of this, more will be explained later, but it involves imputing untrustworthiness to the other person, rather than trustworthiness.[18]

One possible objection that could be raised to the theory is that it seems to imply that we can trust people whom we have no business trusting. That is not to say that they are not trustworthy

[18] I believe there to be an important difference between being positively untrustworthy and merely failing to be trustworthy. This may become clear in the below discussion, but it is discussed more explicitly in Chapter 3.

people, but that it is not our place to trust them. Suppose, for instance, that you make a promise to someone and I overhear it. I might then rely on you to fulfil that promise – for some reason, I also have some stake in the matter. Yet I do not really trust you; your promise was to another, not to me. I am not mistaken in this; I know that you have not promised me anything. But it seems that I am still relying on your trustworthiness. Since this is so, but I am not trusting you, the theory proposed above must be wrong.[19]

But this is a mistaken view of the matter. It is quite true that I would be relying on your trustworthiness without trusting you, but the more precise version of the account does not imply that I would be trusting you. As trust and trustworthiness are three-place predicates, in order to trust you, I must rely on you to be trustworthy *for me* with respect to whatever it is you promised to do. But in this case, I do not rely on you to be trustworthy for me, but to be trustworthy for someone else. This may be an interesting attitude closely related to trust – perhaps I will become indignant on the other's behalf if you break your promise – but my theory does not imply that it is an attitude of trust.[20]

This theory, then, seems highly plausible thus far. It is worth acknowledging, however, that it is not entirely new. Although they do not always explicitly say so, several philosophers have included trustworthiness as part of the content of trust in their theorising.

For instance, Russell Hardin (2002, 1) takes the view that trust is a belief that another will encapsulate one's interests within their own. That is to say, to trust someone is to believe that they will consider it to be in their own interests to uphold your interests. He goes on to argue that trustworthiness is encapsulating another's interests within one's own (Hardin 2002, 28). It follows from these views that to trust is to believe that the other person will be trustworthy.

Or again, take Karen Jones' view. We have already seen that she takes trust to be optimism about and reliance on the other to be moved by the reliance placed on them. She also argues that being trustworthy is taking that reliance to be a compelling reason to act (Jones 2012, 70-1). Thus, trusting someone is essentially a matter of relying on them to be trustworthy, rather like my theory.

Katherine Hawley's views also fit this pattern. She takes trust to be a belief that someone has a commitment, combined with reliance on them to fulfil it (2014, 10). Elsewhere, she argues that trustworthiness is avoiding unfulfilled commitments (2019, 73). Therefore, we can infer that trust is a belief that someone has a commitment, combined with reliance on them to be trustworthy in this particular case.

Finally, there is Pamela Hieronymi's view (2008, 224) that trust is a belief that another person will do something based on the their practical reasons to be trustworthy. Their practical reasons become our evidential reasons for the belief. Although Hieronymi does not take trustworthiness to be part of the content of trust, only a part of the justifying conditions, the general idea that one must consider trustworthiness in order to have a complete account of trust is clearly present.

[19] For similar considerations, see Moran (2005, 22-3) and McMyler (2013, 1067), who discuss the significance of being addressed by a speaker, as opposed to overhearing them.

[20] The issue of trusting when the promise has been made to someone else is discussed at greater length in Chapter 5, as 'The Third Party Problem'.

In all of these ideas of trust, the one trusted is represented as trustworthy. As I mentioned earlier, the idea of trust as reliance on trustworthiness is likely to be uncontroversial; several philosophers of trust have already had very similar ideas, even if they have not explicitly stated them. The main disagreements arise when we consider what kind of attitude trust is and what it means to be trustworthy. I have argued that it is a kind of reliance, but Hardin and Hieronymi think of it as a kind of belief. Jones and Hawley agree that trust involves reliance on the other to be trustworthy but disagree over what constitutes trustworthiness.

The previous chapter was dedicated to determining what kind of attitude trust is, so I shall not return to that issue in detail here. The basic idea was that belief is justified only by epistemic reasons, whereas trust also permits of practical reasons, which indicates that trust is not a kind of belief. Trust does require belief in the case of trusting testimony, but, it was found, reliance also entails belief in those circumstances. Therefore, reliance is a better fit, since it is justified by both practical and epistemic reasons. In the next section, I shall expound in more detail what I take trustworthiness to be.

The Account in More Detail

Since trustworthiness is a part of the content of trust, it is necessary to understand what it is to be trustworthy before we can fully appreciate what it is to trust. In this account, trustworthiness has a certain priority over trust. This, I think, is as it should be. Trust may be valuable and useful for cooperation, but only so insofar as there are trustworthy people in whom it can be placed. It is worse than useless when it is placed in those who are not trustworthy. It is fitting, then, that we have a keen focus on trustworthiness in our explanation of trust.

Of course, we must also remember that our purpose here is to present an account of trust. To avoid getting drawn too far off topic, I defer more detailed discussion of the nature of trustworthiness to another chapter. Here, however, is the account that shall be argued for:

> B is trustworthy for A with respect to ϕ if and only if B has committed to A that they ϕ and B ϕs.

One thing to note about this theory is that it is a fairly specific notion of trustworthiness. It is meaningful to talk of someone being a generally trustworthy person. However, we are concerned with trustworthiness on a given occasion. This is what usually matters to us, after all, when we trust someone in that instance.

With this theory in place, we can also state a fuller version of the proposed theory of trust:

> A trusts B to ϕ if and only if A relies on B to fulfil their commitment to A that they ϕ.

Like Hawley's view, this is based on commitments. However, I hope to demonstrate that it nonetheless avoids the problem of neglecting the perceived motives of the one trusted.

It can also be reiterated that the advantages outlined above do indeed hold. It rules out the con artist's 'trust', since the victim has no genuine commitment. As was argued for Hawley's view, the commitment is not valid if it was extracted through trickery. By the same token, it requires the recruitment of agency; plausibly, only agents can make and fulfil commitments. It is

vulnerable to betrayal because, as with Hawley's view, the one trusted could deliberately break their commitment and this would be an instance of being betrayed.

Since it is similar to Hawley's theory of trust, it shares the strengths of that account. But for the same reason, my theory must answer to the same problems as hers. The main one, as we saw, was that it does not rule out the possibility of manipulation and coercion being counted as trust, since someone might be manipulated or coerced into fulfilling a commitment that they have previously made voluntarily. So, can one be compelled by a trustor to fulfil their commitment on this view?

At first glance, it looks as if they can. Just as we said with Hawley's theory, someone could fulfil their commitment purely out of fear of what the other person might do otherwise, which is clearly not the same as being trustworthy, nor would it count as being trusted.[21] However, I think it is possible to avoid this implication by looking more closely at what a commitment requires.

In making a commitment, one is, in a sense, making an offer. One offers to do the thing in question for someone else. But offers need not be accepted. Someone may reject the offer, in which case there is no obligation to carry it out. Commitments require acceptance, or uptake, if they are to be binding.[22] What I suggest is that forcing (or coercing, or manipulating) someone to do that which they have voluntarily committed to doing represents a rejection of the commitment. The coercer essentially makes it clear to the one who commits that they are not prepared to take their word for it; that they are going to take steps themselves to ensure that it gets done. To signal that one will not take another's word on whether they will do something is to reject their word; to refuse to accept the commitment. Since a commitment requires acceptance to be binding, it is voided if the agent in question is coerced into fulfilling it. Therefore, the view I have presented does not imply that those who are coerced or manipulated into fulfilling their commitments are either being trusted or being trustworthy.

Hawley could use the same defence, of course. If, on her view, commitment requires uptake and coercion amounts to rejection, then she can also say that being coerced into doing what one committed to doing does not really count as fulfilling a commitment. So her commitments-based account survives the potential problem that was raised for it. However, I do think that my theory has two advantages over hers. First, it does not require belief in the other's commitment, only reliance on them to fulfil it. This is simpler and also more accurate for, as will be argued in the next section, trust without belief in a commitment can still be genuine, but mistaken, trust. Second, I believe that my theory of trustworthiness, the content of trust, has certain advantages over Hawley's, as will be shown in the next chapter.

To return to our main line of argument, this aspect of the theory allows it to rule out cases of coercion, manipulation, and other instances of being compelled against one's will. In such cases, one is not fulfilling a commitment or being relied on to do so. Rather, one's commitment has been rejected and so one no longer has a commitment to fulfil.

[21] Note that this is a case in which one fails to trust because the other is not (seen to be) trustworthy. This lends further support to the view that trustworthiness has a certain priority over trust.
[22] See Thomson (1990, 296-8).

There is more to be said about what it means to have a commitment and to be trustworthy. Once again, I shall defer this matter, lest we stray off the main topic of this chapter. I will say more on the subject in its proper place. For now, I turn to the issue of distrust. It was mentioned above that my theory can be extended to cover the attitude of distrust. Now, with a theory of trustworthiness in place, we can begin to see why.

Distrust

In this section, I will consider how the account of trust I have been advocating can be extended to the notion of distrust. As discussed earlier, it is desirable that a theory of trust can also provide the resources to explain the attitude of distrust. To start, I shall make some general remarks about the nature of distrust and how it relates to trust, reliance and non-reliance. I will then apply those ideas to the theory of trust as reliance on trustworthiness.

We should begin this discussion by noting that distrust is an attitude of a different character to mere non-reliance, as noted by Hawley (2014, 3). In this way, it is analogous to trust. The context needs to be suitable for trust. It is not appropriate for Kant's neighbours to trust him to walk by at the same time every day (Baier 1986, 235); nor is it appropriate to trust a shelf to bear the weight of a vase (Hawley 2014, 2). Rather, only reliance is appropriate, though for slightly different reasons. In the one case, Kant is not being in any sense recruited, but merely used by his neighbours. His failure to turn up one day would not count as a betrayal. The shelf, on the other hand, should not be trusted because it is an inanimate object with no agency of its own.

Similarly, distrust is not appropriate in these cases. If Kant's neighbours come to doubt that he will turn up one day, or that he will arrive at a different time and potentially mislead them, or if we think that the shelf will probably fall or break under the weight of the precious vase, there will be a refusal to rely. But this non-reliance will not amount to distrust. There is a sense in which trust and distrust are both appropriate in the same kinds of context: where we are considering whether to recruit another's agency in such a way that we would be vulnerable to their betraying us. We should therefore be asking not only when the context is appropriate for trust rather than mere reliance, but when the context is appropriate for trust-or-distrust rather than mere reliance-or-non-reliance.[23]

To apply these general thoughts to the account in question, the kind of reliance one has in trusting is reliance on another to be trustworthy. Therefore, the kind of reliance that is withheld in distrusting is also that in another's trustworthiness. More than that, distrusting involves withholding reliance because of their perceived untrustworthiness – having made a commitment but not being disposed to fulfil it.[24] As for the context appropriate to trust-or-distrust, our theory can also provide an answer. Given the account of trustworthiness that has been stated and the fact that this is embedded within the account of trust, we can say that it is appropriate to trust-or-distrust just in case the other person has made a commitment. Without a commitment, trust would be misplaced. If trustworthiness involves fulfilling a commitment and trust is reliance on trustworthiness, then how can we (appropriately) trust someone if they have not made a commitment to us? Take the examples of Kant and the shelf again. On the

[23] This approach is also taken by (Hawley 2014, 6; 9; 2019, 19)

[24] As it stands, this is somewhat vague. More will be said of untrustworthiness in the next chapter.

theory I have proposed, these can be explained by the fact that in neither case has a commitment been made. Kant never said that he would help his neighbours to know the time, nor did he promise that he would take a walk past their houses at the same time each day. That is just something he did for his own purposes. The shelf, as an inanimate object with no agency, is incapable of having made any commitments. Incidentally, this again highlights the fact that this theory fulfils the requirement that trust involves the recruitment of another's agency. A commitment is a self-imposed obligation owed to another party, so making one and following through on it of one's own volition surely displays normative agency.

Does this mean, then, that trust must involve a belief that the other has a commitment, like Hawley says? Not necessarily. Trust may be mistaken in a variety of ways. The clearest is trusting the untrustworthy, but as we have seen, it would also be a mistake to trust a shelf, or for a certain Prussian community to trust Kant. We can mistakenly trust (or distrust) by thinking that someone has made a commitment when they have not, but we may also mistakenly trust without such a belief; we might realise that the other has no commitment but, being confused about what trust entails and perhaps not realising that 'I never committed to that' is a valid excuse when accused of betrayal, we trust them anyway. Recall the case in which I was expecting a letter of reference from you when you have not committed to writing one. In that example, I was not really trusting you, but guilt-tripping you. However, suppose that I do not have much insight into what I am doing and do not consider my attitude or behaviour to be manipulative. I think, mistakenly, that your being trustworthy requires you to do me this favour and I rely on your trustworthiness. It seems that in this case, I trust you, though I have no right to. This might be a fairly niche example which will not occur very often. But I think it an advantage that this theory allows for it, whereas Hawley's does not. We should not make the mistake of thinking that irrational or inappropriate trust does not count as trust at all.

We can say, then, that distrusting is withholding reliance in another's trustworthiness when they have made a commitment. This is different from merely not relying, since, as with the distinction between trust and reliance, it involves a commitment. Furthermore, what is not relied upon is not merely the other's performance, but their trustworthy performance. In this way, we can extend our ideas about trust to account for distrust in a way that mirrors the attitude of trust in a satisfying way. The distinction between distrust and mere non-reliance is in some ways analogous to that between turning down an invitation and never receiving one in the first place. The primary result may be similar, but the circumstances and the kinds of attitudes involved are quite different.

My theory can, then, provide the basis for a plausible account of distrust as well as of trust. The two fit together in a way that might well be expected, being, as Jones (1996, 15-6) puts it, contraries but not contradictories; it is possible to neither trust nor distrust. What is more, both take into account the trustworthiness of another; all three concepts are part of a holistic theory. Once again, this seems to be an advantage. They are the kinds of concepts that ought to fit together and our thinking about one ought to influence our thinking about the others.

Trusting Testimony

In the previous chapter, it was argued that trust is a special kind of reliance that entails belief in the specific case of trusting testimony. Let us now consider whether the more precise account of trust that has now been developed can fulfil this requirement.

Recall that we took testimony trust to be a kind of act trust. When we trust someone in what they say, we trust them to do something specific: speak knowledgeably and sincerely (Holton 1994, 73-4). Or, more accurately, we trust them *to have spoken* knowledgeably and sincerely. It is possible to speak truthfully without speaking knowledgeably and sincerely; one might speak on a subject on which one knows nothing, or even deliberately try to deceive, but still accidentally say something true here and there. However, I will here treat the two as equivalent; we are not currently concerned with cases of accidental truth-telling. For present purposes, testimony trust is trusting someone to have told the truth.

On the theory of trust as reliance on trustworthiness, this amounts to relying on someone to have been trustworthy in their speech. This seems highly plausible. Being trustworthy in speech is presumably what is meant by honesty and it seems right to say that to trust someone's testimony is to rely on them to be honest. To be more specific, though, our theory implies that testimony trust is relying on another to fulfil their commitment to telling the truth, since that is what follows from unpacking the embedded account of trustworthiness.

This gives rise to a potential problem. Many people make assertions or give testimony without promising, or otherwise making commitments, to tell the truth. There are exceptions, of course. Think of witnesses about to testify in court. But that is not the usual practice. Can this theory account for trust in ordinary cases of testimony, given that it is heavily dependent on the idea of a commitment having been made?

Fortunately, this objection is easily dealt with. Speaking – or signing, or writing, or otherwise asserting – does ordinarily come with a commitment to honesty. It was mentioned earlier that a commitment need not be explicit; it can remain unspoken and even unacknowledged. It is commonly accepted among philosophers that the norms of assertion do include an implicit commitment to speak honestly, at least as far as one is able (Hawley 2019, 50-4). Again, there are exceptions. Consider an actor speaking a line on stage. The line may well express something false, since it relates to the fictional world of the play and not to the actual world. Indeed, it may even be false within that fictional world, if the script has that character lie at this point in the play. But clearly the actor is not guilty of lying. The reason is that at least some of the usual norms of assertion, including the commitment to honesty, are waived in these conditions. There is a general understanding among the audience and the actors that the usual commitments to honesty do not apply (Kenyon 2010, 353). Such cases highlight the fact that ordinary assertions do involve commitment; the presence of a commitment in most cases explains what is missing in the case of the actor that makes it perfectly acceptable for them to knowingly speak falsely.

Let us return to the central question of this section. Does the account of trust given above entail belief in the case of testimony? Making the appropriate substitutions, we might state the implied account of testimony trust formally like this:

> A testimony trusts B if and only if A relies on B to have fulfilled their commitment to speaking truthfully.

This does indeed entail belief. On this understanding, A will not really be trusting B unless they believe B. The argument that was given in more detail previously applies here with little emendation, but I will briefly rehearse it.

If A is relying on B to fulfil their commitment to speaking honestly, then they are relying on B to speak honestly. This is an epistemic reliance; what is at stake is the truth of A's beliefs, rather than some material loss or harm. Therefore, in order to be genuinely relying – putting the truth of beliefs at stake – A must believe what B asserts. By the same tokens, this instance of trust will not be under one's direct voluntary control, since one cannot directly choose to have or not have beliefs. Nor will this instance of trust be rationally responsive to practical reasons, but only epistemic ones. Therefore, the proposed theory can indeed fulfil the requirement of testimony trust entailing belief.

This is not unique to my theory, of course. It was mentioned in the discussion of the participant stance approach that the arguments I used to show that reliance can entail belief could be used by Holton to adapt his theory. Any reliance-based account could likewise appeal to the points I have appealed to in order to show that it can encapsulate belief in the case of testimony trust, so long as the particular details of the account do not preclude them. I have demonstrated how the arguments concerning reliance on testimony fit with the theory of trust as reliance on trustworthiness. There is a commitment to tell the truth in ordinary cases of assertion, and reliance on trustworthiness in assertion entails belief.

Conclusion

We began this chapter where we left the previous one: with the view that trust is a kind of reliance. We also assumed that trust is a way of recruiting another's agency for cooperative and beneficial action, in such a way that it leaves one vulnerable to betrayal from the one trusted. We then went on to consider four influential theories of trust. Although each had a certain appeal, none were free from objections.

Baier's goodwill account was not able to handle the case of the con artist; it implied that con artists genuinely trust their victims by relying on their goodwill. Holton's view faced the problem that its basis, the participant stance, was not sufficiently precise to distinguish trust from other attitudes, like those a comedian or advertiser adopts towards their audiences. The double-layered reliance view advocated by Jones was could not properly distinguish between appropriate trust and manipulative reliance. Hawley's commitments-based theory seemed to work best. There was no decisive objection to it, since the problem of being coerced or manipulated into fulfilling commitments can be solved. However, I believe that there is a simpler approach, which also encapsulates a more suitable theory of trustworthiness.

Learning from each of these, the account of trust as reliance on trustworthiness avoids the significant problems, but builds on the strengths of the theories considered. Like the goodwill and double-layered reliance accounts, it can explain the view that trust often comes with some positive disposition on the part of the one trusted; what they are trusted to do must always be undertaken willingly. It avoids the vagueness of the participant stance approach, leaving it open whether or not a trustor might take such a stance. No doubt they can and often do, but it need

not be part of the definition of trust. Finally, it adopts and simplifies Hawley's view that commitments are central to trust and trustworthiness. The theory manages to avoid the problems faced by others, while remaining intuitive and relatively simple. It can capture the starting idea, that trust is a kind of reliance that recruits another's agency and is vulnerable to betrayal, as well as meet the requirement that testimony trust entails belief.

We began with two main questions: What distinguishes trust from mere reliance? What exactly are we relying on when we trust? We now have ready answers to both. Indeed, both have the same answer. What distinguishes trust from reliance is that, in trusting, we rely specifically on another's trustworthiness.

There is still more to do, of course. The theory of trustworthiness that I have proposed, being central to the theory of trust, needs to be further justified and explained. We will cover this issue in the next chapter. Later, in Chapters 4 and 5, we will also consider what it takes to justify trust.

All of these considerations follow from this being a holistic theory. It does not merely pick out trust in isolation, but locates it among a set of other attitudes and concepts. Trust is one of several interlocking concepts which fit together like pieces in a jigsaw. Although I think that the theory I have proposed should be accepted, it should not be expected to stand on its own, any more than a piece of a jigsaw puzzle ought to be considered a complete picture. Indeed, for this very reason, I think we should be suspicious of any theory of trust that does stand alone; it is part of the nature of trust that it connects to other concepts like trustworthiness, without which it cannot be fully understood.

This does mean that I have not yet made the complete case for my theory of trust. That cannot be finished until the other crucial elements have been established. For now, however, I believe that a reasonable start has been made. The place occupied by trust within this wider theory has been found and even if my account of trustworthiness proves incorrect, it seems likely that trust as reliance on (the correct idea of) trustworthiness is the right starting-point.

3

Trustworthiness as Person-Specific Reliability

Introduction

In the two previous chapters, it was argued that trust is a kind of reliance. Specifically, it is reliance on another to be trustworthy. In order to complete the account, we must therefore give a theory of trustworthiness. This, I hope to show, is an interesting and valuable question even without its crucial place in the theory of trust.

Trustworthiness pervades our everyday lives on almost all levels of interaction, from intimate personal relations to business agreements to large-scale societal organisation. Being trustworthy, in various contexts, entails honestly sharing our knowledge, keeping our promises and doing our part in joint endeavours rather than taking advantage of others' efforts. The concept of trustworthiness clearly has implications for ethics, epistemology, and theories of rational action.

Despite its importance, it is not as much discussed as its counterpart attitude of trust.[25] The purpose of this chapter is to give an account of trustworthiness that increases understanding of the concept and which accommodates the main features attributed to it. I draw on Katherine Hawley's (2014; 2019) work on the subject, proposing a commitments-based view of trustworthiness. Being trustworthy, at its heart, is a matter of fulfilling the commitments one has made and avoiding unfulfilled commitments. This chapter provides a theoretical foundation for such a view, but differs in some respects from Hawley's version of it. It will be found that the version I propose is better able to deal with certain challenges that commitment-based views face.

I begin with the idea that, as trust is a kind of reliance appropriate only for persons, so trustworthiness is a kind of reliability specific to persons. Drawing on the idea that the crucial difference between persons and objects is the autonomy of the former, this person-specific reliability will be shown to entail the keeping of commitments. The aim is to not only argue that commitments play this central role in the concept of trustworthiness, but also to show how it is based on the simple and relatively uncontroversial idea that to be trustworthy is to be a

[25] That is not to say that there is no discussion on the matter. See, for instance, Hardin (1996), Jones (2012) and especially Hawley (2019), whose work is a heavy influence on what follows.

reliable person. The result is an intuitive theory that is distinct from Hawley's, but vindicates her core view.

I begin by taking an overview of the main features of trustworthiness and motivating the idea of trustworthiness as person-specific reliability. Next, I consider what it means for something to be reliable in a more general sense, arguing that its meaning depends on the kind of object to which it is applied. The reliability of a car, for instance, is different to the reliability of a boat; the one consists in driving well, the other in sailing well. The concept will then be applied to persons. Given the foregoing discussion, this will not be a straightforward matter: in what does reliability consist for a person? It will be argued that person-specific reliability – that is, trustworthiness – is a matter of fulfilling one's commitments. We will then return to Hawley's commitment-based account and argue that this version is an improvement over it. This will be done by considering how we each respond to certain challenging cases, specifically the statuses of the uncommitted and of those who are willing but unable to fulfil their commitments. It will be shown that the account presented here can give more satisfactory responses.

Part 1: Trust, Reliance, and Reliability

In this part, I will lay start building the theoretical foundations for my theory of trustworthiness. I begin by motivating the idea that trustworthiness is person-specific reliability. I will then offer an account of reliability in general, which will later be applied to persons. This part ends with a consideration of two objections to my preferred approach.

Motivating the Account

It is common for trust to be considered a kind of reliance (Baier, 1986, 240; Holton, 1994, 66-7 Hawley, 2014, 2). Specifically, it is a kind of reliance that is only appropriate for people. We may reasonably rely on objects for various purposes, but we do not properly trust them. This is not a distinction that we always make in our language; we do often talk of trusting things like our cars or our phones. But such talk is generally loose or metaphorical; if the matter is highlighted, we are quick to see that the 'trust' we place in impersonal objects is not the same as the trust we place in our friend, colleague, or doctor. Following this, I take trustworthiness to be a kind of reliability that can only be had by a person, not by an object. This idea of trustworthiness as person-specific reliability is the starting-point for my account.

It is also common to take trustworthiness to be a three-place predicate: 'A is trustworthy for B with respect to ϕ', where A and B are agents and ϕ is a behaviour.[26] This mirrors the common view that trust is three-place: 'B trusts A to ϕ'. We do not simply trust, but trust a person to behave in a certain way.[27] Similarly, one can be trustworthy for some people and with respect

[26] For instance, see Hardin (2002, 60) and Jones (2012, 70-1).

[27] See, for instance, Baier (1986, 236), Jones (1996, 5-6) and Hawley (2014, 1) for examples of three-place theories of trust. Domenicucci and Holton (2017) are exceptional in arguing that trust is fundamentally two-place. One assumes that they would say the same of trustworthiness.

to some behaviours but not others. A *mafioso* may display a great deal of trustworthiness to their mafia family and others involved in their criminal enterprises, but very little to the police, business-owners, and others in the local community. Incidentally, this also highlights the fact that trustworthiness, though it is often thought of as a moral good, can be turned to destructive and immoral purposes. To take another example, we may well be inclined to consider our doctor trustworthy in matters relating to our health; we may be less inclined to consider them trustworthy if we are seeking financial advice. Being trustworthy with respect to certain actions does not entail being trustworthy in all one's dealings.

In a similar vein, the scope of trustworthiness is variable, as it is for trust. We can talk of trusting someone in a broad sense, or of trusting them to do something specific. Similarly, it makes sense to talk of someone simply being a trustworthy person, or of them being trustworthy with respect to a certain domain of interaction[28], or with respect to a specific action or behaviour. For instance, a colleague may be a generally trustworthy person, or they may be trustworthy with respect to work-related interactions, or trustworthy with respect to some specific thing you have asked them to do.

Another noteworthy feature is that trustworthiness applies in both speech and action. We can trust someone to do something, but we can also trust someone's testimony. Typically, trusting testimony is taken to be a case of trusting someone to do something specific: roughly, to trust them to speak honestly (Holton, 1994, 73; Hieronymi, 2008, 219; Hawley, 2014, 16-7). I say 'roughly' since there is disagreement over the precise relationship between act trust and testimony trust, but this is not the place to explore that issue in detail. Based on this, we can say something similar about trustworthiness. We can be trustworthy regarding our actions and regarding our speech. The latter is a special case of the former; when we speak honestly, that honest speech is the trustworthy action.[29]

It will be shown that the theory of trustworthiness I offer can encapsulate each of these features. In giving examples, I will usually focus on cases in which the agent in question is trustworthy (or not) with respect to a specific action for ease of exposition, but 'ϕ' can stand for something more general or vague, as when a doctor is trustworthy regarding health-related matters, or if someone can be trusted to do 'the right thing'. We will consider the specific issue of trustworthy testimony towards the end of this chapter.

Having given these preliminaries, I will now motivate the idea of trustworthiness as person-specific reliability. Recall the examples of trustworthy behaviour mentioned above: sharing knowledge, keeping promises, doing our part. Each of these is an instance of being reliable in some way. It also seems correct to say that trustworthiness belongs only to people. Consider the following example:

> Suppose I trust you to look after a precious glass vase, yet you carelessly break it. I may feel betrayed and angry; recriminations will be in order; I may demand an apology. Suppose instead that I rely on a shelf to support the vase, yet the shelf

[28] This is Jones' (2012, 62, 70-1) preferred formulation, which she reiterates in her (2017, 99).

[29] There are certain complications. If we have promised to lie in order to help someone else, then there would seem to be a tension between being trustworthy with respect to promise-keeping and being trustworthy with respect to being honest. However, the tension resolves when we consider that trustworthiness is three-place: by lying, we would be trustworthy for one person, but not another.

> collapses, breaking the vase. I will be disappointed, perhaps upset, but it would be inappropriate to feel betrayed by the shelf, or to demand an apology from it.

(Hawley, 2014, 2)

It is clearly not appropriate to blame the shelf or to accuse it of betrayal. We should not say that it has been untrustworthy, though we may say so in the case of the person who carelessly breaks the vase. Similarly, we might be glad if the shelf does not collapse and perhaps relieved if we had entertained doubts. But we should not feel gratitude in the way we might if a person had been very careful with the vase and saved it from damage. Such reactions are appropriate only for persons; objects cannot be either trustworthy or untrustworthy, though they may be reliable or unreliable.[30] We take the person, but not the shelf, to be responsible, since the person possesses agency of their own.

Although only persons can be trustworthy, it is possible for a person to be merely reliable without being trustworthy and to be unreliable without being untrustworthy. Kant is said to have taken walks with such regularity that his neighbours could set their clocks by him. He was very reliable for anyone who wished to check that they had the right time. But he was not being trustworthy.[31] He was not taking his walks in order to signal the time; he may not even have known that he was thus relied upon. Like the shelf on which the vase is placed, it is not appropriate to feel grateful or to praise his trustworthiness in coming by, however helpful it might be. Similarly, if he did not take his walk one day, or took it at a different time, then he would have been unreliable in that instance. Those relying on him would be disappointed and perhaps misled, but they may not complain of being betrayed or accuse Kant of behaving in an untrustworthy manner.

Thus far, I have not asserted anything very controversial. I have merely stated and motivated ideas that almost all philosophers of trust would agree with. The crucial question is what makes trustworthiness distinct from mere reliability, just as the crucial question for trust was what makes it distinct from mere reliance. My own view is that being trustworthy involves fulfilling the commitments that one has and avoiding having unfulfilled commitments. In this way, it is similar to that expounded in Hawley's (2019) *How to be Trustworthy* and much of what I have to say could be used in defence of her ideas. However, there are also some important differences between our views. In drawing these out, I hope to show that my theory has certain advantages over hers.

I will now begin laying the groundwork of my theory of trustworthiness. This will involve finding a definition of reliability that suits our purposes and applying it to persons. It will be found that no alteration is needed to the core concept of reliability in order to move from reliability to trustworthiness.

A General Analysis of Reliability

What does it mean to be reliable? This question needs to be made more precise, since an object or person might be reliable with respect to some tasks but not others. The shelf that is reliable

[30] These points are inspired by Strawson's (1974, 4-20) discussion of reactive attitudes, which Holton (1994, 66-7) also makes use of in his theory of trust.

[31] I borrow this example of reliability without trustworthiness from Baier (1986, 235).

for holding a vase may not be reliable holding a heavy set of weights. Kant may have been a reliable timekeeper, but that does not mean he would have been, say, a reliable cook. We should therefore ask: what does it mean to be reliable for a given task?

To begin, we might reflect that in ordinary usage, being reliable is often synonymous with being something that can reasonably be relied on. If I say that I can rely on something to accomplish some task for me, I mean that I find it reliable with respect to that task. So, as a first attempt, we may say that something is reliable just in case it does what it is relied upon to do. Or, more formally:

x is reliable with respect to ϕ if and only if when x is relied upon to ϕ, x does ϕ,

where x is an object or person and ϕ is a task.

This may seem intuitive, but closer inspection reveals that it does not fit well with ordinary usage. Consider, for instance, a car that frequently breaks down when driven on normal roads. When relied on for driving, this would be an unreliable car. But suppose that, though it has not been designed for it, this car happens to float on water very well. It could be driven into a lake and its occupants would remain safe and dry. Does this make it any more reliable? Would we call it a reliable car? The above definition suggests that we sometimes should: it is reliable for its driver with respect to floating, though not for driving. This seems to some extent correct, inasmuch as someone who relied on it to float would not be disappointed. However, when someone talks of a car being reliable, we do not, unless in very specific circumstances, take them to mean reliable with respect to floating. If we were sold a car having been assured that it is reliable and only later discovered that it was reliable for floating and not driving, we would consider ourselves to have been cheated; that is just not what is normally meant by a reliable car.

The key point here is that, when it comes to judging reliability, it matters what the thing in question is for. There is clearly something wrong with this car, but an advocate of the definition suggested would struggle to explain what. All they could say is that it is reliable for floating but not for driving. But then, how could we distinguish between a reliable car and a reliable boat? We expect different things of reliable boats and reliable cars, but on this definition, we cannot justify those intuitive expectations. To make sense of what it means for something to be reliable, we need to consider what it is for. Since a car is meant to drive but not float, we assess its reliability by how well it drives; whether it floats or not is normally irrelevant to whether it counts as a reliable car.

So, let us make a second attempt. Being reliable is not a matter of simply doing whatever something is relied on to do, but a matter of doing what it is meant to do. For a car, this would be driving rather than floating. We may put this idea as follows:

x is a reliable F if and only if x is an F and x does what an F is meant to do,

where x is an object or person and F is a predicate.

So, we might say that some object is a reliable car just in case it is a car and it drives well. We can also apply this to a person performing some role: someone is a reliable employee just in case they are an employee and they perform their job well. This seems to work better; we can now distinguish between a reliable car and a reliable boat easily, since cars and boats are meant to do different things. A potential difficulty is that what something is meant to do on the

relevant sense of 'meant to' can sometimes be difficult to determine. I take it, however, that there is a fairly clear sense in which cars are meant to drive and boats are meant to float. We shall come back to this question below, when we discuss the reliability of persons.

Note that this conception of reliability dispenses with the action; it is no longer reliable with respect to a particular task, but with respect to some predicate. The car that is reliable for floating but not for driving is thus not reliable *as a car*. This is another important point that shall be discussed further in due course. For now, let us accept this second account.

It should also be noted that adopting the second account of reliability does not imply that the first account is entirely wrong. It does make sense to say of some car that it is reliable for floating – and it may be pertinent if its driver must cross a body of water and has no other means of doing so. But, in addition to fitting more closely with normal usage, the second idea is better suited to our purposes. This is because the idea we are hoping to explore is that of person-specific reliability: what does it mean to be reliable as a person? If reliability were taken to be relative to an agent and a task, we could not answer this question. The personhood would not matter; when determining if someone was reliable, we would only need to know the relevant task and check whether they did it when relied upon. The question of reliability would be addressed in the same way for persons as for objects and so there would be no such thing as person-specific reliability. Therefore, we take reliability to be not a predicate, but a predicate-modifier. In this way, it is like a common usage of 'good'.[32] When we say that someone is a good teacher, for instance, we do not generally mean that they are both a teacher and good; we mean that they are good *at teaching*. Similarly, when we call something a reliable car, we do not usually mean simply that it is a car and also reliable; we mean that it is reliable *as a car*. The reliability of a car consists in driving well. The reliability of a boat consists in sailing well. The reliability of a person will consist in something else.[33]

Two Objections

I will here divert briefly from the main line of thought to consider a couple of potential problems with the proposed idea of reliability. The first, it will be shown, is misguided, but will serve to highlight the plausibility of 'reliable' being a predicate modifier. The second is also unsuccessful, but provides an opportunity to examine the preferred view of reliability more closely.

The first objection is to the above claim that 'reliable' is similar to 'good', in that what it means will depend on the predicate to which it is attached. Consider the following propositions:

(1) Black cats are black.
(2) Good assassins are good.

[32] See, for example, Finlay (2014, 19-21). Whether this is the fundamental meaning of 'good' is a substantial and much-debated philosophical question and not one that will be answered here.

[33] If being a reliable car consists in driving well, then what is the difference between being reliable and being good? The difference is not precise, but can be made somewhat clear. In a car, reliability consists simply in being able to drive under normal conditions. Being good goes beyond this. A good car displays excellence in at least some of a variety of features – comfort, safety, handling, fuel efficiency, speed, luggage capacity and so on. Reliability is plausibly a necessary condition for this non-moral sense of 'good', but it is not the same thing.

(3) Reliable assassins are reliable.

Of the first two propositions, (1) is clearly a tautology but (2) is not – it even seems in tension with itself, since those who are good at killing people for money are plausibly not especially good people. There is thus a clear syntactical difference between (1) and (2), though they superficially have the same form. This difference consists in the fact that 'black' is a predicate, whereas 'good' is a predicate modifier. The one means the same thing whatever it is applied to; the other means something different depending on how it is used. In particular, 'good' on its own tends to denote moral goodness, whereas in 'good assassins', it indicates being good at assassinating.

Now, the objector asks, is (3) a tautology? Intuitively, it seems that it is. To deny it would be a contradiction. Therefore, (3) is like (1) rather than (2). Since it is tautologous, 'reliable' must be a predicate like 'black' rather than a predicate modifier like 'good'. Because the proposed view of reliability depends upon its being a predicate modifier, it cannot be correct.

However, this objection is misleading. Consider (2) again. As was explained, it seems non-tautologous because when 'good' appears before a noun, it is taken to be attributive – in this case, good at assassinating – but when it appears on its own, it is typically taken to be moral. But this latter point does not always hold. Suppose that someone is deciding on which assassin they should hire. They might say, 'We will use this one; she's very good.' Here, the context implies that 'good' is to be understood as attributive, not moral, despite it not being attached to a noun. Or suppose that another assassin, finding themselves out of work in a frustratingly harmonious period, takes up freelance maintenance work in order to make ends meet. Someone deciding which maintenance worker to hire might then say, 'I'll use the assassin; he's very reliable.' Even though he is referred to as an assassin and not as a maintenance worker, the context implies that it is his reliability in the domain of maintenance that is being referred to rather than in the domain of assassinations or even just general reliability. So, the context can alter the meanings of these terms. (2) might be taken as tautologous if the context implies that the second 'good' is attributive; (3) might be taken as non-tautologous if the context implies that the second 'reliable' is attributed to something other than assassinating. We should therefore not read too much into the examples of this objection; the propositions will not always have the meanings they appear to when read in isolation.

A second objection concerns the question of how specific trustworthiness is. Recall that it was observed that trustworthiness is a three-place predicate of the form, 'A is trustworthy for B with respect to ϕ'. That is, one may be trustworthy with respect to some behaviours but not with respect to others, as well as for some people but not for others. In contrast, the preferred account of reliability is relative to a predicate, rather than relative to a task and an agent. If trustworthiness is a kind of reliability, however, surely it ought to have the same form. Rather than 'x is a reliable F', we should have 'x is reliable for B with respect to ϕ'.

There are two reasons to reject this view. The first is that a three-place account of reliability would fall into the same problems as the initial account given above. It would be far too relative; something would be reliable insofar as it does what it is relied upon to do. So, again, we could not say that the car that floats well but drives badly is unreliable if it is relied on for floating. Since this version is relative to agents as well as to tasks, however, the problem gets even worse. Suppose that, by an even stranger quirk of its manufacture, a certain car responds well to some people but not others. Depending on who is in the driving seat, it may refuse to

start, frequently break down, or develop various faults. Again, we would be inclined to call this an unreliable car, since reliable cars should respond well to whoever is in the driving seat, so long as they are a competent driver. But if reliability is three-place, we could not do so without qualification. We could not say that it is reliable *as a car*, only that it is reliable for certain agents and with respect to certain tasks. This is another instance of an account that can make sense, given the appropriate clarifications. But it does not fit well with ordinary usage, or suit our aim of giving an account of being reliable *as a person*.

The second reason is that three-place trustworthiness does not necessarily entail three-place reliability. All that is required is for being reliable *as a person* to be three-place. This can certainly be done. Being a reliable F entails doing what an F is meant to do, which may have further argument-places, depending on what 'F' stands for. So, to satisfy the required form of trustworthiness, we need only show that what a person is meant to do, in virtue of being a person, involves both a task and another agent. This will be demonstrated in due course.

Part 2: Person-Specific Reliability

We must now ask what it means to be a reliable person. Following from the discussion above, this involves asking what it is that a person is meant to do. A person, of course, may occupy many roles, with varying reliability. Someone may be, for instance, a reliable employee but an unreliable parent. That is, they may perform well in their job for their clients or superiors, yet do less well in the tasks relevant to parenting for their children. Or to take another example, Kant was a reliable timekeeper, but this does not entail his being reliable in any other respect. People can be reliable or unreliable as teachers, plumbers, scientists, students, or prime ministers, since these roles involve different tasks. The question that concerns us here, however, is not what makes someone reliable in any particular role they may have, but what makes them reliable simply as a person. This means that we will need to determine what one is, in the relevant sense, meant to do in virtue of being a person.

One way to approach this is from a general perspective, by exploring the issue of what determines what an F is meant to do. It is fairly intuitive that a reliable car drives well; this is the most familiar usage of the phrase 'reliable car'. But how might this be applied to persons? We will consider first what seems to be an intuitive answer, based on intentions. However, this shall be found to have several problems and rejected, so we will take a different approach instead.

Basing Reliability in Intentions

As an initial attempt, then, we might say that if some object or person x is meant to ϕ, then there must be someone who intends x to ϕ. After all, saying that something is meant for some task is often equivalent to saying that it is intended for that task. If that is the case, then reliability is dependent on an agent having a particular intention. This leads us to ask, who is this intending agent who determines what something is for? It cannot be the relying agent, since

that leads us back to the problem of something being reliable for anything; the car that floats in water would be perfectly reliable even in the second sense, so long as the driver intended it to float well. In the case of a car, it seems more intuitive that the intending agent will be the designer. Part of being a designer is plausibly deciding what the thing designed is meant for. A car's reliability consists in its driving performance because its designer intended it to be used thus. But who plays the role of the designer/intender for persons? Theists may be inclined to say that God, as our creator, is the one who designs us and determines what we are for. However, the concept of trustworthiness should not depend on God; whether theism or atheism is true, it is possible to be trustworthy. Furthermore, a divinely ordained purpose seems a rather too weighty matter to fit with how we usually talk of people's trustworthiness. A more promising idea is that for a person, unlike cars and most other objects we may rely upon, the relevant agent is the person themselves. Given our autonomy, what we are meant to do is up to us, as what a car is meant to do is up to its designer. This is rather neat if we accept the intention view: what a person is meant or intended to do is just what they intend for themselves to do. They get to decide the tasks which others may rely on them to perform.

This attempt does give an intuitive answer – an answer, moreover, that I do believe to be very close to being correct. However, the view that reliability is dependent on an intending agent comes up against a number of difficulties. Firstly, in the case of persons, it relies on a move from 'they are meant/intended to ϕ' to 'they mean/intend to ϕ'; if a person is meant to ϕ, it is claimed, then they themselves intend to ϕ. But this is not always true. Consider, for instance, the case of an arranged marriage. The two people may not want to get married, but are, in a certain sense, intended to be married. Their families intend for them to get married, but one or both of the 'couple' may have every intention of refusing to do so. So this crucial move in the reasoning of the first attempt does not work.

Secondly, there is sometimes no salient agent at all. It certainly makes sense to talk of reliable and unreliable parents, for example, but there may be nobody who is intending the parents to do certain things, such as ensuring their children get to and from school safely. In the case of reliable parents, one might say that the agents are (as with persons) the parents themselves; reliable parents presumably do intend to ensure the safety and wellbeing of their offspring. But, by the current view, we would also need to say that unreliable parents are intended to do this. Otherwise, their failure to do so could not make them unreliable, as a car's failure to float does not entail its reliability. But of course, an unreliable parent may have no intention of, say, picking up their child from school. There is no guarantee that anyone would fulfil the position of intender for parents in general. Similar things may be said of many other roles a person might occupy.

Thirdly, this view gives an inaccurate picture of trustworthiness. It implies that reliability in persons consists in their doing what they intend to do. If trustworthiness is person-specific reliability, this implies that being trustworthy is a matter of following through on one's intentions. This is quite clearly incorrect. Suppose that I make a promise without any intention of fulfilling it. I am still untrustworthy for breaking it. Conversely, if I currently intend to have fish for supper, but then change my mind, I am not thereby being untrustworthy, even if someone were to rely on my having fish. As far as I was aware, it was no one's business but my own. Following through on intentions could be called being *resolute*. Being resolute may overlap with being trustworthy, but they are not the same thing.

This idea will not work, but we can learn from its failure. There is a correct idea lurking within it: a person has some measure of control over what they are meant to do. As shown by the third problem, this is not full control, but we do have the power to alter what we are meant to do in a way that cars and boats cannot. We must therefore find a way of incorporating this measure of control into person-specific reliability, without allowing it to simply be a matter of following through on intentions.

A Better Approach to Reliability

My preferred approach is based on what makes persons relevantly different from other things that we may rely on. In the current context, this means what it is that makes it appropriate to trust persons, but not objects. Fitting with common themes in the literature, I will suggest that the key difference is a person's autonomy. As Jones (2012, 65) points out, trust is a way of recruiting another's agency, not merely getting them to do something. Or as Domenicucci and Holton (2017, 151) say, trusting involves granting another a certain amount of discretion. Both are ways of appealing to the autonomy of the one trusted. It is plausible, then, that an important feature of person-specific reliability is an appeal to autonomy. Nevertheless, there is an important balance to be struck. Being trustworthy is not a case of exercising our autonomy by doing whatever we like. On the contrary, trustworthiness places certain constraints on our behaviour. I will strike this balance by showing how one's autonomy can be the source of those constraints.

For example, take a paradigm case of being trustworthy: keeping a promise. In making a promise, we make it the case that we ought to fulfil it. We cannot avoid promissory obligation simply by deciding that we will not honour the promise, but it is nonetheless a voluntary obligation. What a person is meant to do is to some extent under their control.

In general, I suggest that an F is meant to ϕ just in case it follows from what it means to be an F that it should ϕ. This works well for things like cars and boats. If it is not the case that some object should drive well, for instance, then that object can hardly be called a car. Similarly, something is only a boat if it should float and generally operate well on water. How this teleology is grounded is a question I leave aside. So long as the given kind of object has some purpose built into it, whatever the metaphysical explanation of that is, it is the kind of thing that can be reliable in the preferred sense and performance of the built-in purpose is the metric by which its reliability can be measured. This means that my account of reliability does not apply to objects that have no particular purpose. A rock, for example, cannot be reliable in this sense, since there is nothing that it should do that follows from what it means to be a rock. For such objects, we might revert to the initial account of reliability – which was not incorrect, just not so useful – and say that it is reliable for certain tasks. Or, we might say that rocks are reliable or unreliable not *as rocks*, but as certain other things – they may be reliable building materials, or hammers, or shelters. If you say that you have a reliable car, I will usually know what you mean without further clarification; if you say that you have a reliable rock, then I will need more information – in what way reliable? There is no rock-specific reliability in the way that there is car-specific reliability.

What is the sense in which a person may be reliable as a person, rather than in a role? Are there any kinds of behaviour and such that it follows from being a person that one should perform

such behaviours? Now, we must be cautious here. We have already mentioned that divinely ordained purpose is not what we are dealing with. Similarly, this is not an attempt to define what it means to be a person, nor an argument for any deep teleological claim about what humans are for, nor a discussion of the foundations of morality. Our aims are more modest. Whatever the merits of such discussions, they are beyond the scope of this chapter. Nevertheless, I hope to show that it does make sense to talk of our being meant to perform certain tasks, in the sense of it following from personhood that we should do them.

In the preferred account of reliability, being a reliable *F* entails a characteristic task. The problem is that there is no clear sense in which a person is meant to do something simply because they are a person. Certainly, there does not seem to be anything analogous to driving for cars or sailing for boats. We might be tempted to say, then, that a person is like a rock: one may be reliable only in the sense of the first account, or one may be reliable by the second account only within a given role rather than *as a person*. We have already mentioned that persons can be reliable or unreliable in certain roles, so perhaps that is the only way in which they can be reliable, just as a rock may be reliable as, say, a hammer, but not as a rock. If this is the case, then there is no person-specific reliability at all – and therefore, no trustworthiness.

Acquired and Innate Purposes

However, we need not take this view. What a person is meant to do in virtue of being a person may not be immediately clear, but it can be worked out. To help do this, consider the following thought experiment. Imagine that, perhaps far in the future, there exists a device known as the Infinitely Adaptable Machine (IAM). These machines are capable of doing or being virtually anything, depending on the programming that they are given. We can imagine that they possess advanced AI and can transform themselves into whatever form is required. They might be programmed to drive around like cars, or to sail on water like boats. They might be required to perform complex calculations, to care for the sick, or to come up with sound policies for governing a nation. In principle, what they can do is limited only by the laws of physics and logic.

Now, these IAMs are the kinds of things that can be reliable or unreliable in the sense we are using. If some of them were to start breaking down and not fulfilling their programmed tasks effectively, we would consider them to be unreliable. The better the IAMs perform in the tasks given to them, the more reliable they are. This is because there is something that they are meant to do: *whatever they are programmed to do*. This is, of course, an extremely vague specification, but by hypothesis, it gives us the criterion by which to judge the reliability of an IAM. If one of these machines does other than what it is programmed to do, then something has gone wrong and it lacks reliability.

The crucial point is this: when an IAM first comes off the factory floor, having never yet been programmed with any task, there is nothing that it should do. It is meant to do *whatever it is programmed to do*, but that set of tasks is empty. A well-functioning IAM without programming will therefore do nothing. Yet it is not like a rock; it can be reliable in the relevant sense. There is such a thing as IAM-specific reliability.

What this shows us is that there is a difference between what we might call an ‘innate purpose’ and an ‘acquired purpose’. A car has an innate purpose: it is meant for driving. An IAM has an

acquired purpose: it is not meant to do anything until a task is given to it. Yet that purpose still follows from the kind of thing it is. Unlike a brand-new car or boat, a brand-new IAM does not have a specific kind of task until it acquires one. However, if this task is acquired in the right way (by being programmed), then, unlike with a rock, it is a part of what it is to be an IAM that it should perform that task. Of course, it may still be reliable or unreliable in a way not specific to IAMs if it is used for something without being programmed for it. An unprogrammed IAM might make a reliable paperweight, for instance. It is not thereby being reliable *as an IAM*, just as the floating car we discussed earlier is not reliable *as a car*. The kind of thing that an IAM is does not directly tell us what specific tasks it is meant to perform, but tells us how such a purpose is acquired: by being programmed.

To come back again to our main line of thought, I suggest that a person is in some ways like an IAM: they start off without anything that they are meant to do, but can acquire purposes. I will argue that a person acquires a purpose, not by programming, but by choice. Unlike the other kinds of purposive objects which we have been discussing, be they cars, boats, or IAMs, a person is autonomous. As such, the kind of thing a person is tells us how a specific purpose can be acquired.

Again, I am not here concerned with morality in general. There are certain things that we ought morally to do and others that we ought morally to refrain from doing without us choosing what they are. Trustworthiness is no doubt closely connected to the broader moral framework – promissory obligation carries moral weight – but I here remain neutral on what this connection is. Reliability has a normative force to it; terms like 'meant to' and 'should' are not out of place. The extent to which this normativity becomes moral when applied to persons is not a matter I shall delve into.

I shall now lay out my argument concerning what person-specific reliability consists in.

1. For any x and for any F, if x can be reliable as an F, then there is some task ϕ such that x is meant to ϕ in virtue of being an F.

This first premise is merely an implication of the definition of reliability that was argued for earlier.

2. Given such an x, ϕ is either an innate purpose or an acquired purpose.

The above thought experiment and subsequent discussion of IAMs shows that this is true. Objects like cars and boats have innate purposes – specific things that they are meant to do that follow from what they are. IAMs, however, have acquired purposes – there is nothing that they are meant to do until they are programmed, but these programmed tasks are still things that they should do in virtue of being IAMs. A rock, on the other hand, is not such an x; it may be used for various tasks and have a degree of reliability for those tasks, but it is not the case that it should do anything just because it is a rock.

3. No person has an innate purpose.

This is a more precise statement of the point that caused the current problem. Unlike cars and boats, persons do not have anything specific that they are meant to do just because of the kind of thing that they are. It follows from these last two premises that, so long as there is such a thing as person-specific reliability, a person must be capable of acquiring a purpose or purposes.

4. Purposes may be acquired either voluntarily or involuntarily.

This I take to be an application of the Law of Excluded Middle. There might be vague borderline cases, in which it is not clear whether something is voluntary or not. However, the fact that it is not easy to know where the line is does not mean that it is not there, or that it does not matter. The issue of voluntariness will be discussed further in Part 3.[34]

5. A person cannot acquire a purpose involuntarily.

This is a particularly crucial premise, so I shall discuss it at greater length. It appeals to the fact that it matters how a purpose is acquired; that this is determined by the kind of thing the purposive object is. An IAM, for instance, is a programmable machine. Its purpose is to do whatever it is programmed to do. If it is relied upon for something other than that for which it has been programmed, then it is not being relied upon *as an IAM*. An unprogrammed IAM that is used as a paperweight, for instance, is not being used *as an IAM*, since that does not fall within the scope of *whatever it is programmed to do*. However, an IAM that has been programmed to act as a paperweight and is used accordingly is being used as an IAM, since in that case it is doing what it is meant to do. There is no difference in what the device is doing, but there is a difference in how it came to be doing it. The purpose must be acquired in the right way in order for what it does to count as being reliable as an IAM, otherwise the purpose does not follow from the kind of thing that it is.

Now, suppose that a person is a purposive object like an IAM: they have acquired purposes that follow from the kind of thing that they are. How do we discover how these purposes are acquired? We must consider what kind of thing a person is. An IAM is a programmable machine. A person, I assume, is an autonomous agent. An IAM acquires purposes by programming; a person must acquire purposes through their own autonomy. That is to say, a person chooses what they are meant to do and is reliable insofar as they do those things. All this is assuming that that a person can acquire purposes and so can be reliable in the sense relevant here; I am discussing what person-specific reliability would consist in *if* there is such a thing as person-specific reliability. This is the next premise, which will be justified in due course.

The point here is that an involuntarily acquired purpose is inconsistent with the kind of thing a person is. Persons can be put to various uses without their permission, but this would be like using an unprogrammed IAM as a paperweight. Reliance on a person to do something that they have not chosen to do is reliance that does not respect their autonomy. It is not reliance on them *as a person*. An extreme example of this is slavery. Someone might be used as a slave and may be highly reliable in that capacity. But this use does not respect their autonomy or their personhood. They are therefore not exhibiting person-specific reliability. In contrast, someone who does exactly the same things as the slave having committed to doing them – for instance, they are a hired worker – is being reliable as a person. Their autonomy is being exercised, because they have chosen what they are meant to do.

This demonstrates the correct idea within the intentions-based theory discussed above. It is true that persons, as autonomous agents, must have some level of control over what their purposes are – what makes them reliable. Once a purpose is acquired, however, it cannot be set aside at

[34] See the section entitled 'The Question of Competence'.

will. This will be discussed further below. For now, it suffices that if a person has a purpose, it cannot be acquired involuntarily.

6. Persons can be reliable *as persons*.

This is intuitive, but for now is just an assumption. It essentially states that there is such a thing as person-specific reliability. To justify it, we must show that there is something that persons are meant to do that follows from their being persons. As just argued, this means that they must have the capacity to choose their purposes. A reliable person voluntarily makes it the case that they are meant to do certain things and does them. This will be shown early in the next section.

All this entails the following conclusion:

7. A person acquires a purpose voluntarily.

Person-specific reliability thus consists of fulfilling the purposes that one has chosen; a person decides what they are meant to do. Relying on someone to do what they are meant to do therefore involves respecting their autonomy and recruiting their agency in the way characteristic of trust.

A Commitments-Based Account

If this is right, then trustworthiness, construed as person-specific reliability, is based on commitments. One commits to doing something just in case one voluntarily imposes an obligation on oneself to do it. More precisely, it is an obligation owed to a specified other, rather than a broad obligation owed to others in general, or to no one in particular. At least part of the purpose of the concept of reliability is to help identify what may be relied upon for its characteristic purpose. For this reason, reliability must be concerned with a potential relier, although it is not required for anyone to actually rely for something to be reliable. A paradigm example of a commitment is a promise, but there are others. This fits neatly with the requirements that have been set out. Commitments are made voluntarily, but are binding. Therefore, trustworthiness imposes restrictions on behaviour, and the source of those restrictions is one's own autonomy. [35]

I will not give a definitive account here of the conditions under which commitments are made. They can be implicit or explicit; they can be made actively or tacitly. It is possible to commit by knowingly going along with a convention without voicing any objection. This can make it difficult to know when a commitment has been made and it is easy to imagine arguments flaring based on disagreement over whether someone has committed to do something or not. For our purposes, it suffices merely that commitments do exist as I have characterised them. Note that this is all that is required to justify the assumption made in the sixth premise above. Persons

[35] The extent to which a commitment must be voluntary may be disputed. Hawley (2019, 49) writes, 'The social nature of language and communication means that in practice we do not exercise absolute authority over what we do with our words. We may sometimes make commitments we did not anticipate, or fail to make commitments we intended to incur.' It will be shown below that we sometimes do fail to make the commitments that we intend to make. I remain neutral on whether we can unintentionally commit. However, if we do something carelessly, we plausibly still do it autonomously; careless commitment is still, in the relevant sense, voluntary. What my view does require is that making a commitment precludes being forced, coerced, or manipulated.

can be reliable *as persons* just in case there is something that they are meant to do in virtue of being persons. As was pointed out in the justification of the fifth premise, this would have to be a voluntarily acquired purpose; so long as persons can have such purposes, they can be reliable in our sense. Commitments are voluntarily acquired purposes of persons, so the fact that persons can make them entails that they can be reliable on our preferred account.

Note how this account fulfils the ideas mentioned earlier in connection with the intention-based view that was ultimately rejected. The failure was instructive in that it highlighted that persons do not have total control over what they are meant to do. Trustworthiness does not consist in doing what we intend to do. The above argument is accordingly based not on intentions or any other attitude that one might have, but merely on the kind of thing a person is: an autonomous agent, capable of setting their own purposes. This entails that one who makes a commitment and then changes their mind, or who does not intend to fulfil it in the first place, is still bound by their commitment. Their attitudes do not affect their purposes. Furthermore, we observed that a correct idea embedded in that view was that a person should have some measure of control over what constitutes their reliability. On the commitments-based view, this is explained. Commitments are made voluntarily, so people can decide what makes them reliable. If I have made no commitment to doing something, then anyone relying on me to do said thing is like a driver taking their car into a lake. Maybe I will happen to do what they are hoping, but it will not be a measure of my reliability whether I do or do not – without a commitment, this is not what I am meant to do. Kant did what his neighbours relied on him to do by walking past their homes at the same time each day. In this respect, he was a reliable timekeeper, but he was not being a reliable person in our sense, since he had made no commitment to walk by at the same time each day. It was merely something he happened to do.

We are now in a position to state the account of person-specific reliability that I think we should accept. Recall first our preferred account of general reliability:

> x is a reliable F if and only if x is an F and x does what an F is meant to do.

We can apply this to persons, making appropriate substitutions:

> A is a reliable person if and only if A is a person and A *fulfils their commitments*.

As with IAMs, the task is specified only very vaguely, in contrast to cars and boats, which have quite specific tasks. This is to be expected from things with acquired purposes. By their nature, we cannot pin down a single specific task that they are all meant to do, but only state how what they are meant to do is to be determined. For IAMs, this is programming; for persons, this is commitment. For both, there is in principle no limit to what they may be meant to do, depending on their programming and commitments, respectively.

Another implication of this account is that one may be trustworthy for some people and not others and with respect to some actions but not others. For example, a *mafioso* is trustworthy for their don but not for others in the community; a doctor is trustworthy with respect to medical matters but not with respect to giving financial advice. Our account of person-specific reliability explains this. One is person-specifically reliable – that is, trustworthy – for another by fulfilling one's commitments to them. Members of the mafia will generally fulfil commitments that they have to their don, but probably not any commitments that they may have to local business owners. The doctor might fulfil any medical-related commitments that they have made, but perhaps not commitments connected to finance. The kind of commitment

and the agent to whom it is made matter. This account therefore encapsulates the common view mentioned at the outset, that trustworthiness is a three-place relation holding among two agents and a behaviour: A is trustworthy for B with respect to ϕ.

The mafia example raises another question: what is the relationship between trustworthiness and morality? We do typically think of being trustworthy as a moral good, yet it can clearly be used for immoral purposes. As mentioned above, I am not here concerned with morality in general, so I will only sketch a route to a possible answer. Recall that a paradigm case of being trustworthy is keeping promises. It therefore seems reasonable to think that the moral grounds of trustworthiness are those of promissory obligation, or at least similar. There is disagreement about what those grounds are and I will not adjudicate here.[36] But insofar as there is a genuine moral requirement to keep promises, there is, for the same reasons, a requirement to be trustworthy. This will come with similar exceptions as promise-keeping; there are times when morality requires that we break our promises, and there are plausibly times when we ought to be untrustworthy in other ways, such as lying. Although I do not here develop the relationship between morality and trustworthiness, it is likely to resemble that between morality and promise-keeping.

This account also fits well with the ordinary usage of 'reliable' as applied to people. When we call someone reliable or accuse them of being unreliable, it is usually in the context of us taking them to have a commitment or commitments. Typical examples of unreliability in people include missing appointments, being late for work, and not meeting deadlines. On our account, they are unreliable because they had commitments to do certain things that they failed to keep. Similarly, if we call someone reliable, it tends to be because they do what we take them to have committed to doing. This might be in relation to either a general tendency, or a particular instance of commitment-fulfilment.

On a related note, we can now also explain an overlap between trustworthiness and reliability within a certain role. Teachers, doctors, police officers, and various others who might be considered to have role-specific reliability are often, if they do their jobs well, considered trustworthy. This account makes clear why this should be. Since these are roles which they took on voluntarily and which entail certain commitments, by performing them well, they are displaying person-specific reliability as well as, say, teacher-specific reliability. By doing their jobs properly, they are being trustworthy, at least with respect to their role-related commitments. This stands in contrast to somebody who is forced into a role, who might perform it very well but is not thereby being trustworthy. They have made no commitment, so do not exhibit person-specific reliability in their role-specific reliability.

I will now give an explicit account of trustworthiness. The easiest way to do this would be to simply substitute 'A is a reliable person' for 'A is trustworthy' in the above account. Although this would be accurate, it would not be particularly informative. I shall therefore present my theory of trustworthiness in a slightly different format, which highlights the fact, emphasised

[36] Scanlon (1990; 1998, 295-327), for instance, embeds promises into a contractualist framework and argues that promissory obligation is derived from a general duty to not mislead or manipulate others. Thomson (1990, 294-321) thinks of giving one's word as allowing another to acquire a claim against one. Rawls (1971, 344-50) takes promising to be a social convention and argues that, given the existence of such a convention, it would violate the principle of fairness to break a promise. I leave it to the reader to apply their favourite theory of promissory obligation to trustworthiness, if they so wish.

just now, that one can be trustworthy for some people but not others and with respect to some tasks but not others. That there is no substantive difference and only a difference of form should be readily apparent.

> A is trustworthy for B with respect to ϕ if and only if A has committed to B that they ϕ and A ϕs.

Here, A and B are persons and ϕ is a task, action, or behaviour that can be construed however broadly or narrowly as is required, reflecting the flexibility of the concept of trustworthiness. One might be trustworthy with respect to performing some very specific action just once, or with respect to behaving in some way consistently over time. ϕ might even be substituted for something extremely vague, like 'doing the right thing'. As we have already stressed, there is in principle no limit to what behaviours a person might be trustworthy with respect to, so long as it is something they have committed to doing.

It will also be noted that this theory has trustworthiness as a three-place predicate, which satisfies the second objection raised at the end of Part 1. Although reliability is two-place, person-specific reliability, once unpacked, is three-place. A reliable person does what a person is meant to do and that is to fulfil their commitments. Since a commitment has to be made *to* somebody, this entails that what a person is meant to do, in virtue of being a person, involves both a task and another agent, as required.

Although this is the account of trustworthiness that I think should be adopted, following as it does from a plausible account of reliability applied to persons, there are some questions that it is natural ask of it, which have so far gone unanswered. Firstly, if trustworthiness requires having made a commitment, what is the status of the uncommitted? If, for some action, someone has made no commitment regarding that action, should we consider them trustworthy with respect to it or not? More broadly, what should we say of those who have never made any commitments?

Secondly, there is an important distinction between the examples of the *mafioso* and the doctor. The one chooses not to be trustworthy, but the other fails to be trustworthy because the task is beyond their competence. We can suppose that the doctor who (perhaps unwisely) commits to giving financial advice bears their client no ill will and tries their best, but nonetheless cannot do it. It is a different story with the mafia member and the local businesses. They deliberately exploit their victims, with never any intention to keep the commitments they have made. To take a third example, suppose a five-year-old child promises to make the most delicious cake ever for his mother's birthday. He is not competent to keep his well-meaning promise, but is he to be considered untrustworthy? It seems a little harsh to put an honest but failed attempt in the same category as deliberate lying and betrayal. Is it untrustworthy to break a commitment through incompetence rather than intentionally?

Thirdly, what does this account imply about testimony? Speaking honestly is one way of being trustworthy and lying is one way of being untrustworthy. Yet this should be so without any stated commitment to telling the truth. If someone speaks without signalling that they will tell the truth, as is quite normal, we should still be able to judge whether they have been trustworthy.

These questions will be addressed in the next section. I believe that the above theory of trustworthiness can provide a satisfactory answer to each of them. In so doing, I will also draw

out the differences between my theory and that presented by Katherine Hawley (2014; 2019), whose theories of trust and trustworthiness are also based in commitments.

Part 3: Questions and Challenges

In this section, I will address the three questions mentioned above, while also comparing my view to Hawley's. The questions will be taken in reverse order, starting with honest and dishonest testimony and ending with the issue of the uncommitted. My theory agrees with Hawley's as far as testimony is concerned, but I believe that it can offer better answers to the other questions.

The Question of Testimony

A paradigm case of being trustworthy is telling the truth.[37] But telling the truth is surely a case of being trustworthy even without a commitment to doing so. Similarly, lying is a paradigm case of being untrustworthy, whether one has committed to being honest or not. This therefore seems to be a counterexample to the idea that trustworthiness is a matter of keeping commitments.

However, this objection is mistaken. Recall that a commitment may be made implicitly or tacitly, such as when one goes along with a convention. In asserting, one commits to telling the truth, at least as far as one knows. This fits with one uncontroversial norm of assertion to speak truthfully.[38] Not all locutions are like this. Consider an actor who speaks a line on stage. Whether the proposition is true or not, it would not be appropriate to call them either trustworthy or untrustworthy. In that context, they have no commitment to speak truthfully and so their speech does not really count as assertion (Kenyon, 2010, 353). There is usually an implicit commitment to speak honestly, which can be set aside in certain contexts. Therefore, a commitment-based theory can deal with cases of assertion.

There are also examples of the opposite: sometimes, we do explicitly commit to being honest. For instance, a witness in court will promise outright to tell the truth, the whole truth, and nothing but the truth. What my account implies of such cases is that the speaker incurs a 'double commitment' to speak truthfully. Whatever wrong is done by breaching a commitment is therefore done twice by someone who lies under oath. This makes sense of the fact that being

[37] Assertion, telling, and testimony should here be understood as communicating to an audience. I am not concerned with writing in a secret diary or privately soliloquising, as discussed by Owens (2006, 117-9). I take there to be no important difference between asserting and testifying.

[38] There are many variations on this. Williamson (1996, 494-5, 502-8), for instance, advocates a knowledge norm, on which one should assert only what one knows; Lackey (2008, 125) advances a Reasonable to Believe Norm, on which one should assert only what it is reasonable to believe at least partly because it is reasonable to believe it. Whichever way the idea is best unpacked, there is general agreement that we should try to keep our assertions accurate. The precise content of the commitment one makes in asserting can be adapted to one's preferred theory of the norms of assertion.

dishonest in such circumstances seems to be worse than lying normally; that one is bound more strongly to tell the truth than one usually is.

Katherine Hawley has this to say on the matter:

> I claim that asserting as to whether p involves both
>
> (a) promising to speak truthfully as to whether p; and
> (b) speaking truthfully or untruthfully as to whether p, i.e. keeping or breaking the promise.

(Hawley, 2019, 51)

On this view, assertion is a matter of implicitly making a promise and simultaneously either keeping or breaking it. This, I think, is broadly correct, although my view does not require that the commitment be a promise. A promise is a type of commitment, but what makes something a promise instead of a different sort of commitment is not something we need to go into here. The important point for our purposes is that assertion or testimony involves committing to telling the truth, where that commitment is, in the very same act, either kept or broken.

The Question of the Uncommitted

Next, we turn to the matter of whether someone without the relevant commitment should be considered trustworthy. Suppose that one is relied on or trusted to ϕ, but has no made no commitment to doing so. If one does not ϕ, is one thereby untrustworthy with respect to ϕ? If one does ϕ, is one thereby trustworthy with respect to ϕ? At a more general level, if someone never makes a commitment, should they be considered an overall trustworthy or untrustworthy person? Either conclusion seems odd, since a commitment that is not made can be neither kept nor broken.

The problem here stems from a binary view of trustworthiness and untrustworthiness. I suggest that there is a third category that lies between them, which I shall call ‘non-trustworthiness’. To make clear what is meant by this, let us consider what it means to be untrustworthy on my view.

As trustworthiness is a kind of reliability specific to persons, so untrustworthiness is a kind of unreliability specific to persons. The account of untrustworthiness I favour can be derived with very similar arguments to those already given concerning person-specific reliability, so I will here give only a very brief version.

Reliability in general, it will be recalled, is a matter of something doing what it is meant to do. Paralleling this, unreliability is something failing to do what it is meant to do. Or, more precisely:

> x is an unreliable F if and only if x is an F and x does not do what an F is meant to do.

This seems to work well with objects such as cars and boats. A car is unreliable just in case it does not drive well; a boat is unreliable just in case it does not sail well. The boundary between doing something well and not doing it well is vague, but it does not have to be precise for the

account to be accurate. Objects like rocks, on this view, cannot be unreliable, for like reliability, there must be something that the kind of object in question is meant to do.

Applying this to objects with acquired purposes, IAMs will be unreliable insofar as they do not respond to their programming, since that is what they are meant to do. Two IAMs may therefore do exactly the same thing, with one being reliable and the other unreliable. If one is programmed to act as a paperweight and the other to design a new office building and they both just hold down stacks of paper, then only the former is being reliable.

For persons, unreliability will consist in their not performing their voluntarily acquired purposes – that is, in not fulfilling their commitments. We can state this as follows:

> A is an unreliable person if and only if A is a person and A *fails to fulfil their commitments*.

Now, to not fulfil a commitment entails that there is a commitment to be unfulfilled. As was mentioned in the context of rocks, there must be something that an object is meant to do in order to count as unreliable and for persons, this is what they have committed to doing. Following on from this, it is quite simple to give the account of untrustworthiness in the same style as the account of trustworthiness above.

> A is untrustworthy for B with respect to ϕ if and only if A has committed to B that they ϕ and A does not ϕ.

Equipped with this theory of untrustworthiness, we can more easily grasp the status of the uncommitted. On a binary view, they are either trustworthy because they have not broken a commitment, or untrustworthy because they have not kept a commitment. However, we can now see that neither of these apply, since both trustworthiness and untrustworthiness require that one has committed to performing the relevant action.

For this reason, we need the concept of non-trustworthiness, which simply refers to behaviour that is outside the scope of trustworthiness. This should not be particularly controversial; as pervasive as matters of trustworthiness are, it is quite intuitive that it is not relevant to all actions. If one has no commitment to ϕ, then one is not trustworthy by ϕing, nor untrustworthy by not ϕing, but merely non-trustworthy with respect to ϕ. Similarly, someone who makes no commitments whatsoever is non-trustworthy in a more general sense.

Katherine Hawley disagrees. On her view, everyone who is not trustworthy is untrustworthy; she does not have an equivalent of non-trustworthiness. Her view is also constituted negatively: one is trustworthy not by fulfilling commitments, but by avoiding unfulfilled commitments. It follows that those who are uncommitted are trustworthy, for if one has no commitments then one has no unfulfilled commitments. (Hawley, 2019, 74; 79)

The view that the uncommitted are thereby trustworthy does not seem very plausible to me. One important purpose of the concept of trustworthiness is that it helps us to know who to trust. A trustworthy person is, literally, worthy of our trust. It is thus similar to ideas of blameworthiness and praiseworthiness, which indicate who we should blame or praise. The trustworthiness of a person will not be the only factor in determining whether we trust them. If I am happy to do something for myself, then I may decide against trusting you to do it, however trustworthy you may be. Conversely, if I am desperate that something gets done and have nowhere else to turn, I might entrust it to a very untrustworthy person, or to someone of whose

trustworthiness I know nothing.[39] But that someone is usually trustworthy should be a reason in favour of trusting them.

On my view, this is explicable. If somebody always or usually fulfils their commitments in some domain and they have made a commitment to us in that domain, then we have good inductive evidence that they will fulfil their commitment to us. Their past trustworthiness thus counts in favour of our trusting them. On Hawley's view, however, this is not the case. Suppose that someone has made no commitments, at least of a given kind. They then make their first ever commitment of that kind to us. For instance, they promise to look after our children having never taken on childcare responsibilities before. Does the fact that they have made no commitments previously count in favour of trusting them? It seems not. It is true that they have kept every one of their commitments, but only vacuously so. It gives us no indication of whether we should trust them. If anything, it might make us more reluctant to do so. We might still trust them for other reasons, but the point here is that if being uncommitted counts as being trustworthy, then we must say that someone might be perfectly trustworthy, yet this trustworthiness does not count in favour of trusting them. This flies in the face of the role of the concept of trustworthiness in helping us to decide who to trust.

Hawley is well aware that her binary approach, which leaves no space between being trustworthy and being untrustworthy, can seem counterintuitive. She considers whether the term 'trustworthy' really is appropriate for the uncommitted:

> For example, a hermit who has no commitments whatsoever easily manages to avoid unfulfilled commitments; we might be reluctant to describe hermits either as trustworthy or as untrustworthy. Moreover inanimate objects and babies lack unfulfilled commitments, because they lack commitments. Yet we don't call them 'trustworthy'.

(Hawley, 2019, 78)

She goes on to argue that this is not a serious problem because almost everyone has commitments and so the issue will only rarely arise; that 'hermits and babies are not central to our thinking about trustworthiness and untrustworthiness.' (Hawley, 2019, 79).

However, I do not think that this is a good reason for setting the issue aside, for a number of reasons. Firstly, it does not address the argument made above, that classifying the uncommitted as either trustworthy or untrustworthy fails to respect the roles those concepts play in determining who we should trust.

Secondly, it does not deal well with the special case of non-committal speech. As discussed above, Hawley thinks that assertion involves an implicit commitment to tell the truth, and I agree. But if someone speaks without asserting, like the actor on stage, then it would be very odd to say that they are being trustworthy in their speech, on the grounds that, having made no assertion, they did not commit to tell the truth. Rather, we should say that their speech is not assessable in terms of trustworthiness. They are being non-trustworthy.

Thirdly, the fact that certain circumstances, such as someone lacking commitments, are highly unlikely does not mean that they are philosophically unimportant. Consider fake barns and

[39] Whether this counts as genuine trust is debatable. I mention it merely as a possible reason for trusting someone besides trustworthiness.

runaway trolleys.[40] Both are instrumental in important philosophical thought experiments, yet are not things that we are ever likely to come across. They are not usually 'central to our thinking' about knowledge and moral dilemmas in the everyday sense, yet do need to be considered when drawing philosophical conclusions. Similarly, the fact that uncommitted people are rare and do not normally play a role in our thinking about or experience of trustworthiness and untrustworthiness does not mean that we should disregard them from our theorising about the concepts.

Finally, it seems that being uncommitted is not nearly so rare as Hawley supposes. It is quite common for people to make a habit of trying not to commit to things in order to avoid the risk of disappointing anyone. This becomes especially clear when we recall that trustworthiness and untrustworthiness are relative to certain actions. We do not simply commit, but commit to some action. Everyone is uncommitted to something, so the question of whether we are trustworthy or untrustworthy with respect to the actions that we are not committed to applies to us all.

This does not mean that Hawley's theory is straightforwardly incorrect, since it may work well for considering a more general sense of trustworthiness. However, it does limit its usefulness. When it comes to someone's status with respect to a given action, I believe that my theory, allowing as it does for the concept of non-trustworthiness, does a better job of dealing with the uncommitted.

The Question of Competence

The next problem is the one exemplified by the doctor, the *mafioso,* and the cake-baking child: in two of the cases, the commitment remains unfulfilled because the task is just too difficult for the commitment-maker; in the other one, the commitment is wilfully broken. Many other examples could no doubt be given; the distinction between a failed attempt to do what is right and deliberate wrongdoing is a familiar one in ethics.

It must first be acknowledged that my view entails that, in all cases, the agent is not trustworthy. None of them fulfil their commitments. However, we can again make use of the notion of non-trustworthiness to distinguish them. This time, a more nuanced application is needed. The view that I shall defend here is that, while all who willingly break their commitments are untrustworthy in doing so, some of those who try and fail to keep them are untrustworthy and others are non-trustworthy. Specifically, those who know, or who should know, that they will be unlikely to succeed in fulfilling the commitment when they make it are untrustworthy and those who are innocently ignorant of the fact are non-trustworthy.

To begin, recall that a commitment is something that one is meant to do that is acquired voluntarily. The distinction I wish to draw rests on what can prevent a commitment from being voluntary and thus prevent it from being a genuine commitment at all.

Suppose that someone volunteers for a certain duty, but some crucial information about the task they are to perform is withheld. Perhaps it turns out to be more dangerous or to take much longer than they had initially been led to believe. To what extent have they really volunteered

[40] See Goldman (1976, 772-3) and Foot (1967), respectively.

in an ethically meaningful sense? There will no doubt be some grey areas – they may have made unjustified assumptions or the nature of the job might unavoidably change after they have signed up – but it seems that at least some such cases should not count as genuine volunteering. Volunteering is like consenting, in that it must be sufficiently well-informed to be morally valid. Again, how well-informed it must be is vague and may depend on context, but it is clearly a relevant factor. Of course, if they were fully informed but ignored the information, or had ample opportunity to check what was required, then this would not apply. What matters is whether they should have known better. This is also vague, but there are again clear cases on either side. Some are innocently ignorant and some are culpably ignorant. Where to draw the line may be difficult, but it is an important consideration in voluntary action.

Now, let us go back to the doctor giving financial advice. It is highly unlikely that they are innocently ignorant of their incompetence in this area. Maybe they do not realise that the task is beyond them and they might be well-meaning. But we can suppose that they are somewhat arrogant, perhaps having been lucky in previous investments and wrongly put this down to their own skill and good judgement. They should have known better. It is not rational, given their evidence, to believe that they can be a competent financial advisor. Given our view on commitments, their commitment to give financial advice is genuine. Since they are culpably ignorant of their lack of expertise, they made it voluntarily. This makes them untrustworthy by the definition given above; they committed to something and failed to do it.

In contrast, consider the young child promising to make their mother the most delicious cake ever for her birthday. This may be a well-intentioned promise, but the child has no chance of following through on it. Nevertheless, they do earnestly believe that they can do so and, given their age and lack of experience, they do not seem to be at fault for having that belief. Here, it is not the case that they should have known better; the child is innocently ignorant of their incompetence in the domain of cake-making. Therefore, the commitment is not truly voluntary and so is not genuine. Even given the explicit promise, they are incapable of making a binding commitment to make the most delicious cake ever. It is an interesting implication of the account that people do not always succeed in making the commitments that they think they are making.

If this is right, then those who are innocently ignorant of their incompetence in the relevant domain fall into the category of being uncommitted. As was argued in the previous subsection, this entails that they are non-trustworthy, rather than trustworthy or untrustworthy. However, those who either know or should know that they are incompetent in the relevant domain, but nevertheless make a commitment, are untrustworthy. In the special case of assertion, both those who intentionally lie and those who speak confidently on topics of which they know little are untrustworthy; the latter lack the competence to keep the implicit commitment to tell the truth.

Note that what makes the doctor untrustworthy is not primarily breaking the commitment, but making it in the first place. Insofar as ‘ought’ implies ‘can’, we may not say that those who cannot fulfil their commitments due to their lack of ability still ought to. What they did wrong, from the perspective of trustworthiness, is getting into that situation in the first place. This is a point that Hawley (2019, 77) also draws attention to: trustworthiness requires that we take care to not overcommit. We have a duty to consider the limits of our competence and what a commitment is likely to entail. An unforeseen obstacle is no excuse if we could have foreseen it, had we made the effort to look. In practice, untrustworthiness will often involve a little of both factors. For instance, people may commit to doing things that are very hard but not

impossible for them, then make a lukewarm effort and fail. They may reasonably be criticised both for not trying hard enough and for committing in the first place.

The *mafioso*, unsurprisingly, is also untrustworthy. In lying to business-owners, police, and others in the community, they break the commitment to truth-telling incurred by making an assertion. They are not trying and failing to keep a commitment, but deliberately breaking it, with nothing to defeat its voluntariness. Is it too harsh to put the *mafioso* and the doctor in the same category? It might seem so, but I think the difference is analogous to that of deliberate and negligent wrongdoing. Both are cases of wrongdoing, even if the former is morally worse, all else being equal. Similarly, it is plausibly worse to deliberately lie or make false promises than to speak in ignorance or make rash promises that cannot be kept. But both are cases of untrustworthiness.

In practice, innocent ignorance of one's incompetence is likely to be rare. If one has not bothered to consider whether one can really do something before committing to it, then one's ignorance is hardly innocent. This is especially clear when the stakes are high. Rarely are we in a position in which it is entirely rational to for us to believe that we can do something when we cannot. Nevertheless, such situations are possible and should be accounted for.

On the view presented above, a commitment made in innocent ignorance of one's incompetence to fulfil it is no commitment at all. In contrast, Hawley (2019, 79-82) argues that all those who fail to fulfil commitments due to a lack of competence are untrustworthy. This is another result of taking trustworthiness to be a binary matter. She says that 'trustworthiness requires us to undertake commitments only where we have the competence to fulfil those commitments' (2019, 82). If everyone is either trustworthy or untrustworthy, this entails that anyone who fails to meet a commitment is thereby untrustworthy.

It is counterintuitive that those who make an honest mistake about what they will be able to accomplish should be labelled as untrustworthy. Hawley acknowledges that some are inclined to disagree with her here:

> When I discuss these ideas with others, I find that some are reluctant to describe this sort of person as 'untrustworthy', preferring to reserve that term for those who are dishonest, insincere, or intentionally manipulative … Isn't incompetence in these respects just a matter of unreliability, rather than untrustworthiness?

(Hawley 2019, 81)

She goes on to justify her view, saying that 'untrustworthiness-through-incompetence can have much of the same ethical and practical character as does untrustworthiness-through-bad-intentions' (2019, 81) and that 'it should be clear that not knowing that one is over-committed will not serve as an all-purpose excuse for failure, nor as a mark of trustworthiness' (2019, 81). These claims are true, but they do not entail that failing to meet a commitment through incompetence is always untrustworthy. The actions of the doctor who should have known better than to commit to giving good financial advice may have similar 'ethical and practical character' to deliberately tricking a client, but that does not mean that those who are innocently ignorant of their lack of expertise falls into the same category. Being committed beyond one's ability to follow through is not 'an all-purpose excuse for failure', but innocent ignorance of one's limits can be an excuse.

Interestingly, Hawley adds, 'it is morally problematic to end up in a situation in which you are committed to doing something you are not competent to do, *absent a good excuse at least*.' (Hawley 2019, 81 [emphasis mine]). It is not clear what counts as a good excuse on her view, but it could be that Hawley has in mind cases like the child who is optimistic about their baking abilities. If so, we may have little disagreement on this matter after all. The central point I wish to make here is that my account of trustworthiness can accommodate cases of trying hard to meet commitments but failing.

A related issue worth briefly addressing is that of bad luck. Suppose that someone very skilled in the relevant area makes a commitment, but is faced with extraordinarily difficult unforeseen circumstances. They fail to fulfil their commitment, but it again seems too harsh to label them untrustworthy. In such cases, the agent in question is non-trustworthy, for similar reasons to those given above. They were innocently ignorant of the circumstances that would arise and therefore innocently ignorant of the fact that they would not be able to fulfil their commitment. We might say that, in those specific circumstances, they lacked the necessary competence, but justifiably (or excusably) thought that they possessed it. Their commitment therefore ceases to be binding on them. This view can thus encapsulate the idea that commitments often have implicit release conditions.[41] Again, though, we can imagine cases that point the other way. Often, although we cannot foresee a specific obstacle, we ought to realise that some obstacle or other is likely to arise and adjust our commitments accordingly. This is another vague area, but one that this account is able to deal with. In the case of seriously bad luck, the commitment ceases to hold because it no longer counts as voluntary. But we ought to plan for the possibility of things going at least mildly wrong. Where the bad luck is within normal parameters, the commitment holds and we are untrustworthy if we fail to fulfil it. This is another case in which the untrustworthiness is plausibly less bad than deliberate commitment-breaking; it is untrustworthiness by negligence.

It has been shown that trustworthiness as person-specific reliability can deal with these questions. Where the answers differ from those of Hawley's commitment-based account, they compare favourably. No doubt there are further questions and challenges that could be raised, but the fact that these ones can be dealt with lends support to the account I have proffered.

Conclusion

I have presented a theory of trustworthiness based on the idea that it is a kind of reliability that can be possessed only by persons. This person-specific reliability, it was found, entails a commitments-based theory of trustworthiness. As autonomous agents, we get to decide what others may rely on us for by choosing what we are meant to do for them – that is, by making commitments.

It was then shown how this theory deals with various challenges. It can do this more satisfactorily than at least one of the main commitments-based accounts currently in the literature. This chapter has therefore not only provided a theoretical underpinning to

[41] As pointed out by Marušić (2015, 11-2).

trustworthiness, constructing it out of the more basic concept of reliability, but gives an improvement on the view of trustworthiness as commitment-fulfilment.

At the outset of this chapter, we discussed four features commonly attributed to trustworthiness. It is a three-place predicate; its scope is variable, covering a spectrum from individual actions to general behaviour; it covers both speech and action; and most prominently, it is a kind of reliability specific to persons. The commitments-based account that I have presented satisfies each of these, so fits with plausible ideas already expressed in the literature, though with an original approach.

In the previous chapter, it was argued that trust is reliance on another's trustworthiness. Now that we have a theory of trustworthiness at our disposal, we can unpack this more fully. This is the account of trust that I favoured:

> A trusts B to ϕ just in case A relies on B to be trustworthy for A with respect to ϕ.

Given the theory of trustworthiness as person-specific reliability that has been proposed, we can unpack this as follows:

> A trusts B to ϕ just in case A relies on B to fulfil their commitment to A that they ϕ.

This is a relatively simple account of trust and the one that I shall be using in the remaining chapters – although, now that a theory has been established, I will often use the more condensed version, talking of trust as reliance on trustworthiness.

As one might expect, this theory has a strong resemblance to Hawley's (2014, 10): '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.' Like hers, it shows that it is inappropriate to trust someone to do something without a relevant commitment on their part. However, my version does not explicitly include the belief in that commitment. Rather, I leave it open whether such a belief is required; it may be that one can rely on another to fulfil a commitment that they merely *assume* has been made.

Now that we have a theory of trust, supplemented with a theory of trustworthiness, we will turn in the next two chapters to the norms of trust. I will first consider some of the accounts that have been put forward, based on evidence and practical reasons. I will then construct an account of my own, arguing that certain kinds of both evidential and practical reasons can justify trusting someone.

4

Evidence and the Norms of Trust (I)

Introduction

Trust is a necessary part of our ethical, practical and epistemic lives. There are many things that we could not know or do without depending on others. We simply do not have the resources, time or expertise to be entirely independent, either in our actions or our beliefs. We readily rely on others for information about our surroundings, future events and even our own personal histories (who could know, without ever being told, the date of their birth?). Likewise, we need other people to act in certain cooperative ways in order for our endeavours to be successful. If any joint projects are to work, we need others to turn up and do their part. To have valuable relationships with others – from intimate romance to profit-driven business partnerships – one must often find oneself depending on them.

But people are not always dependable. They sometimes lie, cheat, or otherwise deceive. Less maliciously, people can make mistakes, speak carelessly and forget. Lack of dependability can even arise from very good intentions, as when someone takes on more commitments than they can handle in an effort to please others.[42] It is therefore worth asking when it is reasonable to trust others. What factors are there that can rationally justify trust?

Of course, not all of the examples given above need be cases of trust. It is commonly accepted that there is a difference between trust and mere reliance[43] and we may often just rely on others to do things without any need for trust. Similarly, not every case of believing what someone says is a case of trust; sometimes we will have independent evidence for the proposition, so our belief has nothing to do with the other's assertion. Nevertheless, trust is an important sub-class of these instances of practical and epistemic dependence. At least sometimes, we will trust in cases like those sketched out above.

The purpose of this chapter is to mark out the kinds of reasons that should count when we consider whether to trust someone. The precise nature of trust, including the distinction between trust and reliance, is therefore not our main topic, though we will find it necessary to consider it in due course. For this reason, it will also be necessary to repeat certain points from previous chapters at various stages, which I hope the reader will forgive.

The two most common approaches to the question we are considering are what I shall call Evidentialism and Pragmatism. The Evidentialists believe that, when deliberating about whether and who to trust, the only reasons that count are those which have a bearing on whether

[42] This is a point emphasised by Hawley (2019, 76-7).
[43] See, for instance, Baier (1986, 234), Jones (1996, 9) and Hawley (2014, 1-2).

the trust will in fact be fulfilled. The Pragmatists hold that practical, as well as evidential, considerations bear on the question.[44] In what follows, we will consider arguments for each position. It will be found that both sides have useful insights, but none of the views considered can give us a fully satisfactory theory of the norms that govern trust. I will argue that a certain form of Pragmatism is capable of capturing the correct norms of trust and will spell out a more refined theory in the next chapter. For now, however, the task is to consider those views already proposed.

Aside from Pragmatism and Evidentialism, there is another view that has been suggested. This is the idea that trust does not involve justification in the traditional sense at all. Rather, the justificatory burden is shifted to the other person; they are, in a sense, a guarantor of the truth of what they say. If we are asked to justify our trust-based beliefs and behaviour, we are entitled to say, 'Don't ask me; ask them.' Such is the view advocated by Richard Moran (2005, 11) and Benjamin McMyler (2013, 1067) with respect to the specific case of trusting someone's testimony. Berislav Marušić (2015, 192-3) takes up this approach, applying it to trust generally. He claims that one can be justified in taking up a stance of trust, but that the justification is neither practical nor evidential (2015, 194; 196). Rather, there are specific 'reasons of trust', but it is not made clear what these might be (2015, 195).[45]

The general idea of such a view seems to be this: one must just be justified in taking up a certain attitude towards another person – an attitude of trust, or of epistemic buck-passing, to use McMyler's (2013, 1064) phrase. However, once this is done, no further evidence or practical reason is required to justify the trust. 'Why do you believe this?' is a question to be deferred to the one trusted; 'Why do you trust them?' still requires an answer from the addressee. I will not consider this type of view specifically, because I think that it can be subsumed under the categories that I have already laid out. Whatever the reasons one might have for taking up the attitude of trust, they must either indicate that the proposition in question is (likely to be) true, or count in favour of it in some way unconnected to the truth. That is, they must be either evidential or practical.[46] Therefore, the reasons of trust, whatever they may be, will be at least indirectly addressed within the framework that I have set out.

The chapter will proceed as follows. In the next section, I will describe some of the motivations for each of the two main views, as well as the problems that they must overcome. Then, in the following section, I will consider Evidentialism in more detail. This comes in two forms, which I shall call Simple Evidentialism and Restricted Evidentialism. We will find that neither succeeds in overcoming the problems with which they are faced. Next, I turn to Pragmatism. This, too, is split into two distinct versions: Cognitive Pragmatism and Non-Cognitive Pragmatism. Again, neither will be found to be adequate.

[44] A possible third position is taking practical reasons to be the only kind of consideration that counts, excluding evidential reasons altogether. I do not know of any philosophers who hold this view.

[45] Marušić (2015, 195) talks of taking a promise or assertion as 'an offer of an answer' to the question of whether they will perform the promised action or whether the asserted proposition is true. Accepting such an offer is trust. He contrasts this with merely taking the promise or assertion as 'support[ing] the conclusion' that they will do it or that it is true. What the difference amounts to I am unsure.

[46] My usage of the terms 'evidence' and 'practical reasons' thus makes them mutually exclusive and jointly exhaustive. Another way to describe them might be as 'truth-conducive' and 'non-truth-conducive' reasons, respectively.

Throughout, the convention of taking trust to be a three-place predicate will be followed: A trusts B to ϕ, where A and B are persons and ϕ is a behaviour. Assertion will be treated as a kind of behaviour. In the specific case of A trusting an assertion of B's, the behaviour is that of speaking (or signing, or signalling) truthfully.

With these preliminaries out of the way, let us now consider what motivates Evidentialism and Pragmatism, as well as the problems with which they are faced.

Evidence and Practical Reasons for Trust

Let us begin with practical reasons. In this category, I include any reason that does not bear on the truth of the matter. For our purposes, moral reasons and reasons of friendship count as practical reasons, as well as reasons given by some expected advantage associated with trusting somebody and of the supposed value of trust itself.

There are numerous such reasons that seem to support trusting another. We might wish to develop a good relationship with them, or maintain the relationship that we already have, for which we recognise that trust is an important factor. Or, we may consider trust to be valuable in itself and accordingly do our part in fostering it by trusting others more. Or again, we might see that trust can be of material benefit in the long term, since it helps with efficient cooperation; if my business partner and I trust one another, we will save time and make more money, not to mention be delivered from the emotional burden of worry.

Aside from the specific examples that could be mentioned, there are more general arguments in favour of thinking that at least some of the reasons of trust are practical. As Karen Jones (2012, 62-6) suggests, part of the purpose of the concepts of trust and trustworthiness is given by the fact that we can – and sometimes must – depend on others in various ways. In other words, we recruit others' agency to serve our own interests. This need not be coercive or manipulative – indeed, I will argue that such tactics are incompatible with trust – but cooperative. But if this is part of its purpose, then we should expect practical reasons to be among the reasons that justify trust; there is more point to trusting someone if doing so will further our interests to a greater extent. It has also been observed that trust tends to go beyond the evidence; when we trust, we will do so to a greater extent than the evidence warrants. Relatedly, trust displays a certain resistance to counterevidence. If a friend is accused of some crime and protests their innocence, we will likely believe them. As the evidence against them mounts up, we may well continue to believe them past the point of ordinary epistemic rationality – although this resistance is not limitless (Baker 1987, 1-4). Trust, moreover, is essentially *personal*, in contrast to the impersonal character of purely evidence-based attitudes; our trusting attitudes are directed at persons, not propositions (Marušić 2015, 180-2). These considerations indicate that there is usually more to our deliberations about trust than weighing the evidence. As Victoria McGeer summarises the situation:

> [T]rust may be characterized by two related features: (1) it involves making or maintaining judgements about others, or about what our behaviour should be towards them, that go beyond what the evidence supports; and (2) it renounces the very process of weighing whatever evidence there is in a cool, disengaged, and purportedly objective way. The problem, then, is to explain how trust of this sort can be rational.

(McGeer 2008, 240)

It could be, of course, that our usual trusting practices are irrational and ought to be revised. However, we should not dismiss out of hand the idea that our attitudes are at least somewhat rational without giving due consideration to the alternative.

These points motivate the idea that trust can be justified by at least some practical reasons as well as by evidence. However, there are those who take such a view to be mistaken. Pamela Hieronymi (2008, 232) and Berislav Marušić (2015, 184-5), for instance, take practical reasons to simply be reasons of the wrong kind for trust and therefore no reasons at all. They compare them to practical reasons for belief.[47] If a reward is offered for adopting a certain belief, that is the wrong kind of reason for believing. If the belief is formed for that reason, then it is not a rationally justified belief. This is because only reasons bearing on the truth of a belief can justify it, which practical reasons do not. Such reasons are concerned with the attitude itself – it is advantageous to hold the belief – rather than its content. Applying this to trust, only those reasons which bear on the truth of the question of whether the trusted party will perform – that is, the evidence – can justify trust. Suppose, for instance, one were offered a substantial reward for trusting someone. Could one trust for such a reason? More to the point, if one did trust, would the trust be rationally justified? Intuitively, it seems not; that would be a reason of the wrong kind. It may count as a reason to try to inculcate such an attitude in oneself, but it would not justify the attitude itself. Practical reasons of trust, such as those of friendship or those derived from the perceived value of trust, are seen as relevantly similar to the reason of reward: they do not address the content of the attitude, so cannot be genuine reasons at all. Hence Evidentialism: only the evidence counts.[48]

When we turn to the relationship between evidence and trust, however, we are confronted with a further problem. There seems to be an inherent tension in the idea that rational trust requires evidence, which can be spelled out as follows:

(1) If the evidence supporting trust is weak, then the trust is irrational.
(2) If the evidence supporting trust is strong, then the trust is misplaced.[49]

The first conditional is true by hypothesis; if trust requires evidence to be rational, then, for any instance of trust, there must be some threshold below which the evidence is too weak for the trust to be justified (this threshold may be variable, depending on context and whether there are also non-evidential reasons of trust). By 'misplaced' in (2), I mean that the justification is not conducive to trust specifically. When there is good evidence of something, we can simply form an ordinary justified belief, which is not distinctively *trusting* (Marušić 2015, 180). For instance, suppose that you tell me that p and I already have a lot of independent evidence that p. If I now believe that p, this belief seems not to be an instance of trust, since it need not have anything to do with your assertion. Or suppose that you promise me that you will ϕ and I have

[47] Both Hieronymi and Marušić take trust to be a kind of belief, so it is not surprising that they think this comparison appropriate. Evidentialism does seem to fit naturally with a belief-based view of trust, but I will not assume that all Evidentialists hold such a view, or that all those with a belief-based view are Evidentialists. As mentioned, although Marušić takes trust to be a kind of belief, he does not take it to be justified in the same way as ordinary belief, but by specific reasons of trust.

[48] However, on my use of the term, Evidentialism does not require that *all* the evidence counts, as we will see below.

[49] See, for instance, Marušić (2015, 178-83) and Moran (2005, 27).

strong evidence that you will ϕ because I know that you are in the habit of ϕing, or that it is greatly in your interests to ϕ. Again, this does not seem conducive of trust. If I believe that you will ϕ and rely on you to do so, then I might not be trusting you, since my reasons may have nothing to do with your promise. Requiring evidence for rational trust does not seem to fit with a genuinely trusting attitude.

I call this the 'problem of evidence'. If both (1) and (2) are true, then it seems that evidence can never justify trust. Either there will not be enough evidence, or the evidence will leave no room for a distinctively trusting attitude.[50] The problem of evidence looms largest for Evidentialists, but it must be dealt with by any view on which evidence is thought to justify trust. It must be explained how a large amount of evidence can leave room for a distinctively trusting attitude, rather than (or as well as) an ordinary belief. If it cannot be dealt with, then one must conclude that evidence is simply not the measure by which the rationality of trust is to be judged, if trust can be rational at all.

These are the main motivations of each view as I see them, as well as the main challenges that each must face. The Pragmatist must show why practical reasons are, after all, among the right kinds of reasons of trust. The Evidentialist must deal with the apparent validity of practical reasons, given the common features of trust mentioned above. Both must answer to the problem of evidence, though this does seem more threatening to the Evidentialist's position.

We turn now to examining each in more detail, starting with Evidentialism.

Part 1: Evidentialism

As discussed, the Evidentialist view is that only the evidence counts in favour of trust. Put more formally, A's trust in B to ϕ is rational only if A's evidence suggests that B will ϕ. This can be divided into two categories of views: those which take 'A's evidence' to refer to A's *total* evidence and those which take 'A's evidence' to here refer to a *strict subset* of A's total evidence. On the former view, A should consider all the evidence, whereas on the latter view, there certain kinds of evidence that are not relevant to trust. I will call these 'Simple Evidentialism' and 'Restricted Evidentialism' respectively and shall examine each in turn.

Simple Evidentialism

On Simple Evidentialism, all and only the evidence of performance supports trusting someone. This includes considerations of the kind mentioned above, when motivating the problem of evidence – reasons that seem to have nothing to do with another's testimony or promise and so

[50] I here assume that weak evidence and strong evidence are jointly exhaustive; there is no middle ground between them. If this seems counterintuitive, then read 'insufficient for rational belief (or reliance)' for 'weak' and 'sufficient for rational belief (or reliance)' for 'strong'. But even if there must be some middle ground, within which trust is justified, it would be very odd to say that the trust ceases to be justified because the weight of supporting reasons is too great.

do not appear conducive to trust. Nevertheless, it has been suggested that, at least sometimes, such reasons can justify trust.

An advocate of this view is Thomas Simpson. He motivates it with the following case, which he calls 'Antarctic Resupply':

> Lief is preparing to trek to the South Pole alone, pulling his food and equipment with him on a sledge. He is aiming to break the record for the fastest unsupported journey. The weight of his sledge would jeopardize the attempt if it had provisions for the return journey as well. So Katherine – an old Antarctic hand, who runs an adventure support company – promises that she will be contactable via satellite phone. When called, she will drop by parachute a package of supplies at the Pole, and they arrange a contingency plan if communications should fail. If Katherine does not follow the plan, it is all but certain that Lief will die. Lief is very keen to survive the expedition. He sets out south.

(Simpson 2017, 177)

Simpson takes it that, in 'Antarctic Resupply', Lief trusts Katherine. He also holds that Lief's trust is rational only if it is likely, on Lief's total evidence, that Katherine will drop the supplies at the South Pole upon request. (Simpson 2017, 177-8) Justifying this, Simpson invokes higher-order reasons. Since Lief's life is at stake, he has an *exclusionary reason* to not consider any non-evidential reasons. That is, he has strong reason to not factor reasons such as his relationship with Katherine and the value of trust into his deliberations when deciding whether to set out on his expedition (except insofar as they make it more likely that Katherine will drop the supplies at the right time). Thus, he considers *only* the evidence. (Simpson, 2017, 184-7)[51] Lief also has an *inclusionary reason* to consider his total evidence, not just a part of it. Given that he will probably die if Katherine fails to perform, he has strong reason to factor in all his evidence into his deliberations. He cannot afford to ignore factors which may affect whether he gets his supplies, even if they do not seem distinctively *trusting* reasons. Thus, he considers *all* the evidence. (Simpson, 2017, 187-9) Therefore, the high-stakes situation provides Lief with higher-order reason to *only* consider his *total* evidence.

Simpson advocates what he calls the 'Scope-Restricted Evidentialist Constraint' (SREC), which can be stated as follows:

> SREC: In some circumstances, A's trust in B to ϕ is rational only if, on A's total evidence, it is likely that B will ϕ. (Simpson, 2017, 183)

It is scope-restricted because Simpson does not believe that it applies in every case. It applies in 'Antarctic Resupply' because of the high stakes, but in scenarios with lower stakes, it may be rational to trust either for non-evidential reasons or for only some evidential reasons. Where actual performance on the part of the one trusted is not so crucial, the constraint can be relaxed. (Simpson, 2017, 185; 187) Simpson's position is therefore not an extreme version of Simple Evidentialism, on which it is always the case that trust is rational only if performance is likely on one's total evidence.

[51] The idea of exclusionary reasons is borrowed from Joseph Raz (1999, 38-40).

Nevertheless, since 'Antarctic Resupply' is supposed to be a genuine case of trust justified by all and only the evidence, Simpson's view is vulnerable to a number of criticisms. To start with, he seems to bite the bullet on the problem of evidence. Far from showing how evidence can be compatible with a distinctively trusting attitude, Simpson does not seem to consider this a problem at all. He takes it for granted that, since Lief trusts Katherine in setting off on his expedition and since Lief's setting off is rational only if the total evidence suggests that Katherine will perform, trust is compatible with consideration of all kinds of evidence. Simpson thus implicitly denies that evidence can preclude trust.

But this leads us to the next potential problem: does Lief really trust Katherine at all? Perhaps it is rational for him to consider all and only the evidence, but we might, if we are not already Simple Evidentialists, take that to indicate that, in this scenario, trust is just not rational after all. Any decent theory of the reasons that justify trust will acknowledge that trust is irrational under some circumstances and 'Antarctic Resupply' may be just such a set of circumstances. Here, we might think, Lief is justified in believing that Katherine will perform and is thereby also justified in proceeding with the expedition. But he should not *trust* Katherine – not because she is untrustworthy, but because the stakes are too high for an attitude that is not wholly dependent on total evidence.

But even if trust – as opposed to a firm belief – is justified here, would he still trust her in an amended scenario in which the kind of evidence he has is quite different? Consider the following:

> **Habit:** The same as 'Antarctic Resupply', but Lief has not communicated with Katherine. Katherine, however, has an odd habit of dropping bundles of various supplies at the South Pole at regular intervals. After ensuring that her next supply drop will contain what he needs, Lief carefully times his expedition to coincide with it.

In this version of events, it is rational for Lief to set out, since he can reasonably rely on Katherine to drop the supplies and therefore his survival is just as assured as if Katherine fulfils their agreement in the original version. It undoubtedly seems a high-risk strategy, but let us suppose that Lief is strongly justified in believing that he will arrive at the appropriate time. Does his reliance on Katherine to drop the supplies amount to trust? I do not think so. As noted earlier, it is widely acknowledged that there is a distinction between trust and mere reliance. A classic example is given by Annette Baier (1986, 235), in which Kant's neighbours rely on his famously regular walking habits to set their clocks. They are not trusting him to do so; his actions are not for their benefit, but merely something that they make use of. They are therefore not recruiting his agency in the way described by Karen Jones (2012, 65), in the way that is central to the notions of trust and trustworthiness. Similarly, in this amended version of 'Antarctic Resupply', it seems that Lief does not trust Katherine, but merely relies on her habit. He is not recruiting her agency, but only taking advantage of what he knows that she will do anyway. His total evidence does make it likely that she will make the drop; it does not support trusting her to do so.

Depending on the circumstances, acting according to all the evidence does not seem to involve trusting. Certain kinds of evidence appear to just not be very trusting reasons. A further example goes as follows:

> **Freerider:** The same as 'Antarctic Resupply', but Lief has not communicated with Katherine. Katherine has agreed to drop supplies at the South Pole for a team of researchers working in the area. However, it may take them some time to reach the spot after she has made the delivery. Knowing this, Lief times his expedition to coincide with the supply drop and help himself to what he needs before the researchers can get to it.

To avoid moral complication, let us assume that Katherine plans to drop more than enough to fulfil the researchers' immediate needs and Lief will take only what he requires, leaving the rest. He is therefore not endangering any lives. Once again, Lief has strong evidence that Katherine will drop the supplies, but he is not trusting her. Perhaps the researchers trust her, but Lief is not in an appropriate position to do so. He is taking advantage of what he knows she will do, following the evidence but not trusting her.

The point can be pushed even further. In the examples just given, part of the problem lies with Lief not communicating with Katherine and her reasons for action therefore being nothing to do with him. But some ways of influencing another's reasons for action are also incompatible with trust. Consider this scenario:

> **Coercion:** The same as 'Antarctic Resupply', but Lief has made a threat against Katherine's spouse. If Katherine does not drop the supplies when Lief needs them, then an associate of Lief's, with whom he will be in contact, will murder her spouse.[52]

Here, Lief's total evidence strongly supports Katherine's dropping the supplies. He knows that she is extremely unlikely to allow someone she loves to be killed. Yet he clearly is not trusting her. Force, manipulation, coercion, trickery – these are all ways to gain strong evidence that another will act in a certain way, but they are all incompatible with *trusting* them to act in that way. Therefore, the evidential reasons such tactics provide do not justify trust. For at least some circumstances, then, considering all the evidence simply precludes trust, even if it justifies belief and reliance. There does not seem to be a good answer here to the problem of evidence.

It may be objected that this does not affect Simpson's view, since SREC only applies to some circumstances. However, the circumstances that Simpson seems to have in mind are those with very high stakes – it is Lief's desire to survive that provides both the exclusionary and inclusionary reasons concerning consideration of the evidence (Simpson 2017, 185; 187). In each of the above examples, the stakes are still the same as in the original 'Antarctic Resupply'. The inclusionary reason to consider all the evidence therefore, on Simpson's view, still applies. Yet as the three amended cases demonstrate, no matter how high the stakes, there are certain pieces of evidence that do not support trusting someone, though they may support belief and reliance.

Note that it is only the inclusionary reason that is targeted in these examples, not the exclusionary reason. These criticisms are aimed at only part of the Simple Evidentialist view:

[52] Note that in this scenario, the threat has already been made. It is not a case of Lief making a threat in order to gain more evidence; it is a case of him having made a threat and considering the evidence that it gives him. Simpson (2017, 189-90) rightly distinguishes between 'following the evidence' by gathering more evidence (which he thinks does preclude trust) and 'following the evidence' by considering the evidence one already has. All the variations of 'Antarctic Resupply' fall into the latter category.

that rational trust entails taking into consideration all the evidence. Note also that I have not argued that there is no inclusionary reason of the kind Simpson suggests. Rather, the point is that, when it is rational to consider all the evidence and one does so, one's attitude is not one of trust.

As we have seen, certain kinds of evidence preclude trust. Simple Evidentialism has deeply counterintuitive consequences and fails to account for the distinctiveness of trust as opposed to ordinary belief. I have been arguing against Simpson's view as a particular example, but the points made apply to Simple Evidentialism in general. But perhaps a more modest Evidentialism could work. Even if trust is not justified by *all* one's evidence, perhaps it is still justified by *only* one's evidence. It is to such an idea that we turn next.

Restricted Evidentialism

Unlike Simple Evidentialists, Restricted Evidentialists do not take all kinds of evidence that someone will ϕ to support trusting them to ϕ. Only certain types of evidence count, so examples like 'Habit', 'Freerider' and 'Coercion' need not be considered genuine trust, which gives the restricted version an advantage over Simple Evidentialism.

For example, Pamela Hieronymi (2008, 221-4) argues that, while only evidence can count in favour of trust, it must be evidence of the right kind. This is because trust essentially involves treating the other *as a person*. We must take a participant stance towards them – that is, treat them as a participant of the world, acknowledging their agency, rather than treat them as just another feature of the world. This entails a readiness to feel betrayed, not merely disappointed, should we be let down.[53]

If we 'trust' for reasons of coercion or habit, then we are not treating them like a person because such reasons do not appropriately appeal to the other's agency. Instead, we are treating them like a complex object that might do things, or be made to do things, that we find useful. This is not trusting them at all, but merely making use of them. In 'Habit', for instance, Lief's reasons have nothing to do with Katherine's agency, but only with how she happens to behave. Similarly, in 'Freerider' and 'Coercion', although he is clearly relying on her acting in a certain way, he is not taking the participant stance towards her.[54]

Hieronymi goes further. Even if we take their word to be good evidence either that the spoken proposition is true or that they will do as they have said and believe on that basis, we may not really be trusting them. We are still not treating them like a person, but rather as a good thermometer – very reliable as a source of information, but not something with agency of its own (Hieronymi 2008, 222). If we treat a person's testimony merely as evidence, like any other kind of evidence, then our attitude is too impersonal to count as trust. We are not adopting the participant stance towards them.

[53] Hieronymi borrows the idea of applying the participant stance to trust from Holton (1994), who in turn adapts it from Strawson (1974, 9-10).

[54] Annette Baier (1986, 234) suggests, rightly in my view, that fear is an imputed motive incompatible with trust; if we are depending on another to do something because they are afraid of what might happen otherwise, we are not trusting them. Taking up this point, Holton (1994, 66) argues that fear is not sufficiently self-generated to count as a motive that can be relied upon from the participant stance.

But if we cannot rationally trust for practical reasons and treating someone merely as a source of evidence precludes trusting them, as Hieronymi argues, what does justify trust? She answers that trust is justified by the other's perceived *trustworthiness* (2008, 224).[55] The kind of evidence that supports trusting someone is evidence that suggests that they are trustworthy, which is a *personal* form of evidence. This approach – taking the participant stance, combined with justifying trust only with evidence of trustworthiness – precludes the kinds of reasons Lief has in the adapted versions of 'Antarctic Resupply' above. It therefore respects the intuitive nature of trust as not being wholly based in objective, impersonal evidence, while also providing criteria for its being rationally justified.

As Simpson notes (2017, 187-8), Hieronymi's view also implies that Lief does not trust Katherine in the original case.[56] He takes this to be a disadvantage of views like hers, that they cannot account for what he sees as a clear example of trust (2017, 188-9). But as was demonstrated with the alternate versions of the case, an inclusionary reason to take into account *any and all kinds* of available evidence is not compatible with genuine trust. The idea, mentioned earlier, that in such circumstances Lief cannot rationally trust Katherine should not be dismissed. Hieronymi or her advocate may argue that, even with a very large amount of evidence of Katherine's trustworthiness, the dangerous situation means that Lief ought not trust her, but consider all available evidence. The inclusionary reason to take all the evidence into account still applies, but rather than implying that one should consider all the evidence when determining whether to trust, it entails that trust is not reasonable in those extreme circumstances.

How should we assess Hieronymi's view of rational trust as being justified by evidence of trustworthiness? I believe that it compares favourably to Simple Evidentialism in a number of respects. First, it would explain why trust can be rationally resistant to ordinary counterevidence. Take the example mentioned earlier of the friend accused of a crime. If there is a lot of evidence that your friend is guilty, but they insist on their innocence, you can draw on your knowledge of their trustworthiness. If you have a substantial amount of evidence that they are trustworthy in these kinds of matters, then the overall balance of your evidence might weigh in favour of their being innocent. The balance will of course tip at some point, but the claim was not that trust is completely immune to evidence, only that it will be more resistant than ordinary beliefs. Similarly, the personal nature of trust is explained. If I know that you are trustworthy, at least in your dealings with me, then, on Hieronymi's view, I can rationally trust you. My evidence comes from a feature of you as a person – specifically, your trustworthiness.

Furthermore, it puts forward a plausible solution to the problem of evidence. There is never too much evidence for trust, so long as it is evidence of the right kind – evidence of trustworthiness. Hieronymi (2008, 226) suggests that we can take the other's practical reasons for being trustworthy as our evidential reasons for trusting them. Our evidence is therefore based on their agency, fitting well with the idea that trust involves recruiting others' agency. We might also consider our past experience with them or their reputation as providing evidence of

[55] Hieronymi gives a brief account of trustworthiness: being responsible for what one says, reliable in one's judgement and good in one's will (2008, 224). The precise nature of trustworthiness should not matter for our purposes here, however. A pre-theoretic understanding of the concept should suffice.

[56] Simpson's terminology differs from mine. He thinks of Hieronymi as a 'cognitive non-evidentialist' (2017, 188), meaning that she takes trust to be a kind of belief, but one that is not justified by one's total evidence (2017, 179-80).

trustworthiness. On the other hand, knowing their habits or resorting to coercion do not count in favour of trust; they may be evidence that they will act in a certain way, even based on their practical reasons for doing so, but they are clearly not evidence of trustworthiness. Thus, this view can distinguish trust from ordinary belief by restricting the relevant kinds of evidence.

But Hieronymi's Restricted Evidentialism is not free from problems. Although it can explain the resistance of trust to evidence, without claiming that such trust is irrational, it does not seem able to explain the related point of trust going *beyond* the evidence; of trust-based convictions being sometimes firmer than one's total evidence, including evidence of trustworthiness, warrants. It must also be explained why practical reasons do not count, when these often seem to be an important element in our deliberations about whether to trust. This, I believe, is the more significant problem and we turn next to the response that is offered to it.

Hieronymi is aware of this issue and addresses it with reference to an example of a drama class trust exercise borrowed from Richard Holton:

> You are blindfolded. You stand in the middle of a circle formed by the others. They turn you round till you lose you[r] bearings. And then, with your arms by your sides and your legs straight, you let yourself fall. You let yourself fall because the others will catch you. Or at least that is what they told you they would do. You do not know that they will. You let yourself fall because you trust them to catch you.

(Holton 1994, 63)

Holton believes that, in such cases, it is rational to trust the others for reasons that have nothing to do with whether they will in fact catch you. You may trust because you think that it is important to have an atmosphere of trust in a drama class; you may value the relationships that would be damaged if you did not trust your classmates; you may even just be trying to get through the course of which this exercise is a required part (Holton 1994, 69). Hieronymi, although she disagrees with Holton about practical reasons, recognises the need to offer an explanation of why they can seem like genuine reasons to trust. Her response is that you can indeed fall backwards for the kinds of reasons Holton mentions, without believing that you will be caught, but denies that doing so would be *full-fledged* trust. Rather, you merely *entrust* yourself to them. (Hieronymi 2008, 216-7) In other words, one can certainly act as if one trusts for practical reasons and such actions can be rational, but real trust requires evidential reasons in order to be justified.

Hieronymi goes on to argue that practical reasons are of the wrong kind, since they are concerned with the attitude and not with its content. What matters in the reasons Holton suggests is the trust itself, not the fulfilment of the trust. The importance of an atmosphere of trust, the value of your relationships with your classmates, your desire to get through the course – these are all concerned with trust itself, but do not address one's actually getting caught. They tell in favour of falling backwards, but have nothing to say about what happens after that. To make her point, she draws on similar considerations about ordinary beliefs. If some factor makes having a belief useful, convenient or important, then that does not itself justify the belief. What justifies the belief is evidence: reasons which bear on the truth of the belief's content. If you are offered a reward for forming a belief, for instance, then you are not justified in forming that belief. (Hieronymi 2008, 234) Similarly, trust cannot be justified by reasons that make trusting convenient or valuable, like those Holton proposes.

The common terms for the reasons that favour having an attitude and those which justify it by addressing its content are *state-given* reasons and *object-given* reasons, respectively.[57] Hieronymi (2008, 232) contends that, since practical reasons for trust are state-given rather than object-given, they are not reasons for trust at all. They can only be reasons for acting as if one trusts.

There is a problem with this line of argument, however. For some attitudes, practical reasons are object-given, so are genuine reasons. Take intentions, for instance. Gregory Kavka (1983) proposes a famous thought-experiment in which someone is offered a substantial monetary reward if they form the intention to drink a certain toxin, one which will make them feel unwell for a short time, but which will have no long-term effects. Such a person has a state-given reason to have the relevant intention. Kavka asks whether this counts as a reason to intend to drink the toxin, given that they do not have to drink it in order to get the reward. We need not answer that question; what is relevant here is the distinction between state-given and object-given reasons. We can easily imagine an object-given practical reason that would justify the intention: instead of being offered a reward to intend to drink the toxin, one is offered the reward to actually drink it. There will be no lasting damage and it is probably worth the temporary discomfort in order to gain the reward. Thus, one has an object-given practical reason to intend to drink the toxin. It is practical, not evidential, since it is concerned with gaining something rather than with truth. It is object-given, since it is concerned with the content of the attitude – the action – not with the attitude itself – the intention.

The question for us is this: in its justifying reasons, is trust more like a belief or an intention? Hieronymi, as we have seen, argues for the former, taking practical reasons to be state-given and therefore of the wrong kind.[58] But practical reasons cannot be ruled out in this way, since they can be object-given, at least in the case of intentions. If one of my reasons for trusting someone to do something is the fact that their doing it would benefit me, then this would seem to be an object-given reason for trust (whether it is still a good reason to trust is another matter). As a simple example, consider the Trust Game, a game-theoretic scenario sometimes used in psychology experiments.[59] Player 1 must decide whether to give a sum of money to Player 2. Their doing so can be plausibly construed as an act of trust. What Player 2 receives is multiplied by four. Player 2 then decides whether to share that money with Player 1. If Player 1 decides to send money to Player 2, one of their reasons is likely to be the fact that doing so will potentially lead to financial benefit, if Player 2 is cooperative. This reason is concerned with the desired outcome of the trust, not with the trust itself, so is an object-given reason. Hieronymi does not acknowledge object-given practical reasons, so her argument does nothing to exclude them. I believe that she is successful, however, in showing that state-given practical reasons are not reasons of the right kind.

I have been discussing the version of Restricted Evidentialism according to which the appropriate evidence is restricted to that of trustworthiness. There are of course other ways of restricting the right kind of evidence. Annette Baier (1986, 234), for instance, takes trust to be reliance on another's goodwill. On such a view, we might take the appropriate evidence to be

[57] See, for instance, Parfit (2001, 21-2), Way (2012) and Schroeder (2010). Hieronymi does not use the same terminology, calling state-given reasons *self-referential* reasons (2008, 232), or *extrinsic* reasons (2005, 448).
[58] Indeed, she takes trust to be a type of belief (2008, 228).
[59] There are a number of versions of this. See, for instance, Faulkner (2017, 110-1) and Kosfeld *et al* (2005).

that which shows the other to bear us goodwill and that they are likely to act on it.[60] But any such view would face the same challenge as Hieronymi's: how can practical reasons be excluded in a way that explains their apparent appeal?

The evidence-of-trustworthiness view is the version of Restricted Evidentialism that I take to be most plausible. Trust and trustworthiness are complementary concepts, so a theory of what makes trust rational ought to refer to trustworthiness and Hieronymi's does so in a simple and intuitive manner. Moreover, I believe that she is very close to being correct, as we shall see in the next chapter. In particular, the distinction between state-given and object-given reasons will be of crucial importance. Nevertheless, since it can neither account for nor explain away all the practical reasons that seem to favour trust, it ought to be rejected.

Having considered Simple and Restricted Evidentialism, we have found that neither kind works satisfactorily. Evidentialism cannot reasonably preclude all practical reasons while maintaining the distinctiveness of trust from ordinary belief. We turn next to the broad class of views that permits practical reasons among the reasons of trust.

Part 2: Pragmatism

The draw of Pragmatism is that it is rather more permissive than Evidentialism. Since it holds that both evidence and practical reasons can count in favour of trust, it avoids the problems that plague Evidentialism, while also acknowledging the key role played by evidence in rational trust. It should be noted that Pragmatism, as it is understood here, is compatible with the view that, under some circumstances, only evidence counts[61] – although we will not examine those circumstances until the next chapter. Pragmatism is merely the denial of the Evidentialist claim that practical reasons never justify trust.

As in the previous section, I divide this view into two sub-categories, which I call Cognitive Pragmatism and Non-Cognitive Pragmatism. Cognitive Pragmatists take trust to be a kind of belief, which can nonetheless be justified by practical reasons as well as by evidence. Non-Cognitive Pragmatists, while they may agree that trust is often found in the company of belief, deny that it is the same attitude. Trust is taken to be something else, most commonly a special kind of reliance. We will consider each in turn.

Cognitive Pragmatism

[60] This is an example based on Baier's view, but I do not mean to imply that Baier herself is an evidentialist about trust.

[61] Thus, Simpson's SREC is technically compatible with Pragmatism, but as argued, I do not think that he is correct that very high stakes are such circumstances.

It is quite common for philosophers to take trust to be a kind of belief.[62] However, it may seem odd to think that a belief can be justified by practical reasons. It is more usual to think of belief as being justified only by evidence; as Hieronymi argues, beliefs taken on because they are good, convenient or valuable are not well justified.[63] A belief, on the more common view, is rational insofar as the weight of reasons support its truth, which is not the domain of practical reasons.

Nevertheless, the Cognitive Pragmatist is impressed both by arguments which suggest that genuine trust must come with belief and by the apparent involvement of moral and practical reasons in decisions to trust. Accordingly, they take the view that trust is a special kind of belief that is rationally responsive to practical, as well as evidential, reasons.[64] A common candidate source of reasons for belief is friendship. Simon Keller (2004), for instance, takes the view that the fact that someone is our friend is a reason to believe that they will be successful in their endeavours. If a friend wants our opinion on how competent they are, we ought to be more confident in them than the evidence warrants. In a similar vein, Sarah Stroud (2006) argues that we should be epistemically partial towards our friends, forming and maintaining beliefs that portray them in a positive light. We have already encountered Judith Baker's (1987, 1-4) view that we ought to believe what a friend tells us and that these beliefs ought to have some resistance to counterevidence.

Believing someone because they are a friend or we have some other valuable relationship with them clearly allows for some of the intuitive practical reasons of trust. We will trust in order to develop and maintain those relationships and to cooperate effectively with certain others. The value of trust, as well as its being personal, is accounted for on this view. Furthermore, the tendency of trust to go beyond the evidence and be resistant to counterevidence is explained by the purported fact that evidence is not the only kind of relevant consideration.

An interesting feature of reasons connected to friendship in particular is that they do not always seem to be the sorts of reasons Hieronymi was criticising when she argued that practical reasons are the wrong kind for beliefs. She considers belief for the sake of the belief itself – it is useful or important to have that belief, regardless of whether it is true. This seems to be the case for Stroud's epistemic partiality; there is something valuable about having positive beliefs about someone if they are our friend and this value has little to do with those beliefs being accurate. However, for certain reasons of friendship, it often does matter whether the belief turns out to be true. In Baker's example, it would harm the relationship if, after steadfastly believing one's friend, it turned out that they were lying all along. The reason of friendship would thus be deprived of its force if the belief were false. In Keller's case, likewise, it matters that the other succeeds in what they are trying to do. Perhaps it would not damage the relationship if they failed, but part of the reason for believing it is wanting it to be true. It is not like a case of receiving a reward for adopting a belief, wherein the reward would carry just as much weight as a reason if the belief turned out to be false; these reasons, although they are not evidence, are in a sense concerned with the content of the belief being true. In the examples given by Keller and Baker, the agents are not indifferent to truth. This being the case, these are not purely state-given reasons. They are about the content of the belief as well as the belief itself.

[62] See, for instance, Marušić (2015, 205-8) and Baker (1987).

[63] Others who have argued this include Shah (2006) and Berker (2018). Some exceptions, who take belief to be justified by practical reasons, include Reisner (2009; 2018) and Rinard (2019).

[64] For more focused discussion on whether trust is a kind of belief, see Chapter 1.

Therefore, reasons for belief based on one's relationship cannot be entirely ruled out on the basis of their being state-given.

This kind of view can also deal with the problem of evidence. As was mentioned above, it is not just Evidentialism that must answer to this issue, but any view on which evidence can play a role in justifying trust. Even if one trusts another in part for practical and moral reasons, as the Pragmatist suggests, there is still the question of whether also having supporting evidence can preclude rational trust in favour of ordinary rational belief. In a similar way to Hieronymi, Cognitive Pragmatists can restrict the kinds of evidence that permit of trust. Evidence that is based on a good relationship with the other person is what matters, which will likely include evidence of their trustworthiness. Relationship-based evidence counts in favour of trust, while other kinds of evidence will count in favour of ordinary belief.[65] For example, in Baker's case, the thought, 'They are my friend and are therefore likely to be honest with me' will count in favour of trusting them in their claim to be innocent. This will also involve believing them to be innocent, but it would be a specifically trusting belief. On the other hand, independent evidence coming to light, such as another person's fingerprints at the crime scene, would count in favour of believing one's friend to be innocent, but there would be nothing distinctively trusting about that belief. Thus, the Cognitive Pragmatist is capable of maintaining the distinctiveness of trust, while also permitting some kinds of evidence to count.

However, there is still a significant problem with this view, which stems from the fact that it takes trust to be a kind of belief. This is that practical reasons, even if they are object-given, are the wrong kind of reasons for believing something. Take the following: 'It would be very beneficial if p were true.' This is not a state-given reason for believing that p; it says nothing of whether it would be beneficial to hold the belief. It is an object-given reason, since it is about the content of the belief. Yet it also does nothing to justify the belief; it could not settle the question of whether p.[66] It might be a reason to intend to make it the case that p, but it is not the kind of reason to believe that p.[67] Insofar as belief is an attitude that aims at truth, the Cognitive Pragmatist cannot answer this objection.

This is a problem with the cognitive, rather than the pragmatist, aspect of the view. We turn next to Non-Cognitive Pragmatism, which of course will not be vulnerable to the same objection.

Non-Cognitive Pragmatism

If we deny that trust must always be some kind of belief and accept that it can be justified by both practical and evidential reasons, then we have the least restrictive view yet considered: Non-Cognitive Pragmatism. Because it is so unrestrictive, it can easily avoid at least some of the problems that plagued the previous views. It is not committed to consideration of only the evidence, so evades the objections to Evidentialism. One who takes this view also need not be

[65] This is one way of restricting the permitted evidence, but a Cognitive Pragmatist could take other approaches, if they think that one's relationship to the other party is not the (only) kind of evidence conducive to trust.
[66] See Hieronymi (2005, 447).
[67] Likewise, if one has strong evidence that one will ϕ, this is not the right kind of reason for intending to ϕ, but a reason to believe that one will ϕ.

concerned about practical reasons being of the wrong kind for belief, since they are not committed to the view that trust is a kind of belief – although belief may often accompany trust.

Perhaps the simplest version of Non-Cognitive Pragmatism is that given by James Coleman (1990, 99), who essentially takes trust to be a kind of bet. When A trusts B to ϕ, A is betting that B will ϕ. So, whether one ought to trust is a function of the potential gains, the potential losses, and the likelihood that the one trusted will perform. Trust is rationally justified insofar as the expected outcome is positive. Specifically, where the value of A's gains if B ϕs is G, the value of A's losses if B does not ϕ is L and the probability of B's ϕing is p, the calculation is as follows:

$$\text{A ought to trust B to } \phi \text{ just in case } \frac{p}{1-p} > \frac{L}{G}.$$

This entails that the reasons that count in favour of trusting would be anything that increases the value of G, decreases the value of L, or increases the value of p. The first two would be practical reasons, the third evidential reasons. For example, it may be worth trusting someone, even if it is highly unlikely that they will perform, if one will lose very little from their non-performance but gain very much from their performance.

However, this view on the reasons of trust falls into a familiar problem: there is nothing distinctively trusting about it. As with Simple Evidentialism, there is no restriction on which kinds of evidence can be taken into consideration. Coleman's view is therefore faced with the problem of certain kinds of evidence which are not conducive to trust – those based on someone's habits, on what they are going to do for someone else, or on coercion – not being discounted for rational trust. The same can be said of the kinds of practical reasons that are taken into consideration. They need have nothing to do with friendship, or the value of trust, or with a desire to cooperate; they need only be considerations that make certain behaviours by another beneficial. It is all about maximising expected gains and minimising expected losses; there is nothing distinctively trusting about such a bet.

This is not to say that it is not a rational way of approaching certain situations. It may be reasonable, at times, to 'bet' on another's behaviour, in the sense of making it the case that one stands to gain if they behave in one way and to lose if they behave in another. When this is the case, the above calculation will be appropriate. The point here is that if one does make such a bet, one is not trusting them, but is merely trying to make use of them. One does not recruit their agency or incur vulnerability to betrayal in the way necessary for genuine trust.

Perhaps the central problem with this view is that it does not involve seeing the one trusted as a person, but as just a feature of the world with which one can interact in various ways to try to further one's own interests. It is widely accepted that trust requires in some sense acknowledging the other's personhood or agency.[68] So, to make it distinctively trusting, perhaps the reasons that justify trust must somehow involve the fact that it is an attitude one has to another person, rather than to an object. A more general lesson to be learned from the problems with Coleman's view is that, while certain problems can be avoided by loosening the

[68] Holton (1994, 65-6), for instance, argues that trust involves a participant stance, as we will discuss below; Jones (2017, 99) suggests that it involves imputing a certain motive. Hieronymi (2008, 215-6) and Bennett (2021, 514) express similar ideas.

restrictions on the reasons of trust, we cannot afford to be too liberal, lest we lose sight of the notion of trust altogether.

Let us now turn to a more nuanced version of Non-Cognitive Pragmatism. Richard Holton (1994) presents the idea that trust is reliance from the participant stance.[69] We have already encountered the participant stance in Hieronymi's view,[70] but I shall briefly recapitulate it.

To take the participant stance towards someone is to view them essentially as a person, rather than a mere object.[71] It is to see them as having various aims, motives and desires of their own, as a *participant* in the world, not just a complex feature of it that needs to be managed, made use of, or worked around. In taking such a stance, one adopts a readiness to experience certain reactive attitudes. Depending on what the other person does, we might feel, for instance, grateful or resentful towards them. These would be inappropriate attitudes for features of the world such as the weather. We may feel glad when the weather is good or disappointed when it is bad; we may work the weather into our plans or find ways to avoid being dependent on it. But we should not feel towards it as we might feel towards a person who either helps or hinders us. (Strawson 1974, 7-11)

As has been mentioned, not only in this chapter but previously in this thesis, it is commonly supposed that trusting involves acknowledging the other person's agency; it is appropriate to trust persons, but not objects. It is therefore an intuitive thought that trust involves taking the participant stance. Specifically, according to Holton, trust involves a readiness to feel betrayed and not merely disappointed by another's non-performance. Betrayal is the reactive attitude appropriate to trust.

Trust as reliance from the participant stance avoids the problem faced by Coleman's view of trust as betting on another's behaviour, since it very explicitly offers scope for the distinctiveness of trust. The reasons which favour ordinary reliance will not all be reasons which favour trust. Not all reasons for relying on something will also be reasons for adopting the participant stance. Relatedly, it clearly involves recruitment of another's agency and allows for the possibility of betrayal, rather than mere disappointment. These give it advantages over Coleman's version.

However, I do not think that Holton's Non-Cognitive Pragmatism works completely. It falls foul of a problem with which we are by now familiar: it appeals to the wrong kind of reasons. As we saw earlier, Holton justifies his view with reference to a drama class trust exercise. He thinks that, in that exercise, one can fall for all sorts of reasons: in order to promote an atmosphere of trust in the class; for the sake of one's relationships with one's classmates; or just to get through the drama course. What is more, one's falling can count as trusting if done for any of these reasons. So long as one takes the participant stance in relying on the others to catch one, one trusts them. That is to say, so long as one is ready to feel a sense of betrayal if dropped, rather than, say, a 'grim confirmation of [one's classmates'] alien ways' (Holton 1994, 69), one is genuinely trusting. Such a readiness can presumably be supported by the kinds of practical reasons Holton mentions. In order to promote one's relationships with the others

[69] Holton's view is discussed at greater length in Chapter 1.

[70] It is from Holton that she gets the idea that it may play a significant role in an account of trust.

[71] Strawson calls it the participant *attitude* (1974, 9-11), but I follow Holton and Hieronymi in calling it the participant *stance*, in order to more firmly distinguish it from the reactive attitudes. The participant stance is not one of the reactive attitudes, but the stance in which those attitudes are based.

taking the class, one will see them as agents rather than as objects and react to them accordingly. If one places great value on trust, especially in the context of a drama class, this might not only motivate one to fall, but inculcate a readiness to feel betrayed by the others if one hits the ground.

But these are state-given reasons. We have already seen that not all practical reasons are state-given, so not all are ruled out by Hieronymi's point that state-given reasons are of the wrong kind. In this case, however, Holton presents us with reasons for having the attitude of trust that do not address the content of trust. The supposed value of trust or relationships with other drama students are not reasons directly concerned with one's being caught or dropped. Whether one is actually caught or dropped does not matter for whether one has such reasons. Therefore, they are of the wrong kind.

Like Hieronymi's view, I think that Holton's comes close to being correct. However, his view cannot be accepted, since it embraces state-given reasons for trust. His idea that trust is a special kind of reliance and that it can therefore be justified by at least some practical, as well as evidential, reasons, is, however, correct. I shall make use of it below.

We have seen that neither Cognitive nor Non-Cognitive Pragmatism is fully satisfactory in the forms in which they have been considered. If Pragmatism is to be made plausible, therefore, we need a more sophisticated version. What we need is twofold: a way of non-arbitrarily allowing for some kinds of evidence but not others, to maintain the distinctiveness of trust; a way of non-arbitrarily allowing for some kinds of practical reason but not others, to avoid the wrong kind of reasons problem.

Conclusion

In this chapter, we have explored the two approaches that can be taken in determining the kinds of reasons that justify trust. Neither version of Evidentialism is satisfactory. On the unrestricted variety, reasons are permitted that are not distinctively trusting. The restricted version, meanwhile, fails to rule out object-given practical reasons. The prospects for Pragmatism look little better. If trust is a kind of belief, then practical reasons are of the wrong kind. But even if it is not, Pragmatism seems to allow for state-given reasons to count, which are no reasons at all.

However, I believe that there is a way around these problems. What is required is a way of permitting practical reasons while excluding state-given reasons and permitting evidence so long as it is of a distinctively trusting character. As we have seen, some practical reasons are object-given, as in the case of intentions. Some kinds of evidence are distinctive of trust, as in Hieronymi's restriction to evidence of trustworthiness. In the following chapter, I will argue for a version of Non-Cognitive Pragmatism which encapsulates these ideas.

5

Evidence and the Norms of Trust (II)

Introduction

In the previous chapter, we examined some of the theories of the justifying reasons of trust that have been proposed. Although none were entirely satisfactory, we gained some insight into what the main features of the norms of trust should be. In this chapter, I will propose my own theory, drawing on what we have learned, which will be shown to fit well with the views of trust and trustworthiness proposed in the first three chapters of this thesis. What I propose will also be a reasoned compromise between the main ideas of Evidentialism and Pragmatism, though it will strictly fall into the category of Non-Cognitive Pragmatism, which was discussed at the end of the last chapter.

As has been shown in Chapter 1, trust is a kind of reliance, rather than belief. We should therefore expect it to have the same norms as reliance, but applied in a specific way. These include both practical and evidential reasons; when deciding whether to rely, one considers the practical reasons for using that thing as well as the evidence that it will work. If someone relies on their car to take them somewhere, it is both the evidence that the car is sufficiently reliable and the fact that they need to get to a certain place that makes their reliance rational. Our purpose is to consider how these norms may be applied to trust, rather than reliance in general.

I begin by focusing on a different way of categorising reasons. Rather than evidential and practical reasons, I will be primarily concerned with state-given and object-given reasons. I will then reiterate the simple view of trust that was proposed in Chapter 2: trust is reliance on another to be trustworthy. Applying the norms of reliance, in view of the state-given/object-given distinction, to this view will provide an intuitive theory of the reasons that justify trust, which includes some practical and some evidential reasons, while also excluding some of each type. I believe that this theory will thereby avoid the counterintuitive consequences encountered by Evidentialism and Pragmatism. I will then show how this theory can deal with three potential problems, before concluding with a demonstration of how it fits with my own account of trustworthiness.

One advantage of the account I propose is its simplicity. The difference between trust and mere reliance, I argued in Chapter 2, is just that of content. Trusting just is relying, where the content of that reliance is another's trustworthiness. The theory given in the present chapter simply takes the norms of reliance and applies them to a specific content: trustworthiness.

Part 1: What Justifies Trust

In the first part of this chapter, I will present in some detail the theory of the norms of trust that I favour. To do so, it will be necessary to consider the kind of attitude that trust is and its content; these will tell us which reasons justify it. As we saw in the last chapter, one problem that other theories of trust face is that of the wrong kind of reasons. This I am keen to avoid, so shall begin by discussing that issue.

The Wrong Kinds of Reasons

There are, as we argued in Chapter 4, at least two ways in which a reason can be of the wrong kind. The first is that it is state-given, rather than object-given.[72] That is, it is a reason that shows there to be something good or desirable about having a particular attitude and does not address the content of that attitude. For instance, being offered a reward for believing something or intending to do something does not justify forming the attitude, even assuming that one can. To justify a belief, one needs evidence of its content; to justify an intention, one needs reason for thinking its content to be good or desirable. Likewise, considerations that show trust to be useful, valuable, convenient, or in some other way good do not justify it; one must have a reason that addresses the content of trust.

It is worth noting that the wrong kind of reasons problem arises only in the context of attitudes that have certain standards of correctness. That is to say, attitudes which may make one criticisable on grounds of irrationality if one does not follow the appropriate norms. For example, if one intends to do something that will be harmful, or believes something that is clearly not true, then one's intention or belief will be irrational. But if one imagines that something is the case, then one cannot be criticised on similar grounds. Maybe certain imaginings are not helpful or appropriate, but they are not irrational. There is no standard of correctness on them, in the way that there is for other attitudes. State-given reasons for imagining are therefore not of the wrong kind; there is no failure of rationality in imagining something because one thinks it would be good to imagine it. One way of avoiding this problem in the context of trust, then, is to claim that it is an attitude like imagination, which does not have rational requirements. Perhaps asking when trust is rationally justified is simply the wrong approach; trust may be an inherently arational attitude, with no standards of correctness. We trust and distrust depending on whether we want to, or whether we feel like it, but we are never right or wrong to do so. However, this view does not strike me as plausible. Claims of the form, 'You should not trust them, because …' can be perfectly reasonable. Even if it is not always clear what reasons justify trust, there does seem to be a sense in which some trustings are 'incorrect', while others are 'correct': sometimes, the other person fulfils our trust and sometimes they do not. This stands in contrast to imagination; what I imagine is neither 'correct' nor 'incorrect' regardless of whether what I imagine actually happens. This being the

[72] See, for instance, Parfit (2001, 21-2), Schroeder (2010), and Way (2012). Some, such as Rinard (2019), disagree that state-given reasons – for belief, at least – do not justify attitudes. It would take us beyond the scope of this thesis to address the arguments that can be raised against this view. I will assume for present purposes that only object-given reasons justify attitudes (that have standards of correctness – see below).

case, the distinction between state- and object-given reasons in the case of trust is pertinent. Trust, like intention and belief, can only be justified by reasons addressing its content.

But what kind of reason addressing the content will do? This brings us to the second way in which a reason can be of the wrong kind: it is not conducive to the kind of attitude to be justified. A belief that *p* is not justified by factors that make *p* being the case desirable, even though such a reason is object-given. A belief can only be justified by factors that make its content more likely to be true: evidence. Similarly, an intention to ϕ is not justified by evidence that one will ϕ, but by considerations that make ϕing desirable or good.[73]

So, the right kind of reason for an attitude must both address the content of the attitude and do so in a way that befits that attitude. For belief, evidence of the content is required; for intention, practical reasons for the content are required. It is intuitive that trust will require both evidence and practical reasons to be justified. It is in a sense more active than belief; it is concerned with what might happen – and crucially, with what we *want* or *need* to happen – rather than with what actually is the case.[74] At the same time, we do not have direct control over the outcome. Unlike intention, trust is concerned with someone else's agency, not our own. It therefore also requires evidence, since it is not in our power to ensure that trust is fulfilled, as it typically is for an intention.

To find the reasons that justify trust we must consider what kind of attitude it is and what its content is. It is worth remembering that some of the practical reasons that are often thought to justify trust are object-given. In the last chapter, we saw that this is the case for at least some friendship-based reasons, but it can also hold in less personal relationships, such as those maintained for the sake of mutual profit. I may trust my business partner to do certain things and part of my reason for doing so is the fact that their doing so will make me money. This is a practical reason concerned with the content of my trust; it is not my trusting itself that will be profitable to me.

Trust as Reliance on Trustworthiness

I suggest that to trust someone is to rely on them to be trustworthy.[75] This view is inspired by both Richard Holton (1994) and Pamela Hieronymi (2008). Holton argues that trust is a special kind of reliance, but he includes state-given practical reasons among the potential ways of justifying it (Holton 1994, 69). As I will show, my view of trust as reliance on trustworthiness can avoid that problem. Hieronymi argues that only evidence of trustworthiness counts in favour of trust, but does not acknowledge that some practical reasons are object-given and might therefore be genuine reasons of trust (Hieronymi 2008, 232). My view incorporates the insight that, when it comes to evidence, it is that of trustworthiness, rather than performance generally, that matters. However, since I take trust to be a kind of reliance rather than belief, it includes some practical reasons as well.

[73] We may need other evidence to make intentions rational. For instance, we might need sufficient evidence that the action is possible. However, this is not evidence that the content of the intention is true.

[74] An exception is the case of trusting testimony. This was discussed in Chapter 1 and will be discussed further below.

[75] This view is argued for at greater length in Chapter 2. Here, I merely give the idea some motivation.

I am not here giving a full account of trustworthiness; this theory of the justification of trust is designed to fit with any plausible conception of it. However, it is necessary to establish some minimal requirements of what it means to be trustworthy to guide discussion. To begin with, being trustworthy involves performing of one's own accord, rather than because of fear, coercion, or manipulation. The motivation, as Holton puts it, needs to be self-generated (1994, 65-6). We do not trust by, for instance, relying on the other person's fear of us when we ask them to do things; correspondingly, one who acts out of fear is not being trustworthy. Another, related, requirement, is that the trustworthy person must act in such a way that they can be held responsible for what they do (Hieronymi 2008, 224). They must have sufficient control and awareness of their actions and their consequences that they are valid targets of praise or blame. Otherwise, it does not make sense to see them as potential targets of either betrayal or gratitude, which is necessary when trusting someone. A simple way of summarising both these points is that we do not just rely on another doing something when we trust them; we recruit their agency (Jones 2012, 65). Accordingly, being trustworthy involves placing one's agency to some extent at another's disposal.

One reason to favour the view of trust as reliance on another's trustworthiness is that it is in a certain sense very conventional, with several philosophers' theories taking the trustworthiness of another, not just their performance, to be at least part of the content of trust. For instance, Jones's (2017, 99) theories of trust and trustworthiness are as follows:

> A trusts B to ϕ if and only if A is optimistic that B will be moved to ϕ by A's reliance on B to ϕ and that B is competent with respect to ϕ.
>
> B is trustworthy for A with respect to ϕ if and only if B would be moved by A's reliance on them to ϕ were A to do so and B is competent with respect to ϕ.[76]

These entail that trust is optimism about the other's trustworthiness. The pattern is also followed by Hawley (2019, 9; 76), according to whom:

> A trusts B to ϕ if and only if A believes that B has a commitment to ϕ and relies on B to fulfil it.
>
> B is trustworthy for A with respect to ϕ if and only if B fulfils any commitment to A to ϕ.

Thus, to trust is to believe that another has a commitment and to rely on them to be trustworthy with respect to it. Or again, take Russell Hardin's (2002, 1; 1996, 31) view:

> A trusts B to ϕ if and only if A believes that B will encapsulate A's interests with respect to ϕ within B's own interests.
>
> B is trustworthy for A with respect to ϕ if and only if B encapsulates A's interests with respect to ϕ within B's own interests.[77]

[76] This is a slight simplification of Jones's view, which refers to domains of interaction rather than specific behaviours, and compelling reasons rather than being moved. See also her (2012, 70-1).

[77] I infer this view from Hardin's (1996, 31-9) discussion of trustworthiness being reinforced by incentives that bring one's interests into line with those of the trustor.

From this, we can infer that, on Hardin's view, trust is a matter of believing the other person to be trustworthy.

We have also seen a version of this idea in Hieronymi (2008, 232), in her view that trust is a belief about what another will do, justified by the thought that they are trustworthy, although strictly speaking, Hieronymi does not take trustworthiness to be part of the *content* of trust. For her, the content of trust is merely the other's performance (Hieronymi 2008, 232).

The view of trust as reliance on trustworthiness is therefore a simple application of a very common pattern. Philosophers may disagree about what trustworthiness is and about what kind of attitude trust is (optimism, belief, reliance, …), but many would agree that a crucial part of the content of trust is the trustworthiness of the person in question.

The Reasons of Trust

Since this chapter is primarily concerned with what justifies trust and not what trust is, it might be wondered why the theory of trust as reliance on trustworthiness has been proffered. This might be thought to not be the appropriate place to present an account of trust and that there is little point in repeating the account argued for in Chapter 2. However, there are two points to be made in defending its place here. Firstly, any theory concerning the reasons for an attitude must be based on at least a vague idea of what that attitude is. We could not establish the reasons for belief, for instance, without acknowledging that believing a proposition entails taking that proposition to be true. Similarly, we need at least a broad understanding of trust in order to give the reasons that support it. Secondly, note that this is only a vague theory of trust. I have not asserted what it means to be trustworthy, or what it means to rely.[78] These would need to be unpacked if we hoped to find a full account of trust. The purpose here is to give a theory of the reasons of trust which can then be adapted to any plausible theory of reliance and trustworthiness.

This account entails that the kind of attitude trust is is one of reliance and that its content is the other person's trustworthiness. Therefore, the reasons that justify trust will be those which befit reliance and specifically address the trustworthiness of the one trusted. This gives us restricted sets of both evidential and practical reasons, as we will see.

The reasons which befit reliance are both practical and evidential. If one is considering whether to rely on a rope to bear one's weight, for instance, one will consider the practical reason of the extent to which it would be beneficial if the rope was strong enough, as well as that of the extent to which it would be costly if it was not. One would also consider, if it was available, the evidence for and against the rope actually being sufficiently strong.[79] The rational thing to do would be based on a calculation similar to that given in the previous chapter for Coleman's theory of the reasons of trust. One weighs up the expected benefit of the rope holding, the expected cost of it breaking, and the probability of it being able to hold one's weight.

In trusting someone to ϕ, one is not merely relying on them to ϕ, but relying on them to be trustworthy in ϕing. Therefore, evidence that they will ϕ will not justify trust, unless it is

[78] See Chapter 1 for discussion of reliance. A theory of trustworthiness is proposed in Chapter 3.
[79] This example is based on one in Holton (1994, 68).

evidence that they will be trustworthy in ϕing.[80] Let us consider again the cases of Lief and Katherine in the variations on Simpson's (2017, 177) 'Antarctic Resupply' considered near the start of the last chapter. In the original, Lief is planning an expedition to the South Pole, hoping to set a new record for the fastest unsupported journey. In order to minimise the weight of his sledge and so maximise his speed, he does not take provisions for the return journey. Instead, he asks Katherine, who runs an adventure support company, to drop supplies at the Pole when he contacts her on his satellite phone. This she agrees to do and Lief sets out on his journey. However, we also considered alternative cases in which Katherine has a strange habit of dropping supplies at the South Pole that Lief decides to take advantage of, in which she is dropping supplies for a different group and Lief can get to them first, and in which Lief has coerced Katherine into agreeing to the supply drop.

Now, if Lief goes to the South Pole, then he is relying on Katherine to deliver supplies. But he is not trusting her unless he is relying on her to be trustworthy in doing so. If she drops the supplies out of habit or because she has been coerced, as in the first and third adapted stories, then she is certainly not being trustworthy on any plausible account of that concept. Therefore, if these considerations are Lief's evidence, then he is not justified in trusting her, though he may be justified in relying on her. Conversely, if his evidence consists in knowledge of her reputation for being impeccably trustworthy, or in recognising a strong motivation on her part to be trustworthy – where this motivation is not based on coercion but on something more self-generated – then trust, not just reliance, is justified. Of course, it is still open for someone to suggest that it is not rational for Lief to restrict himself only to evidence of Katherine's trustworthiness, since what really matters to him, with the stakes so high, is her actual performance; whether she is trustworthy or not in following through is immaterial. This may well be true, but it is no objection to the view I have advanced; as was discussed in the last chapter, one can simply say that Lief's is a situation in which trust is not rational precisely because one ought to be considering all of the evidence relevant to Katherine's performance.

What of the other version of Antarctic Resupply, 'Freerider', in which Lief takes advantage of Katherine making a supply drop to researchers in the area? In making the drop, Katherine may well be acting in a trustworthy manner. Lief, in setting out for the South Pole, is relying on her to do so and – let us suppose – is relying specifically on her trustworthiness. It seems that my view incorrectly implies that Lief trusts Katherine, even though she is not delivering the supplies for his benefit, or even knows of his plan. I will address this issue more fully below as what I call the 'Third Party Problem', wherein I argue that it may sometimes be possible to trust people to keep their word to others, but not in cases like 'Freerider'. For now, we can sidestep this problem with a natural refinement of the concept of trustworthiness. Namely, it is three-place: just as A trusts B with respect to ϕ, B (if the trust is well-placed) is trustworthy for A with respect to ϕ. One can be trustworthy for some people but not for others and for some tasks and not others.[81] In the view of trust I am adopting, one does not merely rely on the other

[80] A further criticism of Hieronymi's view is that she takes the content of trust to be performance, rather than trustworthy performance (2008, 232). Her claim that only evidence of trustworthiness counts therefore appears arbitrary (although she acknowledges in a footnote that her view may require emendation if the content of trust is more complex). On my view, however, it is clear why evidence of trustworthiness is the only evidence that counts: that is the content of the attitude (so my view arguably provides the required emendation to Hieronymi's).

[81] This is more fully spelled out in my account of trustworthiness in Chapter 3. For another explicitly three-place account of trustworthiness, see Jones (2012, 70-1).

to be trustworthy, but relies on them to be trustworthy *for oneself*. Thus, Lief does not rationally trust Katherine in 'Freerider', since he knows that she is being trustworthy for the researchers, not for him.

Practical reasons also count, so long as they are concerned with the content of trust, not with the trusting attitude itself. That is to say, they must be object-given, where the object is the other person being trustworthy. The practical reasons of trust must show why it would be of benefit to oneself for another to prove trustworthy in ϕing on this occasion. In practice, there is likely to be significant blurring of boundaries between being trustworthy with regard to ϕ and simply ϕing; part of the reason for trusting someone to ϕ will typically be that their ϕing is beneficial, as in cooperative business arrangements, and trustworthiness does entail performance. If somebody is trustworthy in ϕing, then they ϕ. A practical reason for relying on their performance will thus often be a practical reason for relying on them to be trustworthy and vice versa. When we trust someone to do something and they do it, it is usually an act of trustworthiness, but the two can come apart.

For example, if B has promised A that they will ϕ, then one reason for A to trust B to ϕ would be the fact that it would be useful to A if B proved trustworthy with respect to ϕ. This will often be derived from the usefulness of B's ϕing; A needs ϕ to get done but cannot do it themselves, so enlists B's help. But suppose instead that A found it more expedient to coerce or manipulate B into ϕing, or to rely on B's promise to ϕ for someone else without being party to the agreement, or that B is just going to ϕ anyway and A can make use of B's ϕing. These may still be reasons to rely on B to ϕ, but they are not reasons supportive of trust. Practical reasons that tell in favour of manipulating someone into ϕing or taking advantage of the fact that they will ϕ are often, like practical reasons of trust, based on the value or usefulness of that person ϕing. But such reasons do not tell in favour of relying on someone to be trustworthy, since someone who is being manipulated, or being relied on without their knowledge or consent, is not being trustworthy. In the case of manipulation, B's motivation is not sufficiently self-generated; in the case of uninvited reliance, it is not appropriate for A to feel grateful or resentful – still less betrayed – however B performs with respect to ϕ.

According to this view, Lief does not trust Katherine in the adapted versions of 'Antarctic Resupply' described above. In all of them, Lief has strong practical reason to rely on Katherine's supply drop – he greatly desires this expedition to be a success and, of course, to survive – but, insofar as he does not care about whether Katherine is trustworthy, but only about her performance (which may be perfectly rational given the stakes), he does not have practical reason to trust her. Indeed, the circumstances of the altered stories are such that Lief knows that Katherine will not be trustworthy for him. In Simpson's original case, it is feasible, though not guaranteed, that Lief does trust Katherine. It is stipulated that Lief relies on Katherine's performance and the situation does not preclude her being trustworthy. He therefore may have practical reason to trust her in that case, to rely on her to be trustworthy for him, since her being trustworthy is the most likely way for her to perform. This holds even if it is not, all things considered, rational for him to trust her.

So, being reliance on another's trustworthiness, trust is justified by evidence that the other will be trustworthy and by practical reasons for valuing their trustworthiness.

This, in very general terms, is the answer to the question of what kinds of reasons are of the right kind for supporting trust. They are both epistemic and practical. The epistemic reasons are those which suggest that the other person is likely to be trustworthy. The practical reasons are those which make it good or desirable that they be so. Let us now address this issue more specifically, with some examples of what trust as reliance on trustworthiness implies are the right and the wrong reasons of trust.

The Account in More Detail

To begin with the right kinds of practical reasons, we have said that these must be object-given, so have to appeal to trustworthiness. They must show that it is worth relying on the other person to be trustworthy, as it may be worth relying on a rope to support one's weight. So, one example of a practical reason for trusting someone would be that, if they prove trustworthy, a task that would otherwise be very difficult or impossible would be accomplished (again, 'Antarctic Resupply' comes to mind). Another would be the increased efficiency of cooperation; less time, effort and resources need be expended if the other party is trustworthy. Yet another would be mitigations to potential losses, should the other person turn out not to be trustworthy on the given occasion. If the stakes are very high, such that one would lose a great deal by trusting someone who is not trustworthy, then it would take a great deal to convince one to trust them. This is essentially the point made about the rationality of Lief trusting Katherine; his life is at stake, so it will not be rational for him to trust unless he has extraordinary reason to. So, one factor that would make trust more rational would be lower potential losses.

Turning now to the wrong kind of practical reasons for trust, there are many examples. My view would entail, for instance, that trusting someone for the sake of improving their trustworthiness – so-called therapeutic trust[82] – would be a reason of the wrong kind. Although it mentions their trustworthiness, it is their hoped-for future trustworthiness that is addressed, not their present trustworthiness. It is therefore a state-given reason; it presents a possible benefit of having the trusting attitude towards someone – they become more trustworthy – rather than addressing the content of the attitude. Other reasons of the wrong kind would be trusting in order to improve or maintain a relationship, trusting because it is convenient to do so and trusting because trust is thought to be valuable in itself. Each of these is a state-given reason, not addressing the trustworthiness of the trusted party.

For those who are inclined to think that such reasons do justify trust and that my theory is therefore erroneous in ruling them out, let me briefly highlight some likely features of trusting for such reasons. Firstly, these reasons will still hold even if the trustor knows that there is no chance whatsoever of the other's performance. This means that, if one or more of these reasons is sufficiently strong – if it is extremely important that one's friendship is maintained and trusting one's friend is the only way to ensure that it is, for instance – then one will still be justified in trusting the other when one knows for a certainty that they will not perform. This would be very odd. Even given that the content of trust is not mere performance and that trust is not a kind of belief, surely rational trust in someone to ϕ is incompatible with certain

[82] As defined by Jones (2004, 5).

knowledge that they will not ϕ.[83] Indeed, even if the person in question is known to be highly untrustworthy, the view that state-given reasons can justify trust would imply that it can be rational to trust them. Secondly, consider the dispositions of somebody 'trusting' another under such conditions. They would be inclined to check on the other's progress to ensure that it was being done, or to find opportunities to mitigate the risk of it not being done. If it were possible, they would find some way of doing it themselves, or find someone else to do it. These are not the dispositions of one who genuinely trusts. Rather, this would appear to be a case of only acting as though one trusts. For instance, if one were to therapeutically trust a child, one would take steps to ensure that the costs are minimised should they fail to perform.[84] This is not recruiting the agency of another person, as is required for trust, but rather managing another person. One does not even seem to be relying on them, since their non-performance would not lead to a negative outcome. Therefore, these reasons at best support acting as if one trusts, not trusting itself.

However, it should also be noted that each of these kinds of reasons is very close to being a genuine reason of trust, which helps explain why they seem initially correct. In each case, it is likely that it is in the other's interest to be trustworthy, which would give us evidence of their trustworthiness. Furthermore, their being trustworthy would likely benefit us, which is a practical reason of the right kind for trusting them. Note that these are not the reasons actually appealed to in presenting the supposed force of therapeutic, relational, convenient and value-based trust. Rather, philosophers focus on the attitude of trust itself being in some way good, helpful, or beneficial.[85] But they do tend to accompany them.

For example, take trusting another for the sake of friendship. Another of Holton's examples involves two rock climbers, one more experienced than the other, scaling a cliff face. The less experienced is at one stage faced with a choice: take hold of a securely anchored rope to pull themselves up, or take hold of their partner's outstretched hand. Both might be equally likely to bear their weight, but there is an additional reason to take hold of the offered hand: the gesture of trust enables the relationship between the two to move forward a little. (Holton 1994, 69) Would such a relationship-based reason be permitted on my view? This depends on whether trustworthiness – and so actual performance – matters, or just the gesture of trust. In this case, it seems that the more experienced partner being trustworthy and pulling up the other does matter. Taking the hand is not enough for the relationship to move forward; it would likely move quite far backward if they let go and allowed the less experienced climber to fall. Since trustworthiness matters, this is an object-given practical reason to trust, so is a genuine reason. However, there may be circumstances in which it does not matter whether the other person proves trustworthy or not. If all that is needed is the gesture of trust for the relationship to move

[83] According to Holton, who takes trust to be a kind of reliance, 'One of the constraints on trusting you to do something is the lack of a belief that you will not do it.' (1994, 76). The point made here is even less contentious. Since state-given reasons take no account of the content of trust, they still apply no matter what is known of the content. Unlike Holton, I do not claim that it is not possible to trust for such reasons, only that doing so is not rational.

[84] This is not to be confused with the cost of non-performance already being low, which, as argued, will make trust more rational. This is the practical analogue of Simpson's (2017, 189-90) distinction between considering the evidence that one has and seeking further evidence. Insofar as one is seeking further justification – whether by looking for evidence or by mitigating the risks – then one's trust is either not genuine or not justified.

[85] For instance, Frost-Arnold (2014, 1959-63) argues that there are at least three kinds of trust that we can choose for practical reasons. All the reasons she gives are state-given rather than object-given. For discussion of her view, see Chapter 1.

forward, then the furthering of the relationship is only a state-given reason. In practice, it may be hard to tell the two kinds apart. This is what I mean when I say of the reasons mentioned above, although they are of the wrong kind, that each is very close to being a reason of the right kind. We can imagine scenarios in which a very similar sort of reason is object-given. This may be why such reasons can look so plausible; they are almost right.

What about practical reasons that address performance rather than trustworthiness? Again, although it will often be hard to tell such cases apart in practice, I think my account gives a suitable answer. Suppose that, in Holton's case, the more experienced climber is an instructor who will likely lose their job if they fail to ensure the safety of their partner, whom they are teaching. Personally, though, they do not particularly care about their pupil. Knowing all this, the less experienced climber takes hold of their hand, relying on the instructor's fear of the consequences should they let go. Whether the instructor is trustworthy does not even occur to the pupil, or perhaps they consider them outright untrustworthy. But they are reliant on a quite different motivation. It is unlikely that either the act of reliance or its fulfilment would move the relationship forward, since both know that the instructor only helped in order to maintain their employment. On my view, this is not trust, since what is being relied on is fear, not trustworthiness. That reliance may be rational, but the reasons justifying it do not justify trust.

Let us now consider the evidential reasons that trust as reliance on trustworthiness rules in and out. The right kind of evidence is evidence of the other's trustworthiness. As Hieronymi (2008, 226) points out, this will include their practical reasons for following through. If it is in their interests to keep their promises or to speak honestly, then this provides evidence that supports trusting them. This is similar to Hardin's view mentioned above; trust and trustworthiness are concerned with what is in someone's interests, not just with what they actually do. But this is not the only kind of evidence that is permitted; we do not need to perceive another's motivations or see things from their point of view in order to have evidence of their trustworthiness. A simpler kind of evidence would be knowledge of their reputation. Similar to this, our past experience can also suffice. If we know that this person has always or usually behaved in a trustworthy manner in the past, then this is good evidence that they will act in a trustworthy manner on this occasion, insofar as induction is a reasonable source of evidence. Failing that, we might appeal to situations that do not involve the person we are now considering whether to trust, but are similar in important respects. If people like this person tend to be trustworthy in situations like this one, then we have evidential reason to trust them. Of course, this last kind of evidence is open to abuse; one might trust or refuse to trust based on untrue stereotypes or prejudice. But that does not mean that, used properly, it is not genuine and helpful evidence.

What about the wrong kinds of evidence? Again, the theory I have suggested entails very intuitive results for the kinds of evidence that should not be considered in rational trust. The use of threat or force is ruled out, for instance. These might give very good evidence of performance, at least in part because they provide the other person with very strong practical reasons to comply. But they clearly do not give evidence of trustworthiness. In a similar vein, tricking or manipulating others into doing something does not give evidence for trusting them. Neither does knowledge that they are going to do something for entirely their own reasons or out of habit. These give evidence of performance, but not evidence of trustworthiness. They preclude the required recruitment of others' agency.

I believe that all of these results for the right and wrong kinds of reason for trusting someone roughly track intuitive judgments of how trust can be justified. Furthermore, trust as reliance on trustworthiness is capable of dealing with all the problems discussed at the outset of the last chapter, something that none of the other theories considered there managed. Let us go through them in turn. Firstly, we can accommodate Jones' (2012, 65-6) observation that one of the purposes of trust is efficient cooperation. Being able to cooperate better is a permitted practical reason of trust, since this is one of the benefits of the other person proving trustworthy. Secondly, we can explain why trust can rationally go beyond the evidence and be resistant to it, without being entirely immune to it (Baker 1987, 3-5). Trust may go beyond the evidence, in the sense of being stronger than the evidence merits, because evidential reasons are not the only reasons that may support it. If an instance of trust is supported by both evidence and practical reasons, then we should expect it to be stronger than the evidence alone merits, simply because it is justified by more than the evidence alone. Trust can be resistant to counterevidence for the same reason; it is not just evidence that counts. Nevertheless, a large body of evidence that suggests that someone will not perform will overcome rational trust, since this will indicate that they are, after all, not trustworthy. Thirdly, as has already been made clear, no state-given reasons are permitted, so the main problem faced by Pragmatists is overcome. Fourthly, we can also avoid the problem of evidence. It is possible, on this view, to have a large body of evidence supporting trust, while keeping trust distinct from ordinary rational belief, so long as the evidence is of the right kind. Like Hieronymi, we only permit evidence of trustworthiness, so there is no danger of trust being precluded by too much evidence of performance. Finally, applying this theory would allow Lief to take into consideration the consequences of Katherine proving trustworthy or not when deciding whether he should trust her. It is not just the evidence that matters to him; as was stipulated in the original case, it matters to him greatly that he survives. This desire is not a piece of evidence, but is surely relevant in his deliberations.

At the end of the last chapter, I suggested that we needed a more sophisticated version of Pragmatism to give a satisfactory account of the reasons of trust. Specifically, we needed a theory that would non-arbitrarily include some kinds of evidence but not others in order to maintain the distinctiveness of trust and that would non-arbitrarily include some practical reasons but not others, such that the wrong kind of reasons problem could be avoided. We have now seen that trust as reliance on trustworthiness can accomplish both of these. The evidence it permits is of a kind distinctive of trust. The only practical reasons allowed are object-given. This, therefore, is the more sophisticated Pragmatism that we need. I think that it provides a strong basis for rational trust. Nevertheless, some objections may be raised. In the next part, I consider and respond to three potential difficulties with the theory.

Part 2: Three Potential Problems

In this part, I respond to a series of objections that may be raised against the theory of rational trust that I have proposed. Each will be found to be unsuccessful, but the process of dealing

with them will further illuminate the account. The fact that these issues can be dealt with successfully also adds to its appeal.

There are three problems that I will consider in turn. The first I call the third party problem, which asks whether it makes sense to trust someone to do something for someone other than ourselves – one is a third party to the promise. The second is the problem of testimony, which calls into question the view that trust is a kind of reliance, not belief, given that trusting someone's testimony appears to require believing, not merely relying on, what they say. The final objection is a version of the wrong kind of reasons problem, asking whether we can rationally withhold trust for reasons unconnected to the relevant person's trustworthiness.

The Third Party Problem

Suppose that one knows that a promise has been made to a friend. Can one trust the promisor to keep the promise, given that it was made to someone else? It certainly seems possible to rely on them to be trustworthy, since whether the promise is kept or not might affect us (think of Lief and Katherine's promise to the researchers in 'Freerider'), so my theory appears to imply that one can trust the promisor. But this would be odd; it is surely not the place of someone who is not party to a promise to trust that it is fulfilled. Even if it will affect one's plans, it does not seem right to say that breaking the promise would wrong someone who just happens to know about it in the way that it would wrong the person to whom the promise was made.

One way of dealing with this is to bite the bullet and say that we can trust people to do things for third parties, at least as long as it is relevant to our own plans and goals. Katherine Hawley (2014, 11) gives the following example:

> [S]uppose your daughter's friend promises to her (not to you) that she will stay to the end of the party and give your daughter a lift home. Suppose you rely upon the friend to keep this promise: you drink several glasses of wine, making it impossible for you to safely drive and fetch your daughter yourself. I will take it that you trust your daughter's friend to keep her promise to your daughter.

Hawley's view does seem rather intuitive. After all, if the promise is broken, the job of ensuring that your daughter gets home safely will land on you, the parent. Perhaps because of this, it does seem that the promise-breaker wrongs you just as much – and plausibly more – than they wrong your daughter. The promise may not have been made to you, but the duty to keep it may be at least partly owed to you.

Now, it may well be that this is the case sometimes. In making a promise, it is plausible that one sometimes incurs a duty to people other than the one receiving the promise. It might even be that, in such circumstances, the promise is also made to various others who are not the addressees of the speaker. But whatever the duties of the promisor to third parties and how they are grounded, this does not fully solve the problem. There are cases in which promisors clearly do not have duties to third parties, even third parties who base their plans on the promise being fulfilled.

Suppose that I overhear you making a promise to someone else. The promise has nothing to do with me; it is about something between the two of you which is none of my business. However, I see a way in which it could be to my advantage that you keep the promise and thinking that

you probably will, I form plans around you doing so. In other words, I rely on you to be trustworthy. For example, you might be promising to lend your friend a sum of money. This friend of yours is someone I hope to go into business with and I realise that this loan would enable me to finally follow my dream of opening a shop, if I can persuade them to invest, which I am sure I can. Accordingly, I make the necessary arrangements, assuming that the funding will be available. I stand to lose a lot if you do not keep your promise. Yet you do not owe me anything and do not wrong me if you fail to follow through. I have no right to depend on you in this way.

The difference between Hawley's example of your daughter being promised a lift home and the example of a promised loan seems to be this: in the one case, the promise was about something that already affected you, so it is a legitimate consideration to work into your plans; in the other, the promise was nothing to do with me and I formed my plans based on it. With or without the promise of a lift, you would need to have some plan concerning how your daughter was going to get home, so the promise affects you directly. But the promise of a loan does not affect me directly until after I hear it and decide to make use of it. I choose to involve myself when I need not do so in the second case, but you are already involved in the first.

Now, this is not to say that it would not be rational for me to believe that you will keep your promise, or to rely on your doing so. The point is that it is not an appropriate context for trust. So, the objection survives the initial answer that we can rationally trust as third parties, for there are at least some cases in which we cannot. In Hawley's case of the daughter getting home, it is plausibly appropriate to trust; in the case of the overheard promise, however, it is not. The view of trust as reliance on another's trustworthiness seems to get the wrong result on this; one can rely on the other to be trustworthy for the reasons discussed above. Regardless of the recipient of the promise, one might have strong evidence that the promisor will be trustworthy and it might be highly beneficial if they are. So inappropriate third party trust can be justified on my view, or so it seems.

But this is merely an appearance based on a superficial understanding of trustworthiness. Although I am not here offering anything close to a full account of trustworthiness, it needs to be noted again that it is a three-place predicate. Just as trust is of the form 'A trusts B to ϕ', taking in the two agents and the relevant behaviour, so trustworthiness is of the form 'B is trustworthy for A with respect to ϕ'. One can be trustworthy for some people but not others and with respect to some tasks and not others.

This was mentioned above when I discussed the question of whether Lief trusts Katherine in 'Freerider'. There, I said that it might sometimes be possible to trust as a third party, but not in cases like that. We can now see why this is. 'Freerider' is like the example of my eavesdropping on a potential business partner, rather than like Hawley's case of getting a daughter safely home. That is, Lief involves himself upon hearing of the promise, rather than being already involved. Had Katherine known that Lief was planning an expedition for the near future, and had then promised the researchers (not Lief) that she would make the supply drop, knowing that Lief would also be relying on it, then it might be different. In such a case, Katherine might reasonably be thought to be being trustworthy for Lief as well as for the researchers in dropping sufficient supplies for everyone.

To be more precise, then, the theory of trust that I favour is this:

> A trusts B to ϕ if and only if A relies on B to be trustworthy for A with respect to ϕ.

This is still not a full account of trust, nor does it give a full account of trustworthiness. But it helps us to see why this objection is misguided. Trust involves relying on someone else to be trustworthy *for oneself*. Third party 'trust' involves relying on someone to be trustworthy for someone else, rather than oneself, so does not fit this account. Again, it may sometimes be reasonable to rely on another's trustworthiness as a third party. But doing so is not trusting them and we can now see, with this more refined version, that my account does not imply that it is.

The Problem of Testimony

For the most part, we have so far been discussing trust in the context of performing given actions. However, trust is also applicable in the context of testimony. We can trust someone to do something, but we can also trust what they say. A reasonable theory of the reasons of trust should account for both these kinds of trust.[86]

However, the theory that I have proposed appears to struggle with testimony trust. This is because trusting what someone has said seems to require believing them. Take Baker's (1987, 3) example of the accused friend. You are told that they have committed some crime, but they protest their innocence. Regardless of the rationality of trusting them, it would be very odd to say to them, 'Although I trust you when you say that you are innocent, I do not believe that you are innocent.' Trust entails belief in testimony cases. Yet I have said that trust is a kind of reliance. Relying on something does not necessarily entail believing it, only working it into one's plans. As Baker herself puts it, 'If I trust her in such a situation, I do not merely stand by her, acting in ways that support her, either materially or emotionally. I believe she is innocent.' (Baker 1987, 3) Acting as though she is innocent, using that as a premise in one's practical reasoning, to use Frost-Arnold's (2014, 1966) phrase, is not enough. Sometimes, what is required when we trust is actual belief. Indeed, that may be all that we can offer; perhaps there is nothing we can do about the situation. We cannot merely work the supposition of her innocence into our plans, as it would not affect our plans, so mere reliance does not amount to anything.[87] On the view that trust is a kind of reliance, it may often be accompanied by belief, but belief is not necessary. So, how can a reliance-based theory of trust accommodate trusting testimony? Is there a way in which we can non-arbitrarily claim that trust does come with belief in the case of testimony?

The answer that Holton, whose view of trust is also reliance-based, gives is based on his view that trusting testimony is a specific way of trusting someone to do something: namely, it is trusting another to speak knowledgeably and sincerely (1994, 73). He holds that relying on someone to speak knowledgeably and sincerely entails believing what they say, in the sense that one is not truly relying on them to do so if one fails to believe (1994, 74). We can trust for practical reasons and thereby come to believe. So, in Baker's case of the accused friend, we might initially be undecided. The evidence, as far as we can tell, is inconclusive. But then our friend tells us that she is innocent and thereby invites us to trust her. Because she is our friend,

[86] What I argue in response to this problem is discussed at greater length in Chapters 1 and 3.
[87] It is for this reason that Hieronymi (2008, 219-21) takes trust to be a kind of belief, rather than reliance.

we choose to trust her. We do not choose to believe, but because we have chosen to rely on her testimony, we come to believe that she is innocent.

I believe that Holton is right up to a point. I agree that testimony trust is a type of trusting someone to do something and that, specifically, it is trusting them to speak truthfully. It is also correct that relying on someone to speak truthfully entails believing them. What else could it mean to rely on testimony? However, I think the way he unpacks this idea is incorrect. If we can come to believe in the way he suggests, then it seems possible to choose, albeit indirectly, what to believe. Moreover, we would do so for practical reasons, rather than evidential ones. As Hieronymi (2008, 221) points out in response to Holton, it is not rational to maintain a belief that one knows to be supported by reasons unconnected to its truth. The practical reasons which may support reliance are of the wrong kind for belief.

This is the central force of the problem, for our purposes. I have given a theory of the reasons of trust which is based on the reasons of reliance. Yet here we have a kind of situation, that of testimony, in which a different set of reasons seems relevant, since it is belief and not mere reliance that is required. So, Holton's view does not get around the problem under discussion.

The reason Holton's response fails is that it has the wrong order of priority for belief and reliance. It is true that reliance on testimony entails belief, but we should not infer from that that reliance on testimony gives rise to or causes – still less rationally justifies – belief. Similarly, it may be true that mastering a new skill entails that we have worked hard, since, had we not worked hard, we could not have done so. But we should not infer from this that we can come to have worked hard by first mastering the new skill. This would be to get things the wrong way round.

I suggest that, rather than bringing ourselves to believing what another says by relying on or trusting them, our belief constitutes reliance on their testimony. Reliance still entails belief, but rather than causing it, it is constituted by it.

When considering whether to rely, we must take into account the stakes of the situation. What do we stand to lose if what we are relying on does not happen? What is at stake when we rely on someone to act in a certain way is the success of our plans, our wellbeing, or various other practical matters that are to some extent important to us. For instance, what is at stake for Lief when he relies on Katherine to drop supplies at the appropriate time is the success of his expedition and indeed his survival. Rather more trivially, what is at stake in Holton's drama class example is one's bodily comfort; one will be temporarily hurt and might sustain mild bruising if one is allowed to hit the ground. This is an important question when considering kinds of trust: what will be lost if the one trusted does not perform?

When it comes to testimony trust, which involves relying on someone to be trustworthy in their speech – that is, to speak truthfully[88] – what is at stake is the truth of one's beliefs. What will be lost if the other fails to perform is truth. As Hieronymi (2008, 220) points out, one can voluntarily entrust one's bodily safety to another and do so for practical reasons, but this does

[88] Strictly speaking, being trustworthy in speech is not equivalent to being truthful. It is possible to speak the truth without being trustworthy, for instance by intending to lie but getting muddled, by talking randomly without regard to truth or falsity but still uttering some true propositions, or by speaking authoritatively on subjects one knows nothing about and luckily being correct. However, for current purposes, we will consider only those whose truthful speech is trustworthy.

not work for beliefs – or at least, it is not rational to do so. Relying on testimony requires belief and rational belief requires evidence. So, in the case of testimony, only evidential considerations are permissible.

Note that although only evidential reasons count, that does not mean that all evidential reasons count. It is still restricted to evidence of the other's trustworthiness. Thus, the distinctiveness of a trusting belief from an ordinary belief is maintained.

This view is similar to that which Hieronymi applies to trust generally. Is it therefore vulnerable to the same criticisms as hers? I think not. The restriction to evidential reasons is, on my view, only applicable to testimony trust, whereas Hieronymi takes it to apply to all kinds of trust. Recall that her view was unsuccessful because it could not account for the practical reasons that favour trust in cases like Holton's drama class; for testimony cases, it gives the right results. My view, though, can accommodate for cases like the drama class, since in such cases practical, as well as evidential, reasons count.[89] When what is at stake is a practical matter, practical reasons count; when what is at stake is a doxastic or epistemic matter, only evidential reasons do.

This treatment of testimony trust is a natural extension of the idea that trust is reliance on testimony. Whenever we rely on something, we are putting something at stake.[90] We reached the conclusion that the proffered theory entails that testimony trust requires belief merely by exploring what is being put at stake. Thus, we can accommodate the requirement of belief when trusting testimony without implying, as Holton does, that we can believe for practical reasons.

The Wrong Kind of Reasons Against Trusting

On the view that trust is reliance on another's trustworthiness, the right kinds of reasons for trusting are those which bear, practically or evidentially, on whether the other person will prove trustworthy. One of the features of rational trust that this explains is its resistance to counterevidence. But, as Baker (1987, 3-5) says, this resistance is not limitless. If we have overwhelming evidence to suggest that our friend is in fact guilty, despite their protestations, or that someone will not fulfil their promise, then we cannot rationally trust them. But on the view that I have presented, only evidence concerning their trustworthiness is to be considered. So, can we rationally withhold trust from someone – or relatedly, distrust them – if the overwhelming evidence of their non-performance is of a kind that does not impugn their trustworthiness? For instance, suppose that you know, but they do not, that they will be faced with insurmountable difficulties in keeping their promise? Or if you know that your friend is suffering from a loss of memory, so that she has likely forgotten about the crime she committed and is honestly speaking from the best of her knowledge in claiming to be innocent?

Similar points can be made for practical reasons. Among the wrong kinds of reasons for trust are trusting for the sake of convenience and trusting merely to improve the relationship, where the trustworthiness of the person is irrelevant to the force of the reason. But these do seem to

[89] This is not to say that all the reasons Holton suggests are genuine reasons of trust in that example, for at least some of them are state-given (Holton 1996, 69). The point is merely that some practical reasons do count.

[90] This is a further reason for favouring reliance-based accounts over belief-based accounts. If Lief trusts Katherine, or if one trusts the other members of the drama class, it is not just belief that is at stake. The trustor stands to suffer practical as well as epistemic costs.

be valid considerations when we decide *not* to trust someone. Perhaps we just do not need their help for the task in question and we would find it easier to do it ourselves. Or maybe it would be more convenient to entrust the task to someone other than this person. Or again, we may prefer to not develop a relationship with them and refusing to trust them is a way to avoid becoming closer. The practical reasons for trust are those which show how it will be beneficial if the other person were to prove trustworthy. We might expect, then, that the practical reasons for not trusting, or for distrusting, are those which show that it would not be beneficial if they proved trustworthy. None of these reasons do so, yet they appear to genuinely tell against trusting someone.

I will respond to these two facets of the problem, evidence and practical reasons against trusting, in turn. First, though, it is worth briefly marking the distinction between positive distrust and a mere lack of trust, for there is an important difference between them, just as there is between disbelief and the lack of belief. We lack trust in many people. Those we have never met or interacted with we do not trust. This extends also to objects; since trust is not appropriate to objects but only people, we lack trust in all non-person features of the world. By the same token, however, we also lack distrust in objects and in people with whom we do not interact. Or again, consider the cases of relying on others without trusting them. Clearly, while we lack trust in them, we do not distrust them. There are certain conditions that must be met before either trust or distrust become appropriate. What these are will vary depending on the theory of trust under consideration, but a minimal requirement is that we take ourselves to be the recipient of something like a promise or an assertion.[91] We will distrust those who we think are probably lying, or who are likely going to deliberately break their promises. But if they have said nothing and promised nothing, or if we are indifferent to their assertions and promises, then we neither trust nor distrust them.

So, what kind of evidence tells against trusting someone? The first thing to note is that evidence of non-performance, even if it does not directly bear on the question of whether the other is trustworthy, can be indirect evidence that the other person is not trustworthy. It is often the case that, if someone is trustworthy, they will perform, by keeping their promise or telling the truth. Insofar as this holds, any evidence that they will not perform will also be evidence against their trustworthiness. However, this will not always hold. There can be explanations of non-performance other than a lack of trustworthiness. Two examples we have already given are unforeseen and insurmountable difficulties in keeping one's promise and a loss of memory causing one to have false beliefs about one's past actions.[92]

For these kinds of cases, we need to draw on the distinction just made between not trusting and distrusting. As mentioned, certain conditions need to hold before either trust or distrust is appropriate. I suggest that evidence that does not bear on the other's trustworthiness can nonetheless tell against trusting them by showing that it is inappropriate to either trust or distrust. It is not so much that it makes trust irrational; rather, it shows that the context is not suitable for trust. To demonstrate this, consider first a fairly extreme case: suppose that someone promises you that they will draw a square circle. You know that this is impossible,

[91] Katherine Hawley (2014, 10), for instance, takes trust-or-distrust to be appropriate just in case we believe that the other person has made a commitment to us.

[92] The difference between accidentally and deliberately failing to do as one is trusted to do is discussed in Chapter 3. Here, I am talking of trustworthiness in a broad sense, rather than being trustworthy on a particular occasion, which always entails performance.

but let us suppose that they are innocently ignorant of the finer points of geometry and believe that they can do it. Is it reasonable to trust them, assuming that doing so would involve some potential loss for you?

Presumably, it would not be rational. You know that they cannot do what they have promised to do. They will not merely face difficulties; they will find it literally impossible. So you should not trust them. But what should one's attitude towards them be? Specifically, should we distrust them to draw a square circle? Again, it would seem not. Distrust involves seeing the other as less than innocent and we have stipulated that the promisor in this case is being as honest as they can.

The main issue is less with the trustworthiness element of trust and more with the attitude of reliance itself; it would clearly be irrational to rely on someone to draw a square circle. It is possible to rely without trusting and in such cases there are reasons connected to trustworthiness for not trusting – for instance, it is an object and so cannot be trustworthy, or they are someone who has not agreed to do the thing in question so their trustworthiness is not at issue. But where reliance is irrational, so is trust, since trust is a kind of reliance. It is this that explains the asymmetry of the right kinds of reasons: trust entails reliance, but reliance does not entail trust. Thus, reasons for relying are not always reasons for trusting, but reasons for not relying will always be reasons for not trusting.

In the case of the square circle promisor, we need not impugn their trustworthiness; it need not be a reason specific to trust that prevents us from trusting them. We can actually make sense of our response without appealing to trust at all, only reliance. It would clearly be irrational to rely on this impossible event happening. Since any reason against relying is also a reason against trusting, we also have reason to withhold trust from them.

By the same token, we might withhold trust in less extreme scenarios for reasons unconnected to the trustworthiness of the person in question. If you know that someone will be faced with serious difficulties in keeping their promise, though it is still possible for them, then it may be irrational to trust them, even if they are considered perfectly trustworthy. It would not be reasonable to rely on them. Or if we know that someone is suffering from memory loss, then, even if they are speaking as honestly as they can, we should not rely on them for accurate beliefs; though they are trustworthy, we should not trust them.

Similar considerations apply to the practical reasons we might have for not trusting. If we do not require assistance, or it would be more convenient to place our trust in someone else, or we have no desire to move the relationship forward, then we lack a reason to rely. Were we to trust them, it may or may not be beneficial to us if they proved to be trustworthy. We might therefore have or lack trust-specific reasons to trust them. But we would also have more general reasons of reliance to not do so. Once again, this need not amount to positive distrust; we do not have to regard them with suspicion or as driven by nefarious motives in order to lack a trusting attitude.

What, then, would justify distrust? Here, we must return to reasons more specific than those of mere reliance or non-reliance. Just as rational trust is reliance on the other to be trustworthy, for reasons connected to trustworthiness, so distrust is non-reliance on the other for reasons of

untrustworthiness.[93] If it is particularly likely on the current evidence that someone is untrustworthy and they can do a great deal of harm by going back on their word or lying, then distrust, not merely the lack of trust, is justified. Note how this leaves space for cases in which neither trust nor distrust is rationally justified. What is more, this space does not consist only of borderline cases in which a person's trustworthiness is unknown or subject to doubts; when we neither trust nor distrust, it may not be down to us being undecided about their trustworthiness, or even just because we have no particular dealings with the person in question. It could be that we are heavily reliant on them, or that we have a strong aversion to reliance on them, yet we do not either trust or distrust. I take it to be an advantage of this account that it allows for such cases. One of the principal tasks that philosophers of trust undertake, after all, is to mark the distinction between mere reliance and genuine trust.[94] On this account, we can make a firm distinction between rational trust and rational reliance, as well as between rational distrust and rational non-reliance. Before we are justified in either trusting or distrusting, we must consider the trustworthiness or otherwise of the person in question. Otherwise, the most we can reasonably do is rely on them.

Having considered each of these objections, we have found that the theory of trust as reliance on trustworthiness can answer them all. In so doing, the theory has become clearer and – I hope – more plausible. I will now conclude by considering how this theory fits with my own theory of trustworthiness.

Conclusion

I have now presented and defended a relatively simple theory of the reasons that justify trusting someone. Trust is justified by evidence that another is trustworthy, along with their trustworthiness being to one's advantage. Thus, the purpose of this chapter has been fulfilled.

It will be recalled that that purpose does not include providing a theory of trustworthiness. The idea of the view I have proposed is that it will be compatible with any plausible theory of what it means to be trustworthy. This notwithstanding, I do have a theory of trustworthiness which can be stated as follows:

> A is trustworthy for B with respect to ϕ if and only if A has committed to B that they ϕ and A ϕs.

This is argued for in Chapter 3. I will now demonstrate the implications of combining this account with that of the norms of trust that I have been advocating above. By applying my theory of the reasons that justify trust to my theory of trustworthiness, I hope to show the plausibility of the entire picture, as it were, rather than that of one piece of the jigsaw.

On my view, the theory of trust as reliance on trustworthiness entails that trusting someone is relying on them to fulfil whatever commitments that they have made to one. If you have promised me something, then I trust you in that promise just in case I rely on you to fulfil it.

[93] A more precise account of distrust, as opposed to a lack of trust, is given in Chapter 2.
[94] See, for instance, Baier (1986, 234), Holton (1994, 65), and Hawley (2014, 1-2).

What would justify trust on such a view? As a kind of reliance, it permits of both practical and evidential reasons, as we have been arguing. Let us consider the evidential reasons first.

Some kinds of evidence could rationally be taken into consideration, but not others. Specifically, evidence that suggests that they will keep the commitment is of the right kind. For instance, past experience with that person will matter. If they have always kept their commitments in the past, then this is good evidence that they will this time (barring some important difference in the kind of commitment being made), but if their record in this respect is poor, then this will be evidence against trusting them. Or, if one lacks personal experience with them, knowledge of their reputation will help, since this again provides inductive evidence one way or the other. If they have a reputation for keeping their commitments, then this provides evidential reason in favour of trusting them.

On the other hand, some kinds of evidence would not count. For instance, if you have strong evidence that someone habitually ϕs, then this would not count in favour of trusting them to ϕ in the absence of any commitment to do so. You have evidence that they will do it, but that is not the same as evidence that they will fulfil a commitment. By the same token, if you are planning to force them into ϕing, this would provide evidence that they will ϕ, but no evidence of commitment-fulfilment. So, these kinds of evidence would not support trusting them, given the commitments-based view of trustworthiness.

Similar considerations hold true for practical reasons. If their fulfilling the commitment would have particular benefits, then this gives us a reason to trust them. It may be something especially important or useful that they have promised to do. However, if it is just easier or more convenient to trust them, such as in Holton's drama class example, this would not provide practical reason for trust. In Holton's case, the deliberating about whether to fall does not include the issue of the other students actually fulfilling their commitment to breaking one's fall (assuming that they do have such a commitment). It is about the benefits of falling – of trusting, or appearing to trust – itself. It is not about their trustworthiness, here understood as commitment-fulfilment.

These implications of relying on trustworthiness do seem to be broadly correct. Factors like someone's reputation and personal past experience with them are the kinds of things that we would expect to be taken into account by a rational trustor. We would also expect them to consider the potential benefits and costs associated with the trustee being or failing to be trustworthy on the given occasion. As for the other kinds of reason, we might expect them to be considered by a rational actor, but they are not the kinds of reason that pertain specifically to trust.

Let us bring together the threads of the argument. In the last chapter, it was shown that neither Evidentialism nor the standard forms of Pragmatism are adequate. Noting the problems that each face, we have come to a reasoned compromise between them, showing how it is possible to non-arbitrarily include some of each kind of reason while excluding others. The theory of trust and its reasons that I have proposed implies rational grounds for trusting that are intuitively correct; they are the considerations that we would expect a rational trustor to take into account. In making the distinction between object-given and state-given reasons, as well as the content and attitude of trust – the other's trustworthiness and reliance, respectively – we have provided plausible grounds for these indeed being the right kind of reasons. I therefore

think that this theory, that trust is justified by evidential and practical reasons which bear on the trustworthiness of the other person, ought to be accepted.

In the next chapter, I will switch my general focus from trust and trustworthiness to the issue of commitments. As we have seen, that of commitment is a highly important concept in the overall set of ideas I have been trying to build and fit together. In ending, I wish to show its importance beyond the scope of my main topic by applying it to another substantial ethical matter: that of responsibility.

6

Commitment and the Responsibility Problem

Introduction

So far in this thesis, we have explored the questions of what trust is, what it means to be trustworthy, and when trust is rationally justified. A central concept in these interrelated issues is that of commitment. It is in the context of a commitment that one can be trustworthy or untrustworthy; it is another having made a commitment that makes them an appropriate target of trust. In this final chapter, I will consider commitments in a different context, to show their wider applicability in the field of ethics and so situate the foregoing discussions of trust and trustworthiness in a broader philosophical environment. Specifically, I wish to examine the relationship between commitment and responsibility.

There is a problem that arises when considering who is responsible for a harm or loss that is the result of reckless behaviour on the part of the one who suffers it. Normally, it is natural to assign responsibility to the one who behaved in that way. Having taken a clear and unnecessary risk, one is responsible for the resulting harm. But this seems to change in the particular case of risking harm from another person. If someone is victimised, the responsibility is generally thought to lie solely with the perpetrator, however reckless the victim was. They are fully absolved of responsibility for the harm they suffer.

The purpose of this chapter is to explore what accounts for this shift in responsibility. What difference does it make when the source of a harm is another agent, such that the risk-taker is not responsible for the harm? To be clear, we are not asking why or whether the perpetrator is responsible. That is not particularly puzzling; they intentionally and voluntarily caused harm to another, so any plausible account of responsibility would imply that they are responsible for it. Our concern is what prevents any responsibility from landing on the shoulders of the victim.

It will be argued that the difference is one of commitment. Breaching commitments not only entails responsibility for resulting harms, but also absolves the one to whom the commitment was made of responsibility. For instance, suppose that you promise me that you will pick up my medication, and, relying on your promise, I decide not to get it myself. If you fail to fulfil your promise, then you are responsible for the consequent worsening of my condition, and I am not, even though I could have avoided the harm by acting differently. It will be argued that all members of a society have an implicit, minimal commitment to not harm one another unprovoked. Therefore, by inflicting unprovoked harm on another, one not only incurs

responsibility for oneself, but also absolves the other of responsibility. One has not only wronged them by harming them; one has broken a commitment to them.

First, we will lay out a set of cases that serve to illustrate the problem. We then consider several responses that may initially seem plausible. Each of these will be found inadequate for various reasons, but the process will be instructive and will inform the conclusion that I ultimately draw.

Part 1: The Problem and Initial Responses

I here explicitly lay out the Responsibility Problem, showing why it is of philosophical importance. I will then consider some responses that may be made to it. None of these will be found to be satisfactory, but consideration of whether and when victims are responsible for their own harms will provide us with valuable insights for solving the problem.

The Responsibility Problem

We begin with three cases which shall guide our discussion. These are designed to bring out the key issues with which we will be concerned.

> **Lottery:** A person buys a lottery ticket, knowing that it is highly probable that it will be a losing ticket, but that they stand to gain a large amount of money if they win. The lottery is carried out fairly, without any deception on the part of the organisers. Their ticket turns out to be one of the many losing ones. They have lost the money that they spent on the ticket and received nothing in return.
>
> **Walker:** Somebody is walking home after work and has the choice of two routes. One follows a well-lit road with little traffic and no serious hazards. The other is a narrow, slippery path that winds its way along a cliff edge. They know the second to be more dangerous and there is nothing compelling them to take it. Nonetheless, this person opts for the second route and becomes injured as a result.
>
> **Mugger:** The same as ‘Walker’, except that the unsafe route is not a clifftop path, but winds through darkened parks and alleyways frequented by muggers on the lookout for unwary victims. Knowing this, the agent takes the unsafe route and is consequently mugged.

In ‘Lottery’, it is clear that the agent bears responsibility for their loss. They voluntarily took the risk, knowing all the relevant facts. They were not tricked or coerced and are not entitled to have their money back. Similar considerations apply with ‘Walker’. They did not need to take that path and they knew that it was dangerous, but they took it anyway. When they are injured, it is no one’s fault but their own.

But things seem different with ‘Mugger’. It is still the case that the agent takes an unnecessary risk, knowing what might happen as a result. Yet this time, when the harm occurs, we do not

consider the victim to be responsible. Responsibility for what occurred rests solely on the perpetrator and the victim is certainly entitled to be given back whatever was taken. Most philosophers agree that this is the case; it is wrong to hold the victim responsible. However, the reasons given for why it is wrong are often indirect. We should not consider the victim responsible because doing so leads to negative consequences – it reinforces false ideas, it lets the perpetrator off the hook, it leads to crimes not being reported, or it constitutes a further harm to the victim.[95] I wish to address the matter head-on: regardless of the consequences of doing so, why is it incorrect to say that the victim is responsible, given that in other cases of voluntary risk-taking, the one who comes to harm is responsible?

What the three cases have in common is the voluntary and informed risk-taking on the part of each agent. The difference between 'Mugger' and the other two appears to rest on the involvement of another person. Yet it cannot be *just* the involvement of another that negates responsibility, since 'Lottery' also involves others – the lottery organisers. In that case, though, it was still a matter of chance. The salient difference seems to be that another deliberately acted against the agent's interests, without their consent. The relevance of consent is a matter to which we will return in the next section. For now, I shall consider two responses that might initially seem reasonable, but which prove upon inspection to be inadequate.

The first suggests that the bare fact that the perpetrator is responsible in 'Mugger' is what absolves the victim. If responsibility lies with one person, the argument goes, then it cannot lie with another. This certainly highlights a difference with the other cases. In 'Lottery', the lottery organisers left it up to chance. Moreover, the players chose to enter, rather than being chosen by the organisers. In 'Walker', there is no other agent involved at all. So in neither case is there another reasonable candidate for responsibility.

However, this response ignores the fact that multiple people can bear responsibility for an incident. If more than one person is involved in bringing about a harm, then each will typically be at least partly responsible for the harm. Suppose, for example, that several criminals work together to rob a bank. Surely each bears at least some responsibility. It is not reasonable for any to point to their fellows and say, 'Since they are responsible, I cannot be.' The mere fact that another is clearly responsible does not absolve oneself, so this response does not support the intuition that the victim bears no responsibility for the harm inflicted on them.

Of course, the way in which the victim in 'Mugger' is involved in the crime is different to the way in which each robber is involved in the bank robbery. What the robbers did was clearly morally wrong, but, while, the victim's choice of route may have been crucial to the mugging taking place, they did not do anything wrong by taking it.

This brings us to the second response: since the victim did nothing wrong, they are not responsible for what happened. The only immoral actor was the mugger, so they are the only party responsible. However, this response is also inadequate, for two reasons. The first is that it depends on the view that doing nothing morally wrong is sufficient for lacking responsibility. This has incorrect implications for the first two cases. The agents in 'Lottery' and 'Walker' also do nothing immoral, so on this view are not responsible for their loss and harm,

[95] See, for instance, Brison (1993) Grubb and Turner (2012), and Hayes *et al* (2013). Much of the current literature focuses on victim-blaming for sexual assault, but the scope of this chapter includes all kinds of victimisation.

respectively. But that does not seem at all right. Having not acted immorally does not absolve one of responsibility for the consequences of one's actions.

The second is that it reverses the order of priority for responsibility and morality. At least part of what makes the mugger's behaviour immoral is the fact that they are responsible for the harm, not the other way around. If one harms another without being responsible for that harm, then one's action is not immoral. For instance, if one trips and accidentally pushes someone, the pushing is not immoral. Yet on this view, we must determine whether someone's action is immoral before deciding whether they bear responsibility for any harm that results.

Neither of these initial responses satisfactorily solves the problem. We turn next to a more sophisticated attempt, though one that is also unsuccessful.

The Consent Response

A large part of the wrong of mugging is the lack of consent. Indeed, this is plausibly all that is wrong with it; with consent, it would be an instance of gift-giving, rather than of theft.[96] This distinguishes the risk of being mugged from participating in a lottery; the lottery player has consented to the potential loss of their money by participating in the lottery. Similarly, the agent in 'Walker' has plausibly given consent regarding the risk of injury by voluntarily choosing the dangerous path.

Of course, we are not primarily concerned here with what makes a mugging wrong, but with the responsibility for the harm. It must also be remembered that this is not about choosing to suffer a harm, but about choosing to *risk* a harm. With this in mind, the consent response proposes the following principle. Where S is an agent and P is an event harmful to S:

> S is responsible for P only if S has consented to P or to the risk of P.

The power of consent is that it renders permissible what would otherwise be wrong (Hurd 1996, 123-4). This principle states that, by thus rendering (the risk of) something permissible, an agent also takes on responsibility for its occurrence. In 'Lottery' and 'Walker', it is suggested, consent to the risk of loss and of harm has been given, by buying a ticket and by taking the dangerous route respectively. The agents in those cases are therefore not absolved by this principle. In 'Mugger', however, consent is never given by the victim, either to the risk of harm or to the harm itself. This difference in consent accounts for the difference in responsibility. The above principle allows us to say that the agents in 'Lottery' and 'Walker' are responsible for their own misfortune, but that the assailed agent in 'Mugger' is not.

We might wonder if the principle proposed is plausible, since it may be contentious that consenting to the risk of something entails consent to the thing itself.[97] However, even setting this aside, problems with this view rapidly become clear upon reflection. To begin, consider the position of the agent in 'Walker' at the time of their alleged consent. They choose to take a

[96] Precisely what actions or attitudes constitute consent do not here matter. For discussion of the nature of consent, see, for instance, Raz (1981, 118-25), Alexander (1996), and Liberto (2021).

[97] David Boonin-Vail (1997, 290-300), for instance, disputes this in the context of the ethics of abortion, arguing that consenting to sex, which carries the possibility of creating a foetus, does not amount to consent to the existence of that foetus. I discuss a related issue below – namely, whether responsibility for (as opposed to consent to) the risk of a harm entails responsibility for that harm, should it occur.

dangerous path and this is supposed to constitute consent to the risk of harm. But why not say the same of the agent in 'Mugger' when they choose the more dangerous route? It is a different kind of harm, but if taking a route known to be more hazardous constitutes consent to risk in one case, why should it not constitute consent in the other?

A reply might be that, since in 'Mugger' the risk is posed by another person, consent must be given to that person. But the victim gives no consent – either to risk or to actual harm – to the mugger, having not even met them when they make their choice of route. Therefore, that choice does not count as consent. But this reply merely exposes a deeper problem with the whole view. In 'Walker', there is no other person at all. How can the agent in that case give consent by making a choice of path, but not the agent in 'Mugger'?

The underlying point here is that consent is an act of communication.[98] It cannot just be given; it must be given *to* someone. The consent response therefore accounts well for the agent's responsibility in 'Lottery', since there are other agents involved – the lottery organisers – and the purchase of a ticket plausibly counts as an act of consent to them regarding the risk of loss. But the agent in 'Walker' cannot have consented to the risk they take, since there is no one to whom they can give their consent.

If the advocate of this response insists that it is possible to consent to a risk without the consent being to anyone in particular, then we must puzzle over their conclusion about 'Mugger'. Why has the victim not consented to the risk of harm by choosing the mugger-infested path, without consenting to the mugger whom they later encounter? To be consistent, there must be consent here as well. But this backfires upon the purpose of this response – to show that the difference in 'Mugger' is the lack of consent.

This attempt to absolve the victim also fails. We next consider the counterintuitive idea that the agent in 'Mugger' is, after all, as responsible as those in the other cases.

Is the Victim in 'Mugger' Responsible?

The idea that agents who risk harms imposed by others are responsible for those harms when they occur is a contentious one. It seems not only incorrect, but morally problematic to lay responsibility on the victim. Two of the concerns associated with holding victims responsible are that doing so may be construed as victim-blaming, and that it might imply that the perpetrator is less blameworthy. For instance, if we were to argue that the victim somehow invited the greed and aggression of the mugger, and is therefore responsible for what ensues, we would be in morally dubious territory to say the least.

However, that is not the approach I will be considering. It is possible to suggest that the victim is responsible without implying these consequences. It should be noted, first, that arguing that a victim bears responsibility is not strictly victim-*blaming*, since one can be responsible for a harm without being blameworthy for it – as is the case in 'Lottery' and 'Walker'. Second, it need not entail that the perpetrator is any less responsible or blameworthy for the harm they cause. As has already been pointed out, the mere fact that one agent is responsible does not in

[98] As discussed by Archard (1997, 275-6) in his distinction between consent, an intentional act, and assent, a state of mind.

itself absolve anyone else. To say that the victim is responsible is therefore compatible with saying that the perpetrator is also responsible and that only the perpetrator should be blamed.

Although this view will ultimately be rejected, it ought to be considered for the sake of completeness. It will also be illuminating to examine the reasons that may be given for it. We begin by stating a sufficient condition for responsibility (SCR):

> If S voluntarily brings about P, then S is responsible for P.

Again, S is an agent and P is a harmful event (though it need not be harmful to S). SCR seems uncontroversial and I think we should accept it.[99] It explains why the mugger is responsible for the harm to the victim: they voluntarily brought about that harm, so should be considered straightforwardly responsible for it. However, this principle applies to those who bring about an actual harm; it is silent regarding those who bring about only the risk of harm. We must therefore adapt it if it is to be helpful in our cases.

Three possibilities come to mind when we consider agents who merely risk harm:

> SCR-1: If S voluntarily brings about the risk of P, then if P occurs, S is responsible for P.
>
> SCR-2: If S voluntarily brings about the risk of P, then if P occurs, S bears a degree of responsibility for P corresponding to the degree of the risk.
>
> SCR-3: If S voluntarily brings about the risk of P, then S is responsible for the risk of P.

Each of these looks quite plausible, though not all to the same degree. This is not an exhaustive list, but merely what seem the most intuitive ways of adapting SCR to cases of risk. Let us consider them in turn, beginning with SCR-1, the first and strongest claim.

SCR-1

According to SCR-1, the agents of all three cases bear responsibility for what happens to them. All have voluntarily risked a harmful event, so are responsible for it when it occurs. Even mitigated by the considerations mentioned above, this is highly counterintuitive – and not just in the case of 'Mugger'. Every time we drive, for instance, there is a risk that some harm will occur – a pedestrian might suddenly step in front of the car, another driver might fail to allow braking distance and hit us, a cyclist might unexpectedly pull out in the dark without lights. These are non-negligible risks that good drivers will be on the lookout for, but in none of those cases would the harm be our fault. We cannot eliminate the possibility of others being unsafe.

[99] Apparent counterexamples to SCR are moral dilemmas. If, faced with a terrible choice, we bring about the lesser of two evils, are we responsible for what has occurred? Presumably not, but SCR appears to imply that we are. I suggest, though, that such cases are not fully voluntary. I will not go into the matter here, but it seems plausible that certain sets of options, even if we are free to choose among them, preclude true voluntariness. See Joseph Raz (1988, 379-80).

Nevertheless, SCR-1 implies that we would be responsible for any collisions that occur. This does not seem a correct result, even if we can console ourselves that our responsibility does not entail blameworthiness.

It would therefore take a strong argument to convince us of SCR-1's veracity. One attempt to construct such an argument is based on a 'modest assumption' made by Judith Thomson (1986, 179): if a proposition says that something will happen and it does happen, then that proposition was always true. This suggests, for instance, that buying the ticket in 'Lottery' was always going to result in a loss and that taking the dangerous route in 'Walker' was always going to result in injury.

Let ϕ be the risky action that leads to the harmful event P. The argument for SCR-1 goes as follows:

1. ϕing = Bringing about P.
2. S ϕs voluntarily.
3. Therefore, S brings about P voluntarily.
4. If S brings about P voluntarily, then S is responsible for P.
5. Therefore, S is responsible for P.
6. Therefore, if S ϕs voluntarily, then S is responsible for P.

Here, the first premise follows from Thomson's modest assumption. If ϕing in fact brings about P, then it was always the case that ϕing would bring about P. ϕ just is that action which brings about P. The second is a stipulation of all three cases – the risky action is done voluntarily. The third follows from these. The fourth states SCR, which we have accepted. The final two lines follow from the others, with the final conclusion being a restatement of SCR-1.

If successful, this argument would show that SCR-1 is true and so we should accept it. However, it fails due to two problems. Firstly, Thomson's 'modest assumption', upon which the first premise is based, is really not very modest at all. It is essentially a limited version of fatalism, the view that propositions are true even before the events they describe occur.[100] It entails that the difference between bringing about an event and merely risking it is illusory – if it happens, then it was always going to happen; if it does not, then it was never going to happen. This is a strong philosophical claim that stands in need of justification. Furthermore, even granting that fatalism is true, the problem is not solved. For there is an implicit assumption here that responsibility depends on the truth of fatalism, which does not seem correct. We should not require fatalism in order to make sense of the concept of responsibility. Maybe our world is fatalistic, but in a non-fatalistic world, it should still be possible to say that someone is responsible for an event or not, using our ordinary notions of risks and hazards.

Secondly, the intermediary conclusion 3 depends on the rule of identity substitution. 'ϕs' in 2 is substituted for 'brings about P', which 1 tells us is equivalent. However, the use of this rule of inference is suspect. A famous counterexample concerns its application to the contents of beliefs. Hesperus is identical with Phosphorus – both are the planet Venus – but a certain person may not know this and believe the Morning and Evening Stars to be separate celestial bodies. If they form a belief that Hesperus is bright, so that the proposition 'This person believes that Hesperus is bright' is true, it would be illegitimate to combine this with the equally true premise

[100] I base this on the description of fatalism given in Conee & Sider (2014, 23-5).

'Hesperus is Phosphorus' to infer, 'This person believes that Phosphorus is bright'.[101] This point also holds for voluntary actions. One might voluntarily buy a lottery ticket that turns out to be a losing ticket without voluntarily buying a losing ticket. Or one might voluntarily take a route home that turns out to be a choice that gets one injured without voluntarily choosing to get injured. So, even if we accept Thomson's fatalism, the inference upon which the argument depends is illegitimate.

The underlying point of both these problems is that responsibility is not primarily concerned with deep metaphysical facts about reality. Rather, it is about the relation between subjects' mental states and the world, including what we could reasonably have foreseen. When we perform some voluntary action, we are absolved of responsibility if we could not have known or reasonably suspected the result in advance. Perhaps a more modest form of the argument might appeal to this point, arguing merely that one is responsible for that which one could reasonably foresee. The first premise could be replaced with 'ϕing = Bringing about a strong and foreseeable likelihood of P'. To get to the conclusion, the fourth premise would then need to be altered to 'If S brings about a strong and foreseeable likelihood of P and P occurs, then S is responsible for P'. This would make the agents in all three of our main cases responsible for what happens to them, since they all knew the likely consequences of their actions and proceeded anyway. But this does not solve the problem. The fourth premise is no longer SCR, but something much more controversial. It assumes that people are responsible for consequences that they knew to be likely, which is precisely the claim under scrutiny. Even in a more modest version, then, this argument will not work.

SCR-2

If SCR-1 is misguided, then what of SCR-2? This is a little less ambitious, since it only ascribes a degree of responsibility to the agent, that corresponding to the degree of risk. By 'degree of risk', we mean the probability that the harm will come about, rather than the severity of that harm. But how is 'degree of responsibility' to be understood and what does it mean for the two degrees to correspond to one another? To understand this, we would need an account of partial responsibility. This is a coherent notion and it seems particularly apt in cases in which various different people contribute to some outcome. For instance, if a china ornament is placed precariously on a mantelpiece and I carelessly knock it off, breaking it, it seems reasonable to say that I am only partially responsible for the breakage. I was careless, but then it was also careless of the other person to place the ornament such that it could be so easily dislodged.

However, SCR-2 is false not because one cannot be partially responsible, but because whatever makes it partial cannot be *just* the probability of the event in question. However the correspondence relation between risk and responsibility is to be understood on this view, a very low level of risk presumably entails a very low level of responsibility. But this means that if I take a bet, having correctly calculated that the odds of losing are very small, I am mostly absolved of responsibility if I do lose money. Is this correct? I think not. Gamblers are responsible for their losses whatever the odds.

[101] This example is based on one used by Frege (1948, 210).

Might a more modest version of SCR-2 do better? Suppose that, rather than being sufficient for the corresponding degree of responsibility, voluntarily bringing about the risk is only a contributing factor. Such a principle would seem highly plausible, but it misses the point here. If it is not a sufficient condition for any degree of responsibility, then it fails to deliver the result that any of the agents are responsible. Our aim is to explain the responsibility of the agents in our cases. Such a modified version of SCR-2 does not help us do this; it merely does not rule out responsibility in any case.

SCR-3

Neither of the principles so far considered stands up to scrutiny. This leaves us with SCR-3: If S voluntarily brings about the risk of P, then S is responsible for the risk of P. This should be uncontroversial, since it follows from SCR. It is essentially an application of the idea that one is responsible for that which one voluntarily brings about. Accepting SCR-3, then, we can say that the victim bears responsibility for the risk of being mugged, since they do voluntarily bring about that risk. If this is right, then what can we say of the victim's responsibility for the actual harm? Unfortunately, SCR-3 does not tell us. To say that voluntarily bringing about a risk is sufficient for responsibility for the manifestation of that risk begs the question. Our very purpose in this section has been to determine whether it is plausible that the victim bears responsibility given that they have brought about the risk to themselves. It seems, though, that none of the approaches we have taken entails that taking the risk entails responsibility for the outcome. I will therefore revert to the assumption that the victim in 'Mugger' does not bear responsibility for their harm.

Given that the victim is responsible for the risk of harm, and assuming that the victim is not at all responsible for the harm that they suffer, then we have a case in which responsibility for a risk of harm seems detached from responsibility for the harm itself. This is clearly unusual – it is the responsibility for the risk that compels us to think that the agents in 'Lottery' and 'Walker' are responsible for the harms that befall them. But it is precisely this strangeness that we are trying to explain and which makes this matter philosophically intriguing.

Responsible Victims

Despite the conclusion of the previous section, there is something more to be said for the general view that victims can be responsible for what happens to them, even if this is not the case in 'Mugger'. Consider, for instance, the following case:

> **Terrorist:** A high-level security official is negligent in their duties and this permits a terrorist attack to occur. The attack results in the injury of two people. One is the security official themselves; the other is an ordinary bystander.

It seems that the official in this case bears responsibility for the attack. They are of course not responsible in the same way as the terrorists themselves, since they are responsible only

through negligence, but their status is clearly different to that of the injured bystander.[102] The bystander, who, let us imagine, still knew that there was a significant risk of an attack in that area, seems more like the agent in 'Mugger'.

Take another example, this one an adapted version of 'Mugger'.

> **Trusted:** Person C gives their friend D something precious to look after and D promises to take care of it. D then proceeds to take the unsafe route home, through mugger-infested parks and alleyways. They are robbed and the precious item is stolen. C holds D responsible for its loss.

C does not seem unjustified in holding their friend responsible for losing their precious item. They trusted D, but D was careless, not showing adequate caution when they were supposed to be protecting something. Again, there seems to be a reasonable case for holding the victim, D, responsible. Note that this case is almost identical to 'Mugger'. The only difference is the involvement of an extra party, who also has something at stake when the victim takes their risk.

It might be argued that D is not really a victim at all. Only C suffers a loss, so only C can be considered a victim. We can assume, though, that the experience of being mugged is itself a harm, even if it is not one's own items that are stolen, so it does seem reasonable to consider D a victim in this scenario.

There are two responses to these examples that might seem plausible. One is to say that responsibility is relative. So, C might say that D is responsible for losing what they were given because they were negligent. D, however, would claim that the mugger is responsible, since they were the one who took the item in question. Both claims have some plausibility. Likewise, the security official might be held responsible by their superiors and by the bystander, while they consider the terrorists to be responsible.[103] The other response is to say that the risking of another's safety or possessions, rather than their own, is what makes the victim responsible. Because what D risked was not their own to do with as they pleased, they should have exercised greater care. The security official, meanwhile, risked not only their own safety by their negligence, but also that of the bystander. Let us consider each in turn.

Relative Responsibility

The first response portrays responsibility as essentially a three-place predicate. Rather than having the form, 'S is responsible for P', it has the form, 'R holds S responsible for P', or perhaps, 'S is responsible to R for P', where R is another agent.[104] This fits well with common

[102] I will not go into detail on the different kinds of responsibility here. It may be that, although the security official is responsible, they are not blameworthy in the same way as the terrorists themselves. But that is not our primary concern.

[103] This does not rule out the official's superiors, the bystander, and various others from also holding the terrorists responsible, as they undoubtedly would.

[104] This is arguably the view given by Strawson in his 1962 'Freedom and Resentment' (reprinted in 1974 and 2008 – I will reference the (1974, 1-25) version), in which he introduces the idea of reactive attitudes in response to the challenge posed by determinism to the idea of moral responsibility. Even given the truth of determinism, Strawson argues, we would still be resentful or grateful towards people based on their actions; our attitudes towards them would be so as to hold them responsible, even if determinism implies that they are not responsible

locutions concerning responsibility and related concepts: 'We hold you responsible for what happened'; 'I am accountable to my superiors in this'. It also fits with the way in which we use praise and blame, the assigning of which is one of the ethically significant purposes of the concept of responsibility. A person is not just praised or blamed; they must be praised or blamed *by* someone.

This does appear to neatly explain what is going on with the above two cases. In the first, the security official is responsible to the bystander, their superiors, and perhaps also to the public in general. The terrorists, meanwhile, are responsible to everyone else involved, including the security official who should have prevented the attack. In the second, the mugger is responsible to D for the theft, while C holds D responsible. This does not preclude C also holding the mugger responsible; multiple agents can be held responsible by the same person.

Nevertheless, this response is problematic. It makes responsibility much too relative. It should be possible to say that someone is responsible for something in an objective sense, and not have it depend on anyone holding them responsible. For instance, compare 'Mugger' to the similar case of 'Trusted'. If, in the latter, the victim can be responsible from C's perspective, what are we to say of the victim's responsibility in the former? Only that whether they bear responsibility is a matter of perspective. But this cannot be right – surely those who would hold the victim responsible are incorrect. We might even imagine that the victim considers themselves to be responsible, but if so, they are mistaken. Pushing the point further, the perpetrator might hold the victim responsible – from their perspective, they were going to mug whoever happened to come into the vicinity at that time; that is the fixed background against which their victim made their choice. But again, they are wrong.

Are there, on this view, resources to accommodate the wrongness of some instances of holding responsible? If it is to be at all plausible, then there must be. But once we set correctness conditions on holding responsible, we are back in the realm of responsibility being fundamentally two-place. For if it is correct for a given agent to hold S responsible for P, why not just say that S *is* responsible for P? If it is incorrect, why not just say that S *is not* responsible for P? This also relates to the issue of praise and blame that was mentioned in support of this conception of responsibility. To say that someone is praised or blamed, there must be someone doing the praising or blaming. But this is not the case when the question is whether someone is praiseworthy or blameworthy, which can be the case without anyone taking an attitude or reacting to the given person. Praise and blame are mistaken when the person in question is not praiseworthy or blameworthy. Similarly, it is incorrect to hold someone responsible when they

in any objective sense. Taking their cue from him, some philosophers, such as Bennett (1980, 20-5), Watson (1987, 256-86), and Wallace (1994, 74-83), have argued that the idea of *holding* responsible is explanatorily more basic than that of *being* responsible. However, this interpretation of Strawson has been challenged – see McKenna (2012, 46-49). In any case, I do not think that such a Strawsonian response would be appropriate here, since the reactive attitudes seem to be focused on others' good or bad behaviour – gratitude, resentment, moral indignation – and I wish to account for responsibility for the consequences of morally neutral behaviour as well, as in 'Lottery' and 'Walker'. Furthermore, in 'Walker', there is no one involved who might take a reactive attitude towards the injured agent.

are in fact not responsible. So although there is such an attitude as *holding* responsible, it does not eliminate the need for an account of *being* responsible.[105]

Another problem with this response is that it cannot deal satisfactorily with 'Lottery' and 'Walker'. In both, the agents are responsible for what happens to them. But to whom are they responsible? Who, on this view, is holding them responsible? This is similar to a problem we encountered with the consent response. If we require another agent to be involved in order to correctly attribute responsibility, then there is no way to explain the cases in which the agent is alone.

It does not, upon reflection, seem reasonable to consider responsibility to be quite so relative. Let us turn to the other likely response. As it is, we will find that it is not quite adequate. However, a modified version does give rise to a plausible principle.

Risking Harm to Another

According to the second response, the difference between 'Mugger' and the newer cases lies in the fact that what the victim puts at risk is not (just) their own safety or property, but those of others. The security official is responsible for the terrorist attack that injured them because part of what they were risking in being negligent was the safety of others. D is responsible for being mugged and losing C's precious possession because they were not risking their own possessions. The intuition here seems roughly to be that one has a right to take risks with oneself or what one owns, but no right to do the same with other people or what they own.

However, we are not here concerned primarily with people's rights, but with who bears responsibility for the negative outcomes of taking risks. Clearly, risking only one's own money or safety does not absolve one of responsibility, even if there is nothing morally wrong with it, as illustrated in 'Lottery' and 'Walker'. But the idea here is something like this: being *victimised* absolves one of responsibility when one has taken a risk *only* when what is risked is one's own.

However, there are examples which suggest that this view is not quite right. Suppose that D had not been lent the precious object, but it is their own. However, it is important to others that it is delivered safely to its destination and not stolen. It might be, for instance, a piece of artwork that is of historical, cultural, and aesthetic significance. Now, when D takes the unsafe route and is mugged, it seems that they do bear responsibility for the loss of the object. They are the victim of a crime and were not risking anything but what belonged to them, yet they are not (fully) absolved. An advocate of the second response might argue they are risking something that is not their own to risk, since others have an interest in the safety of the precious item and so D is not the only one who suffers some kind of loss because of their careless actions. This entails that D has a special duty to protect the artefact that they own. They owe it to others to

[105] Certain Strawsonian views, such as Bennett's (1980, 14-47), hold that the need is eliminated by reactive attitudes such as resentment and gratitude. Strawson himself says, in response to the idea that there must be something more to justify praise and blame than mere utility, that '[t]he vital thing can be restored by attending to that complicated web of attitudes and feelings which form an essential part of the moral life as we know it, and which are quite opposed to objectivity of attitude.' (1974, 22-3) Whether he should be interpreted as saying that reactive attitudes can fully explain responsibility without any sense of someone being 'objectively' responsible, and whether he is right if that is what he meant, are issues that will take us too far afield here.

look after it, because, even without it belonging to anyone else, they still have a genuine interest in its being kept safe.

Following this line of thought, it seems plausible that the security official had a special duty to prevent the terrorist attack. In the case of C entrusting D with a precious object, D similarly has a special duty to not take unnecessary risks. It is part of the security official's job to not permit such things to happen; they are not just another victim. D agreed to look after C's property; they owed it to C to take care of it. If this is right, then the following principle appears reasonable:

> If S has a special duty to avoid P and S is responsible for the risk of P, then if P occurs, S is responsible for P.

According to this principle, a victim can be responsible for a harm, provided that they are responsible for the risk and they had a special duty to avoid the harm. In 'Mugger', the victim is responsible for the risk, but do they have a special duty to avoid the harm? It appears not – they do not seem to be in a position analogous to either D or the security official.

Earlier, we considered three ways of applying SCR to cases of risk. Only SCR-3 was found to be appropriate: the most we can say of someone who voluntarily takes a risk is that they are responsible for the existence of that risk, not for its manifestation. But in the cases of 'Terrorist' and 'Trusted', it does seem that responsibility for risk entails at least some degree of responsibility for the harm itself, since both cases involve agents in positions in which they have a particular duty to be careful; it is part of the security official's job to avoid the risk of terrorist incidents and D promised C to keep their property safe. Similarly, the agents in 'Lottery' and 'Walker' bear responsibility for their misfortunes, although they have no particular duty to avoid them.

In all the cases apart from 'Mugger', then, responsibility for risk seems to entail at least some responsibility for the harm. 'Mugger' is distinguished from 'Lottery' and 'Walker' by involving victimisation by another rather than a chance occurrence leaving them worse off. 'Terrorist' and 'Trusted' also involve victimisation, but again, 'Mugger' is distinct from them. This time, the difference comes from the lack of a special duty to avoid harm of the given kind. What these cases tell us, then, is that responsibility for harm entails no further responsibility where one both is victimised and has no specific duty to avoid the victimisation. Where one of these is lacking, the harmed party is responsible for not only the risk, but also the harm itself. What we need is some philosophical explanation for this.

The principle stated above does not hold the victim responsible, but neither does it absolve them – it states a sufficient, not a necessary, condition for responsibility. To get an answer to our original question, we will have to examine what it means to have a special duty in the sense of that principle, and what it implies about the responsibility of other parties involved. This will be the task of the next section.

Part 2: The Commitment Response

I turn now to my preferred solution to the Responsibility Problem, which I call the commitment response. In the course of explaining and justifying it, it will become clear what constitutes a special duty in the sense relevant in the above cases and why it plays the required responsibility-giving role.

Commitments and the Transfer of Responsibility

According to the commitment response, the relevant difference between 'Mugger' and the other two main cases is that it involves a broken commitment on the part of the perpetrator. 'Lottery' and 'Walker' do not involve commitments, so responsibility remains with the risk-taker. Commitments, it will be argued, have the power to transfer responsibility from one party to another.

We start by observing that, in addition to our ordinary moral duties, we can incur duties by making commitments. For instance, when we promise to ϕ, we thereby incur a duty to ϕ which we may not have previously had. Although promises are perhaps the clearest cases of commitments, there are others, which can be implicit or tacit. There are commitments that we can incur through our behaviour, by remaining silent, or by taking on some role. An important feature of a commitment is that it is made *to* somebody. One does not simply make a commitment that one will ϕ; one commits *to another* that one will ϕ. Furthermore, if one has made a commitment, one cannot simply withdraw it at will. It is not like consent, wherein one may decide at any point that one will not after all do or permit to be done what one has previously consented to. To be released from a commitment, one typically requires permission from the one to whom the commitment was made.[106]

We here make use of the Hohfeldian framework of rights. The important features for our purposes are claims, duties and powers.[107] Claims and duties are correlatives: an agent B has a claim against another agent A that they perform some action ϕ just in case A owes to B a duty that they ϕ. A power is the ability to alter the 'Hohfeldian incidents' of oneself or another, for instance, by granting someone else a claim against oneself (thereby incurring a duty to them). (Hohfeld 1978, 35-8; 50-60) What I wish to highlight is that it is within one's own power to make a commitment, but it is not within one's power, nor in any third party's, to get oneself out of it. This suggests the following principle regarding commitments:

> If A commits to B to ϕ, then A owes a duty to B to ϕ and B alone has the power to release A from their duty to ϕ.

It is worth noting that there may be further release conditions. If, for instance, it turns out that ϕ is an impossible task, then, insofar as 'ought' implies 'can', A has no obligation to ϕ. Or if A suffers an injury, or becomes unavoidably engaged elsewhere, so that ϕing becomes much more difficult, then they may plausibly be released from the commitment. However, this is

[106] There are different types of commitment; we might be committed to a person in the sense of being devoted to them, or we might be committed to an idea in the sense of it being entailed by our other beliefs. My concern here is with commitments to do (or to refrain from doing) certain things. For more on the nature of commitments, their various kinds, and possible philosophical implications, see Chang (2013), Shpall (2013, 726-35), Shpall (2014), and Walsh (2017).

[107] There are five more 'Hohfeldian incidents', but they are not directly relevant to the present discussion (Hohfeld 1978, 36).

compatible with the above principle, since such circumstantial release conditions are not due to the exercise of anyone's power regarding A's duty. The point is that no agent other than B can decide that A is to be released from their duty, including A themselves.

The next important point is that linking duty to responsibility. A very simple link is suggested by the following conditional:

> If A has a duty to ϕ and A fails to ϕ, then A is responsible for the foreseeable consequences of their not ϕing.

Although this claim is very plausible, it might also seem to require a plethora of qualifications. What is the limit on consequences that count as foreseeable? Surely there are some valid excuses that A might be able to make for failing in their duty – what kinds of considerations might these be? Although it is worth acknowledging these questions, as well as the consequent fact that this principle, as stated, is an oversimplification, I will not delve into them here. The simplified version will suffice for our purposes, since it will be primarily applied to cases in which it is clear that there is no excuse and the relevant consequences are foreseeable on any plausible view of that notion.

These two principles, connecting commitment to duty and duty to responsibility, jointly entail that commitments can transfer responsibility. By making a commitment and failing to follow through, one takes on responsibility for the consequences.

The picture is completed by a corresponding argument concerning the rights of the one to whom the commitment is made. Consider the following:

> If A commits to B to ϕ, then B has a claim against A that they ϕ and B alone has the power to waive this claim.

This follows, in the Hohfeldian framework, from the principle of commitments already stated. Indeed, the two principles are equivalent. The correlate of A owing a duty to B is B having a claim against A. If A fails to meet their commitment, then it is B who is wronged; B has a right against A that is violated.

Furthermore, it seems that we can again make a link with responsibility. For if some harm obtains because one's rights are violated, then one is not usually responsible for that harm. An exception seems to be if the harm is also caused by a failure on one's own part to fulfil a duty. For example, the security official bears responsibility for the harm that befalls them, whereas a bystander who is also injured does not. The harm is caused by the attackers, of course, but there is also a relevant sense in which it is caused by the failure on the official's part to prevent it. This suggests the following:

> If B has a claim against A that they ϕ, A fails to ϕ and B lacks a duty to ϕ, then B is not responsible for the consequences of ϕ not being done.

These principles jointly entail that another's commitment can absolve one of responsibility, at least in the absence of any particular duty to do oneself what the other has committed to doing. Thus, commitments transfer responsibility: they make the commitment-maker responsible for the consequences of what they fail to do and absolve the one to whom the commitment is made of such responsibility. For instance, recall the example in the introduction regarding your promise to pick up my medication. These principles show, firstly, that if you fail to follow

through, then you are responsible for the consequences – in this case, a worsening of my medical condition – and secondly, that I am not responsible for those consequences. This holds even if I could have easily done it myself. I may be criticisable for being so foolish as to leave such an important task to someone else, just as the victim in 'Mugger' might be considered foolish for taking that route, but ultimately, it is not my fault.

This transference of responsibility can help us to explain the salient differences among the cases we have been considering. All put together, the above principles entail the following:

(a) If A commits to B to ϕ and ϕ is not done, then A is responsible for the foreseeable consequences of not ϕing.
(b) If A commits to B to ϕ, ϕ is not done and B lacks a duty to ϕ, then B is not responsible for the consequences of not ϕing.

Neither (a) nor (b) is applicable to 'Lottery' or 'Walker', since there is no commitment involved in those cases. However, that does not mean that we must say that no one bears responsibility. (a) is a sufficient condition for being responsible for certain consequences; (b) is a sufficient condition for being absolved of responsibility. Neither is necessary, so where neither applies, we can still determine responsibility (or lack thereof) by other means. In these cases, as mentioned earlier, the general idea that the risk-taker bears responsibility for the risk's manifesting seems applicable. This is the assumption that we have been maintaining for typical cases and to which 'Mugger' is an exception we seek to explain.

In the case of 'Terrorist', we can say that both the terrorists and the security official are responsible, though in different ways. As we have mentioned, the security official has a duty to prevent just these kinds of harms from occurring. This duty is derived from a standing commitment that they have as a part of their job. To whom is this commitment made? It might be to their superiors, or it could be to the public in general. Either way, unlike the bystander, they have committed to protecting the public from terrorist incidents. Therefore, by (a), they bear responsibility for the harm to themselves and to the other victim – they committed to preventing attacks and the attack was not prevented, so they are responsible for the consequent (and quite foreseeable) harms. The terrorists, meanwhile, are responsible in the ordinary way, as indicated earlier by SCR: they voluntarily brought about a harmful event, so are responsible for it.

It is no deficiency to imply that multiple parties bear responsibility for the same event, since it is quite common in our ethical discourse to suggest that different people are responsible in different ways and for them to be treated differently accordingly. The penalty for the official who should have prevented the attack will not be the same as that for the terrorists who instigated it. The fact remains, though, that both parties are responsible – both will be called to account and are likely to face some form of punishment for the attack if they are caught and the relevant facts are known. We are dealing here merely with whether a party is responsible; the degree or kind of responsibility is another matter.

Similar considerations apply in 'Trusted'. Despite being the victim of a mugging, D bears responsibility. Since they committed to keeping their friend's property safe and said property was not kept safe, they are responsible for it. When they take the unsafe route, they go back on the commitment. Again, there is room for different kinds of responsibility, or at least different sorts of reasonable reactions on the part of the friend whose item was stolen, depending on

whether they were forgetful, or careless, or malicious in taking the more dangerous path. The mugger, like the terrorists, will also be responsible, but in the sense of SCR, which does not depend on commitment.

The Mugger's Commitment

What, then, are we to say of 'Mugger', the only case in which the one who suffers the harm seems to not bear any responsibility? It is worth mentioning again that the question here is that of what absolves the victim, not of why the perpetrator is responsible. That, as in the two cases just considered, is sufficiently explained by SCR.

The key factor that removes all responsibility from the victim in 'Mugger' is, I shall argue, a commitment on the part of the mugger to not harm them. We have already seen how commitments can transfer responsibility away from a harmed party. Now, it may seem implausible that the mugger has such a commitment. They certainly have not promised anything to that effect. Nor is prevention of harmful acts, including their own, part of their job, as is the case with the security official in 'Terrorist'. Surely their wrong consists simply in their violation of the moral requirement not to steal and their responsibility is accounted for by SCR. There is no need to suggest that they breach a commitment as well. My contention, however, is that there is such a commitment and it derives simply from their membership of society. All members of a society have a standing commitment to one another not to harm them unprovoked. This can be made clear by examples of exceptions that are occasionally made to these commitments.[108] Consider the following case.

> **Boxing:** E and F are boxers fighting in a ring. At the end of a round, E's coach tells E to pull his punches and stop hitting F in the head. E's coach can see that F is at risk of serious injury if the fight continues as it has done thus far and is keen to prevent it if possible. E ignores the advice and continues attacking his opponent as he did before. F ends up with a serious head injury as a result.[109]

People usually have claims against one another not to be harmed. This should be intuitively clear, however it is to be explained. By entering a ring with an opponent, boxers to some extent waive this claim. The waiver is of restricted scope; it applies only to the opponent, so does not permit others – the referee or audience members, say – to join in. It also only applies to certain kinds of attack; there are rules restricting what boxers may do to one another. Nevertheless, there is a morally significant difference between hitting someone in a boxing ring and hitting someone elsewhere.

In 'Boxing', F voluntarily stepped into the ring to participate in the fight knowing that, even within the restricted scope of what is permitted, he could get seriously hurt. His doing so constituted waiving his claim against E that E not harm him. Even so, E's actions are plausibly immoral. Knowing that serious damage could be done, he continued to attack his opponent in such a way as to induce injury. Given that F had waived his claim against E that E refrain from

[108] This is roughly analogous to the example of the actor speaking a line on stage given in Chapter 3. The fact that it would be incorrect to assess their speech in terms of honesty highlights the fact that there is a general commitment to speak honestly.

[109] This is based on the real-life case of Nick Blackwell and Chris Eubank (Nick Blackwell: Chris Eubank Sr says he would have stopped fight, 2016).

attacking in certain ways, and that E attacked only in those ways, how can we account for the immorality of E's actions?

Before answering, let us take another case, this one from Katherine Hawley's paper, 'Trust, Distrust, and Commitment'.

> **Wild West:** Suppose we are in the Wild West. In town, there is an uneasy truce, a semblance of law and order. Out in the desert, as everyone knows, there are no holds barred; you take your life into your hands if you venture there. You and I meet by chance in the desert, and you see that I am armed (of course). Before we exchange words, you may try to predict whether or not I will let you live.

(Hawley, 2014, 18-9)

In this case, there is a difference highlighted between the situation in the town and the situation in the desert. Hawley argues (2014, 19), correctly in my view, that whereas in town everyone has a commitment not to shoot anyone else, no such commitment applies in the desert. This is not to say that it is morally permissible to shoot someone without provocation outside the town's borders. Murder is still wrong there, but has a different moral character to it, in that it is more to be expected. The ordinary norms and values of living in a society are put aside when you venture into the desert; you know that you are taking your life into your hands.

'Wild West' highlights the difference between living within and without a society by putting the two side by side. 'Boxing' also shows a case in which ordinary norms are suspended, yet doing certain things is still wrong. My contention, which the two cases make plausible, is that there is a standing commitment on the part of all members of a society to not harm one another unprovoked. This is arguably a part of what it means to be in a society, though it of course falls far short of a full definition. Giving such a definition is not our purpose here. However, this commitment to non-harm seems a reasonable minimal requirement. Sometimes aspects of it can be set aside, as in 'Boxing', and sometimes it can be set aside wholly, as in 'Wild West'.

Perhaps it will be argued that these cases do not present morally relevant differences from ordinary cases of punching or shooting. Alternatively, it may be thought that there is a difference, but it is to be explained by general moral norms rather than by commitments. Neither seems plausible upon reflection, however. For the first, we would have to say that the usual moral requirement of non-harm holds good in the boxing ring, in which case there is no moral difference between a boxing match and an unregulated brawl. But this cannot be the case; whatever ethical concerns there may be around the sport of boxing, surely there is a difference between a boxing match and straightforward assault. Although it is less clear-cut, something similar holds for the desert in 'Wild West'; it is generally understood that those who venture there are surrendering the protections of ordinary society.

For the second, we would have to say that the relevant moral requirements do not hold. This may at first seem plausible for 'Boxing'; certain ordinary moral norms are put aside, including the requirement to not punch the other person in the head. Maybe E does nothing wrong in subjecting F to severe injury. However, this seems wrongheaded. The aim in boxing is not to cause harm to one's opponent, but to win the match, for which harming one's opponent will

often be necessary. This is an important ethical difference.[110] In fact, the situation is designed to avoid, rather than facilitate, serious harm to the combatants – there are rules, a referee to enforce them, and a ringside doctor. Getting hurt is expected, but if someone sustains a long-term debilitating injury, then something has gone wrong. Given this, it seems plausible to draw a distinction between hitting someone in order to win, from which they may, but hopefully will not, become injured, and hitting someone in order to win while knowing that doing so will cause serious injury. F may have voluntarily entered the ring and E may keep within the rules, but knowing the damage he is likely to cause still makes a moral difference to E's position. For 'Wild West', the objector might try to keep the point modest, saying that the moral injunction against murder still holds, but that it may be okay to injure others. But this also is incorrect. Unprovoked violence against others is generally wrong and merely being in the desert does not change this.

The idea of commitments being involved as well as moral norms satisfies the intuition that E and the 'Wild West' shooter still act immorally, yet the situations lend their actions a different moral character.

So, being a part of a society entails a commitment to not harming other members thereof. It might be wondered how the commitment is made and how, if at all, an individual born into a society can opt out of it. This is no doubt an important and interesting question of political philosophy, but it is not one that I will delve into here. It suffices for our purposes to merely point out that the commitment is generally present, as cases like 'Wild West' and 'Boxing' highlight.[111] Similarly, I will leave the scope of the term 'society' ambiguous – whether it is properly applied to a certain town or city, a whole country, or even some global community. I take it to be sufficiently well-understood for current purposes as being distinct in certain ways from a collection of people who happen to live near each other.

Like any commitment, it is not in one's own power to get out of it. That rests with those to whom it is made. It is not entirely clear to whom the commitment to non-harm is made, but whoever it is, one cannot simply decide that one no longer has the commitment. Perhaps, if the mugger's victim had given them special permission to harm them (for whatever reason), this would release them from the commitment on that particular occasion, as E and F temporarily release one another from their commitments. But of course, no such permission is given.

If the perpetrator is bound by a commitment to not harm others without provocation, then the solution to the Responsibility Problem becomes clear. The perpetrator has committed to not harm the victim (as they commit to not harm anybody) and so, by (a), is responsible for the harm that they cause in breaking that commitment. Thus, their responsibility is overdetermined, since it also follows from SCR. The more interesting result, however, follows from (b). Since the perpetrator broke their commitment to non-harm and the victim has no particular duty to prevent such harm, the victim is not responsible for the consequences of that broken

[110] The difference between intentionally causing harm and causing harm as a necessary means to a certain end is commonly recognised. Foot (1967) provides several examples.

[111] In a similar vein, we might wonder what constitutes acceptance or uptake. In order to be binding, a promise must be accepted, or at least not rejected, by its recipient – see, for instance, Thomson (1990, 296-8). It seems most likely to me that, since this commitment is made tacitly, it is also accepted tacitly. The default position is accepting the commitment to non-harm of everyone else in the society (doing so is arguably another prerequisite for being part of said society). In order to reject it, one must either do so explicitly, or take some other action that precludes the acceptance of the commitment.

commitment – that is, they are not responsible for the harm that they incur when they are mugged. This holds despite their having voluntarily taken the unsafe route, thereby risking themselves and their property.

No doubt the mugger would deny the existence of any such commitment. They would point out that they did not promise, or otherwise explicitly agree, that they would not harm anyone else. It was not up to them that they found themselves part of a wider society with the supposed accompanying commitments, so they should not be held to them. An argument to this effect goes as follows:

1. Commitments can only be made voluntarily.
2. Membership of this society is not voluntary.
3. Therefore, membership of this society cannot constitute or entail a commitment.

The problem with this argument is that it is inconsistent in its use of 'voluntary'. In one sense, the first premise is true. Commitments cannot be forced on us; they are obligations that we have some choice about incurring. However, they are not fully under our control. Once incurred, we cannot simply opt out. Furthermore, we may not even have full control over whether and when we acquire a commitment. They can be acquired tacitly or implicitly. As Hawley points out, 'in some circumstances we acquire commitments simply by allowing others to continue to rely upon us, or by allowing others to think that we have commitments' (2014, 14). It is, sometimes inconveniently, not the case that we make a commitment just when we explicitly declare it, or even that we always know when we have committed. Given this ambiguity, commitments can be understood as voluntary only in a very broad sense. We can find ways around them – we can be alert to commitment-incurring circumstances and avoid them. For instance, we might refuse another's offer of help lest we thereby become committed to help them later. We might make a point of telling people not to rely on us, or make it clear that we are not undertaking commitments. Phrases like 'I might', 'maybe', 'if I have time', 'I suppose that' and so on would pepper our conversations as we try to avoid committing ourselves to anything. Few of us, however, actually live in this fashion. Barely anyone is entirely uncommitted.[112] So, commitments are voluntary in the sense that we can technically avoid them, but in practice we may not have full control or knowledge of what we have committed to.

The second premise is true on a narrower sense of 'voluntary'. To become a member of a society, one does not usually need to explicitly declare or request it. Sometimes one might; successfully applying for citizenship of a country is presumably such an explicit joining of a society. But more often this is not the case. Where voluntariness is understood as requiring some deliberate or explicit act, then, membership of a society is not always voluntary, and we can suppose that the mugger is one of the majority of people who did not deliberately choose to be part of the society that they are in.

However, if we are consistent with the use of 'voluntary', then at least one of the premises is false. On the narrow sense, on which the second premise is true, the first is false; commitments do not need to be made deliberately or explicitly. Similarly, it seems that membership of society

[112] See Hawley (2019, 78-9). This is not to contradict the claim, made in Chapter 3, that everyone is uncommitted with regard to something. We all have commitments of some kinds, but we also all lack commitments of some (other) kinds.

is voluntary on the broader sense of the word. It may be difficult to untangle oneself, but, in principle, it is possible to opt out of society altogether, becoming something of a hermit. As mentioned, I will not go into precisely what conditions must be fulfilled in order to achieve this, but all that is required for this point to stand is that it is possible to do it. This opens up conceptual space to say that, in the sense that commitments are voluntary, membership of society is also voluntary. Thus, one way or the other, the argument above fails, so the mugger indeed has a commitment to non-harm.

Now, if this commitment on the part of the perpetrator absolves the victim in 'Mugger', why does it not absolve the person who was entrusted with their friend's precious object? After all, the cases are nearly identical. The perpetrator in that case presumably has a commitment to non-harm which they are breaking by mugging the person who comes down the path.

Here is where the 'B lacks a duty to ϕ' clause of (b) becomes relevant. Since they owe it to their friend to look after their precious item, having made a commitment to that effect and been trusted with its safekeeping, they have a duty to not let it be stolen. They took an unnecessary risk, so do not have an excuse. As emphasised above, this in no way absolves the mugger, who still bears full responsibility for their actions. But their commitment does not this time transfer the responsibility away from the victim.

This, then, is the difference between 'Mugger' and the other cases: a valid commitment to non-harm is broken by the perpetrator, which, in the absence of a duty to prevent such harms, transfers responsibility for the harm away from the victim. In 'Walker', there is no other agent involved, so no one to break a commitment. In 'Lottery', the organisers clearly do not have a commitment to the players that they will not suffer a loss, so cannot be breaking it.

Conclusion

We have seen how the case of 'Mugger' differs from the standard cases of risking a harm and that harm occurring, as exemplified by 'Walker' and 'Lottery'. Although the victim did genuinely take the risk, which usually entails responsibility for the consequent harm, they are entirely absolved in this case. This transfer of responsibility is due to the perpetrator's commitment to non-harm, which is not a factor in the other two cases.

It has also been shown that there can be cases, even cases very similar to 'Mugger', in which the victim bears some measure of responsibility. These can be accounted for by their duty to avoid or prevent the harms in question, though this does not mean that the perpetrator is any less responsible.

Such cases do provide further questions. What is the difference between the responsibility of the perpetrator and that of the victim – or anyone else with a duty to stop them – for the harm that occurs? Do terrorists, who may not be members of the society they target, break a commitment in causing harm, or do they lack the commitment to non-harm? We might also ask about cases like 'Boxing'. Having waived their claim against being punched, might a recipient of harm bear some form of responsibility for being injured when they have voluntarily entered the boxing ring? What is the moral difference between punching someone normally and doing so in the ring? But these matters can be set aside for now. We have explained the

impact that a commitment has on responsibility and in so doing, have exposed the error in blaming the victim.

Furthermore, we have embedded the idea of commitment, which is central to our understanding of trust and trustworthiness, within the broader ethical framework of moral responsibility, thus connecting trust and trustworthiness to responsibility. This connection between important ethical concepts will likely be a fruitful subject of inquiry in future work.

Conclusion

In this thesis, I have tried to capture the nature of trust and trustworthiness and the norms that govern the practice of trusting. Issues of trust have implications for both ethics and epistemology, affecting both how we should act and what we can know. Trust is also an important element in many kinds of cooperation, enabling us to work together more efficiently, and also plays a role in many personal relationships. By the same token, when trust is misplaced, it can have seriously negative consequences, so it is something with which we must take care.

The aim of this thesis was to provide a theory of trust and trustworthiness, and to further our understanding of when it is rational to trust. This, I believe, has now been achieved. By carefully weighing various ideas against one another and learning from those which fail or succeed only partially, I have come to a relatively simple conclusion about what trust is and when it is justified, which can nonetheless encompass the complexities of the diverse contexts in which people can trust each other.

The overall theories of trust and trustworthiness advocated herein may be stated as follows, where A and B are persons and ϕ is a behaviour:

> A trusts B to ϕ if and only if A relies on B to be trustworthy for A with respect to ϕ.
>
> B is trustworthy for A with respect to ϕ if and only if B has committed to A that they ϕ and B ϕs.

This entails that to trust someone is to rely on them to fulfil a commitment. Or, to put it more formally:

> A trusts B to ϕ if and only if A relies on B fulfilling their commitment to A that they ϕ.

In the absence of such a commitment, the trust is misplaced, since its content is vacuous.

The reasons which justify trust are both practical and epistemic, so long as they address the content of trust – the other's trustworthiness. The reasons which support A trusting B to ϕ may therefore be expressed as follows:

> Epistemic: Factors which increase the likelihood, from A's perspective, of B fulfilling their commitment to A that they ϕ.
>
> Practical: Factors which make it more beneficial to A that B fulfils their commitment to A that they ϕ.

Evidence that B will ϕ alone therefore does not give A reason to trust; it must be evidence of B's trustworthiness. Similarly, trust being valuable or convenient does not give A reason to trust B; practical reasons to trust must show how it would benefit A for B to prove trustworthy. Again, trust without commitment is misplaced, so cannot be rationally justified.

Although I think that what I have presented here furthers philosophical understanding of trust, we are by no means finished. There are numerous areas of potential future research to consider and I shall end by briefly discussing some of them. In so doing, I hope also to demonstrate, one last time, the value that theorising about trust might have for philosophy more broadly.

I believe that there is more to be said on the distinction between trusting someone to do something (act trust) and trusting what they say (testimony trust). I have taken the view that testimony trust is a specific case of act trust: it is trusting another to speak honestly. However, testimony trust may be more substantially different. It is essentially backward-looking, whereas act trust tends to be forward-looking; testimony trust is concerned with what someone has done, whereas act trust is about what they will do. Furthermore, testimony trust is concerned with a specific feature of what someone has done, rather than with the actual action they have performed. When we deliberate about trusting someone's testimony that p, that they have uttered p is not in doubt; what we are querying is a specific feature of the utterance – whether it is true. In contrast, when we deliberate about trusting someone to ϕ, we are not concerned with any specific features of ϕ, but only with whether they will ϕ. In general, the relationship between assertions and commitments deserves greater scrutiny and the philosophy of trust may provide a fruitful angle of approach.

There are also a number of philosophical applications that could be made with a theory of trust. Trust is closely connected to assertions and promises. It would be interesting to explore how a theory of trust might shed light on these more well-researched topics. It also seems to me that trust deserves a more prominent place in broader ethical and epistemological conversations than it is currently afforded. It certainly plays a role in our lives as moral agents and as knowers, so perhaps it ought to play a more significant role in moral and epistemic theorising. Then there is the matter of rational action. I have stressed that trust can be rational and given the kinds of reasons that can count in favour of it. This being the case, maybe trust needs to be given a place in discussions of social dilemmas and game theory; cases in which it is not clear why agents do or should cooperate.

Finally, there are practical applications. I have said that trust is relevant in a wide variety of contexts, but I have focused my attention on a fairly narrow set of cases: those in which one agent might trust one other agent with some particular thing. But trust can also be considered in a political context. Under what conditions do and should a people trust their leaders, or a set of leaders their people? To what extent does it make sense to talk of states or other political entities trusting one another? Extending the idea of trust in a group context, we might wonder what philosophy of trust can tell us about large-scale cooperation more generally. To what extent can we trust corporations? What about the many experts whom we do not personally know and whose research we cannot feasibly check? Modern technology extends and complicates matters. What is the impact of the Internet and social media on our trusting relations? Is it reasonable – and does it even make sense – to trust in and via digital technology? More conventionally, trust has clear implications for various fields of practical ethics. Those

who study friendship, medical ethics, and the ethics of caregiving will need to take issues of trust and trustworthiness into account.

Perhaps the most significant question we can ask in the practical domain is this: how do we build and maintain good trusting relations? I hope that it is clear by now that it is generally in all of our interests to be able to do so, but how to go about it is another matter. Connected to this is the question of why someone should be trustworthy. What motivates trustworthy behaviour? Why should I be the kind of person whom others are willing to trust? Perhaps answering this would be a starting-point to generating a greater degree of trust in all aspects of life. After all, people are only likely to place their trust when they are confident that others are worthy of it.

As the above remarks testify, there is no shortage of work still to be done in this area. Nevertheless, it is my hope that this thesis represents some modest advance in the philosophy of trust.

Bibliography

Alexander, L. (1996) The Moral Magic of Consent (II). *Legal Theory*, 2(3), 165-174

Archard, D. (1997) 'A Nod's as Good as a Wink': Consent, Convention, and Reasonable Belief. *Legal Theory*, 3(3), 273-290

Baier, A. (1986) Trust and Antitrust. *Ethics*, 96(2), 231-260

Baker, J. (1987) Trust and Rationality. *Pacific Philosophical Quarterly*, 68(1), 1-13

Bennett, J. (1980) Accountability, in van Straaten, Z. (ed.) *Philosophical Subjects: Essays Presented to P.F. Strawson*. Oxford: Clarendon Press, 14-47

Bennett, M. (2021) Demoralizing Trust. *Ethics*, 131(3), 511-538

Berker, S. (2018) A Combinatorial Argument against Practical Reasons for Belief. *Analytic Philosophy*, 59(4), 427-470

Boonin-Vail, D. (1997) A Defense of "A Defense of Abortion": On the Responsibility Objection to Thomson's Argument. *Ethics*, 107(2), 286-313

Bratman, M. (1992) Practical Reasoning and Acceptance in a Context. *Mind* 101(401), 1-15

Brison, S. (1993) Surviving Sexual Violence: A Philosophical Perspective. *Journal of Social Philosophy*, 24(1), 5-22

Chang, R. (2013) Commitments, Reasons, and the Will. *Oxford Studies in Metaethics*, 8, 74-113

Coleman, J. (1990) *Foundations of Social Theory*. Cambridge, Massachusetts: Belknap Press of Harvard University Press

Conee, E. and Sider, T. (2014) *Riddles of Existence: A Guided Tour of Metaphysics*. 2nd ed. Oxford: Oxford University Press

Domenicucci, J. and Holton, R. (2017) Trust as a Two-place Relation, in Faulkner, P. and Simpson, T. (eds.) *The Philosophy of Trust*. 2nd ed. New York: Oxford University Press, 149-160

Faulkner, P. (2017) The Problem of Trust, in Faulkner, P. and Simpson, T. (eds.) *The Philosophy of Trust*. 2nd ed. New York: Oxford University Press, 109-128

Finlay, S. (2014) *Confusion of Tongues: A Theory of Normative Language*. New York: Oxford University Press

Fischer, J. and Ravizza, M. (1993) Responsibility for Consequences, in Fischer, J. and Ravizza, M. (eds.) *Perspectives on Moral Responsibility*. 1st ed. Ithaca, New York: Cornell University Press, 322-347

Fischer, J. and Ravizza, M. (1998) *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press

Foot, P. (1967) The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review*, 5, 5-15

Frege, G. (1948) Sense and Reference. *The Philosophical Review*, 57(3), 209-230

Fricker, M. (2007) *Epistemic Injustice: Power and the Ethics of Knowing*. New York: Oxford University Press

Frost-Arnold, K. (2014) The Cognitive Attitude of Rational Trust. *Synthese*, 191(9), 1957-1974

Goldman, A. (1976) Discrimination and Perceptual Knowledge. *The Journal of Philosophy*, 73(20), 771-791

Grubb, A. and Turner, E. (2012) Attribution of Blame in Rape Cases: A Review of the Impact of Rape Myth Acceptance, Gender Role Conformity and Substance use on Victim Blaming. *Aggression and Violent Behavior*, 17(5), 443-452

Hardin, R. (1996) Trustworthiness. *Ethics*, 107(1), 26-42

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

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

Hawley, K. (2019) *How to be Trustworthy*. Oxford: Oxford University Press

Hayes, R., Lorenz, K., and Bell, K. Victim Blaming Others: Rape Myth Acceptance and the Just World Belief. *Feminist Criminology*, 8(3), 202-220

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

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

Hohfeld, W. (1978) *Fundamental Legal Conceptions as Applied in Judicial Reasoning*. 3rd ed. Westport: Greenwood Press

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

Hurd, H. (1996) The Moral Magic of Consent. *Legal Theory*, 2(2), 121-146

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

Jones, K. (2004) Trust and Terror. In: P. DesAutels and M.U. Walker, eds., *Moral Psychology: Feminist Ethics and Social Theory*. New York: Rowman & Littlefield, 3-18

Jones, K. (2012) Trustworthiness. *Ethics*, 123(1), 61-85

Jones, K. (2017) ‘But I was Counting On You!’, in Faulkner, P. and Simpson, T. (eds.) *The Philosophy of Trust*. 2nd ed. New York: Oxford University Press, 90-108

Kavka, G. (1983) The Toxin Puzzle. *Analysis*, 43(1), 33-36

Keller, S. (2004) Friendship and Belief. *Philosophical Papers*, 33(3), 329-351

Kenyon, T. (2010) Assertion and Capitulation. *Pacific Philosophical Quarterly*, 91(3), 352-368.

Kosfeld, M., Heinrichs, M., Zak, P., Fischbacher, U. and Fehr, E. (2005) Oxytocin Increases Trust in Humans. *Nature*, 435(7042), 673-676

Lackey, J. (2008), *Learning From Words*. Oxford: Oxford University Press

Liberto, H. (2021) Coercion, Consent, and the Mechanistic Question. *Ethics*, 131(2), 210-245

Marušić, B. (2015) *Evidence and Agency: Norms of Belief for Promising and Resolving*. New York: Oxford University Press

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

McKenna, M. (2012) *Conversation and Responsibility*. Oxford: Oxford University Press

McMyler, B. (2013) The Epistemic Significance of Address. *Synthese*, 190(6), 1059-1078

Moran, R. (2005) Getting Told and Being Believed. *Philosopher’s Imprint*, 5(5), 1-29

Nick Blackwell: Chris Eubank Sr says he would have stopped fight (2016) (online) Available at: https://www.bbc.co.uk/sport/boxing/35915479 (Accessed 12th September 2022)

Owens, D. (2006) Testimony and Assertion. *Philosophical Studies*, 130(1), 105-129

Parfit, D. (2001) Rationality and Reasons, in Egonsson, D., Josefsson, J., Petersson, B. and Rønnow-Rasmussen, T. (eds.) *Exploring Practical Philosophy: From Action to Values*. Aldershot: Ashgate, 17-39

Rawls, J. (1971) *A Theory of Justice*. Cambridge: Harvard University Press

Raz, J. (1981) Authority and Consent. *Virginia Law Review*, 67(1), 103-131

Raz, J. (1988) *The Morality of Freedom*. 2nd ed. Oxford: Oxford University Press

Raz, J. (1999) *Practical Reason and Norms*. 3rd ed. Oxford: Oxford University Press

Reisner, A. (2018) Pragmatic Reasons for Belief. In: D. Star, ed., *The Oxford Handbook of Reasons and Normativity*, 1st ed. New York: Oxford University Press, 705-728

Rinard, S. (2019) Believing for Practical Reasons. *Noûs*, 53(4), 763-784

Scanlon, T. (1990) Promises and Practices. *Philosophy & Public Affairs*, 19(3), 199-226

Scanlon, T. (1998) *What We Owe to Each Other*. Cambridge: Harvard University Press

Schroeder, M. (2010) Value and the Right Kind of Reason. *Oxford Studies in Metaethics*, 5, 25-55

Shpall, S. (2013) Wide and Narrow Scope. *Philosophical Studies*, 163(3), 717-736

Shpall, S. (2014) Moral and Rational Commitment. *Philosophy and Phenomenological Research*, 88(1), 146-172

Shah, Nishi (2006) A New Argument for Evidentialism. *Philosophical Quarterly*, 56(225), 481-498

Simpson, T. (2017) Trust and Evidence, in Faulkner, P. and Simpson, T. (eds.) *The Philosophy of Trust*. 2nd ed. New York: Oxford University Press, 177-194

Strawson, P. F. (1974) *Freedom and Resentment and Other Essays*. London: Methuen

Stroud, S. (2006) Epistemic Partiality in Friendship. *Ethics*, 116(3), 498-524

Thomson, J. (1986) Imposing Risks, in Parent, W. (ed.) *Rights, Restitution, and Risk: Essays in Moral Theory*. Cambridge, Massachusetts; London: Harvard University Press, 173-192

Thomson, J. (1990) *The Realm of Rights*. Cambridge, Massachusetts: Harvard University Press

Wallace, J. (1994) *Responsibility and the Moral Sentiments*. Cambridge, Massachusetts: Harvard University Press

Walsh, J. (2017) Commitment and Partialism in the Ethics of Care. *Hypatia*, 32(4), 817-832

Watson, G. (1987) Responsibility and the Limits of Evil: Variations on a Strawsonian Theme, in Schoeman, F. (ed.) *Responsibility, Character, and the Emotions*. Cambridge: Cambridge University Press. 256-286

Way, J. (2012) Transmission and the Wrong Kind of Reason. *Ethics*, 122(3), 489-515

Williamson, T. (1996) Knowing and Asserting. *Philosophical Review* 105(4), 489-523