1
The Value of Trust

Trust is central to our lives, at both an individual and a societal level. We trust our spouses, our friends, our children, and our work colleagues, though to varying degrees and in varying circumstances. We hope to be trusted in turn. We may restore trust by forgiving others when they betray us. Some people trust their schoolteachers, doctors, the local butcher, their parents, or the neighbour. Few trust bankers now, and journalists are feeling the pinch too. Attitudes towards politicians are a complex mix of trust and distrust, sometimes based on reasons, and at other times based on cynicism or driven by strong passions. Every aspect of social life may see interpersonal relationships based on trust, and in a significant enough proportion of cases, trust actually is the basis for the relationship. Not only may social relationships be based on trust, but it seems to be good when they are. Things are going well when there is a lot of trust around, and they are going poorly when trust is in short supply. The lack of trust is invariably a source of lament, and when times are tough, or a country, organisation, or community is in crisis, the problem is often articulated in terms of trust. For those who would understand human interaction, then, questions of trust are pregnant with importance and implication.

This book is an attempt to answer some philosophical questions that trust raises. My focus is on trust as it occurs between two people, and more specifically, on the reasons which justify that trust. You find yourself in a situation wondering whether you should trust another person. How should you decide this? What reasons should you look for, to determine whether or not it is appropriate for you to trust this person? What counts in favour of trust or withheld trust? If I need to defend my trust, whether to myself or another, what should I say? In short: why should I trust?

Concern for this question arises because of a quotidian problem. It is often unclear whether I should trust a particular person, or whom I should trust, given a choice between different candidates for trust. This decision may be difficult because I do not know someone well enough to assess whether they are trustworthy. But the deepest problems of trust arise when I do know them well enough to make that assessment, and yet there are countervailing considerations which suggest, contrary to my assessment, that I ought either to trust them, or to withhold my trust. A friend claims she is the victim of an unjust accusation, but there is strong evidence supporting the charge; should you trust her? H. J. N. Horsburgh termed these scenarios the 'dilemmas of trust' (1960: 343; echoed by Govier

Trust: A Philosophical Study. Thomas W. Simpson, Oxford University Press. © Thomas W. Simpson 2023.
DOI: 10.1093/oso/9780198855866.003.0001

1998). That they exist forces me, as a socially embedded person who must decide whom to trust and when, to confront the question of why I should trust.

My focus on the normative reasons for interpersonal trust delimits the scope of the enquiry in some important ways. For one, I am not primarily interested here in the political implications of trust. These arise given prevailing patterns of trust and distrust across a population, and I address them only briefly in the penultimate chapter. For another, and as befits the tools of philosophy, the enquiry is also explicitly a normative one, not descriptive. Although the force of the conclusions reached by the different kinds of enquiry differ, with normative enquiries justifying trust and descriptive enquiries explaining its causes, the two projects are not wholly orthogonal. Reasons which would justify trust must plausibly be available to humans as we are, in the kinds of societies in which we live; conversely, if humans are rational agents, the reasons which would explain trust may sometimes be justificatory. The social and biological science of trust, and its philosophy, should be mutually informing. Accordingly, I draw on results from empirical studies of trust to contribute to my account.

This question 'why should I trust?' may seem like a simple one to answer, but the appearance is deceptive. There are two lines of thought about trust that, taken individually, are commonsensical and compelling, and which concern different ways that trust can matter to us. But taken jointly, they may conflict, and give rise to such dilemmas. Starting with these divergent and sometimes conflicting ways that trust is valuable (§§1.1–3), this chapter then shows how the central philosophical-cum-conceptual debate about trust—namely, the project of identifying what the attitude of trust consists in—has been shaped by these competing intuitions. Roughly, cognitive accounts of trust, which take it to be a mental attitude solely or primarily evaluable in terms of truth, try to make sense of the ways in which I benefit when the person whom I trust comes through for me, while non-cognitive accounts of trust, which take it to be a mental attitude evaluable principally in practical terms, seek to accommodate the ways in which trust helps relationships to flourish (§§1.4–5). But, I argue, this debate is ill-formed. There are no strict rules on what is and is not an instance of trust, because we are inventive language users, able to take a root notion and apply it in divergent contexts of use. So, there are plural forms of trust, and the attempt to identify 'the' attitude of trust is a misguided one—a hunting of the snark (§1.6). But this does not mean that there is no philosophical understanding of trust to be had. The deeper issue is not what the attitude of trust consists in, but the ways that trust matters to us, and what those reasons are to which it responds; the kinds of trust that matter fall out from the answers to these questions. When enquiry into trust starts from the axiology, so the two archetypes of trust that emerge are that which follows the evidence, and that which goes beyond the evidence. This prepares the ground for the argument for the principal thesis of the book, that normally, your trust should follow the evidence, so that you trust the trustworthy, and not the untrustworthy

(§1.7). The chapter closes by comparing and contrasting my axiology-first approach with conceptual engineering (§1.8).

1.1 Instrumental Value

We come to know about the world by trusting what others tell us, through talking with friends, colleagues, and strangers, and listening to teachers; by learning the news from websites, the radio, and in print; and reading books and journals. We act on the basis of what we learn from others, and on the basis of trust that people will comply with their commitments. So, both the theoretical and practical dimensions of our lives are impacted by the possibility of trust. As for the former, we can often learn about the world with better results—in terms of accuracy, the speed with which knowledge is acquired, and the range of propositions available to us—by trusting others than we can by trying to find out ourselves. Likewise, with the latter, we often act in the world with better results—in terms of precision, range, and speed—by trusting others rather than trying to do it all alone. In military terminology, artillery and engineers are 'force multipliers'. When they support infantry and armour, the latter are greatly more effective in battle. One unit of investment in combat support may increase fighting power by the equivalent of five units of investment in a combat arm. Similarly, trust is an 'agency multiplier'. It increases one's capacity to get things done. Trusting well gives far better returns than the same investment of effort in doing the project alone. This is in terms of both epistemic and practical projects.

Trust is becoming more pervasive. A dominating feature of modernity is the astonishing increase in how much we now rely on each other. Human life has always been marked by patterns of social dependence. It is possible to be a hermit, but the choice is recherché; everyone else lives with other people, at least in the sense of physical proximity and material dependence. Reliance on others is deeper and broader now than ever before. It is deep, in that we rely on systems built and maintained by others to perform the most basic functions of life. For most of us, it is not clear that we could even fulfil these basic functions for ourselves, should we wish to. We eat food bought from the shop; drink clean water supplied to us from taps, paying plumbers and companies to build and maintain these systems; wear clothes made by others; and so on. Our dependence is also broad, in that there are hardly any activities in which we choose to be entirely self-reliant. Doing so usually indicates a hobbyist's desire to recover a kind of authenticity—no doubt commendable, but with the pleasure of doing it ourselves often meaning worse outcomes. Mundane activities routinely depend on the results of interaction between other people spread across the country and the globe, whom I will never know about, let alone meet. The book you hold in your hand, or the device you read this on, is a result of networks of people interacting

in ways that are almost unimaginably complex and wide. 'Trust is indispensable in order to increase a social system's potential for action beyond these elementary forms [of on-the-spot transactions]' (Luhmann 1973 [2017]: 98).

The science fiction thought experiment in *The Day of the Triffids* makes vivid the extent of our dependence. In John Wyndham's imagined world, a dazzling meteor shower blinds nearly everyone, making them vulnerable to a carnivorous plant, the triffid, and leaving only a handful of sighted survivors (Wyndham 1951). Aside from the drama of surviving among predatory plants, the power of the story is in its invitation to the reader to ask whether she could survive on her own wits, when all the normal structures of life are stripped away. The degree of divergence between our world and a triffid world shows how much our ordinary pattern of life relies on others doing their part.

As a basis for relying on others, trust has advantages that other ways of motivating reliable behaviour do not. The comparative claim above contrasted the results that trust often yields with those I get when I try to do it myself. There is another comparative claim. It is possible to rely on others but in a way that is not trusting. I may enhance my agency in a non-trusting way through deceit, manipulation, and threats of force or intimidation. These are strategies for making others reliable but are not instances of trust. In the paper that catalysed interest in trust in contemporary moral philosophy, Annette Baier observes that it is itself an indication that trust has ceased when security guards are needed to deter others from injecting poison into the food on the shelves in shops. The 'comedian, the advertiser, the blackmailer, the kidnapper-extortioner, and the terrorist' all show reliance without trust (1986: 234). Her point is that trust is distinct from reliance, and this is correct. The present point is that trust also often works better than these strategies, being more stable and more efficient for getting people to do what I want.

Take the stability of trust first. Suppose the person on whom I wish to rely sees through my manipulation, or realises the deceit, or works out how to neutralise the threat of force through resistance or escape. Then her reason for complying evaporates, and I am left high and dry, having no reason to rely on her. In contrast, when trust is accepted and the trusted person sees herself as thereby having reason to be trustworthy, that reason is not so easily overridden by changes in circumstance or the trusted's epistemic position. Trust may lead to that person being reliable across a much wider range of ways that the world could go. When this is so, trust is more stable. This leads to gains in terms of levels of actual reliability, as well as the subjective assurance for the trustor.

It can also be more efficient to establish and maintain trust-based forms of reliance. It is often costly and difficult to make non-trusting strategies work. The credible application of force requires a willingness to use violence, and a sufficient excess of power to make resistance a bad choice for the person threatened. To make threats effective, it also requires that people and time are allocated to 'guard

labour', monitoring compliance and imposing violence on those who do not comply (the term is from Jayadev and Bowles 2006). But guard labour is unproductive. Deceit and manipulation require control to be maintained over the subject's epistemic situation. They demand vigilance and sometimes creativity. In contrast, because it is willingly accepted, trust-based reliance requires no monitoring or enforcement. You simply ask something of another and let them decide whether they want to undertake the commitment. In economic terms, the transaction costs incurred by the other strategies are avoided. As well as being more stable, trust may be more efficient.

The instrumental benefits of trust are illustrated by the wide array of research into trust and its societal correlates. A full survey of the social science literature on trust is beyond the scope of this book, but the central point is that, in general, trust correlates with happy outcomes, and its lack correlates with unhappy outcomes. To indicate such results, in brief, note that surveys that seek to elucidate social values and beliefs have, over many decades and in countries globally, asked the following question:

> Generally speaking, would you say that most people can be trusted, or that you need to be very careful in dealing with people?

The question deliberately removes any clue about the social situation in which one may trust another. There is no indication as to how well the respondent may know the potential trustee, or whether there is any relevant social or professional role, or what may be at stake in the potential trust interaction. Instead, it asks about someone's 'generalised trust' towards unidentified others, regarding an unidentified trust interaction.

Answers to the generalised trust question can then be correlated with different outcomes, at both the individual and the societal level. Individuals with high levels of generalised trust report better physical health (Barefoot et al. 1998; Jen et al. 2010; Giordano et al. 2012); live longer (Barefoot et al. 1998; Giordano et al. 2019); are happier (Kuroki 2011; Helliwell and Wang 2011; Helliwell et al. 2018; Righi and Masserini 2021); are more likely to volunteer (Bekkers 2012), recycle (Sønderskov 2011), engage in jury duty (Uslaner 2002) and civic life (Uslaner and Conley 2003), start a business (Guiso et al. 2006), and participate in the stock market (Guiso et al. 2008); and they do better economically (Butler et al. 2016, although this study also finds that too much trust lowers income). They are also better at detecting lies (Carter and Weber 2010).

Across a society, a high level of generalised trust promotes economic growth (Fukuyama 1995; Knack and Keefer 1997; Zak and Knack 2001; Algan and Cahuc 2010; Bjørnskov 2018) and international trade (Greif 1989; Guiso et al. 2009); increases the productivity of firms (Bloom et al. 2012); leads to efficient public institutions (Putnam et al. 1993; Uslaner 2002; Tabellini 2008), improves

democratic governance (Inglehart 1999; Jamal and Nooruddin 2010) by influencing educational outcomes and the rule of law (Bjørnskov 2012); supports welfare states (Rothstein and Stolle 2003; Kumlin et al. 2018); promotes tax compliance (Scholz and Lubell 1998; Hammar et al. 2009) while diminishing corruption (La Porta et al. 1997; Wang and Graddy 2008; Uslaner 2008; Rothstein and Stolle 2008; You 2018) and reducing crime (Buonanno et al. 2009; Akçomak and Weel 2012). More recently, high levels of generalised trust have been shown to predict compliance with unusual restrictions imposed to restrict the spread of Covid-19 (Petherick et al. 2021). Living in a high-trust society is, then, much preferable to living in a low-trust society.

The above merely scratches the surface of what is now a vast body of empirical and theoretical work, developed especially over the last 30 years. As well as generalised trust, there is a literature on 'political trust'—trust of the central political institutions in a country—and 'particularised trust'—trust of known others. Questions regarding the causes and effects of trust similarly arise in these more specified relationships, parallel to those for generalised trust. The best, recent single-volume survey of the field is probably Uslaner (2018).

Trust, then, has instrumental benefits. Thomas Reid remarked that it is 'the most effectual engine of human power' (1788 [2010]: 666). Trusting reliance on others often yields better results compared to relying on oneself only, and to non-trusting reliance. In both these respects, trust is an agency multiplier.

1.2 Interpersonal Value

Trust matters to us, but not just because of what it enables us to do. Its instrumental benefits are only part of the story. Trust also matters because of what it signifies about our relationships, and the kinds of relationships that it constitutes and helps to enable. Call this its interpersonal value.

If trust mattered only because of its agency-enhancing potential, then nothing more would be needed to identify its value other than calculating how effective it is, in contrast to other ways of getting things done. The results of that calculation would say all there was to say about why trust matters; no residue of unexpressed value would be left. But that is not the case. Trust's role in our lives is richer and more complex than simply helping me to learn about, navigate, and impose my will on a world of objects. An account that explained why trust matters solely in these terms would be inadequate. It would be atomistic, in at least one of the senses that Charles Taylor uses the word: it would betray a view about human nature in which society matters purely instrumentally for the individual (1985: 187–9). Trust is not just an agency multiplier. Two considerations help to bring this out.

First, there are times when trust is less efficient than other bases for relying on people, and yet that does not settle the decision as to whether to trust. While trust

is often more stable and efficient than force, deceit, and manipulation, under some circumstances it is less stable and less efficient. Relationships based on force, deceit, and manipulation can also be resilient. We have a surprising capacity to endure domination and to accept the stories others tell to and about us. Think of how authoritarian regimes and abusive relationships can survive and indeed prosper. Trust is comparatively inefficient when I have forceful measures available to me and little sanction for using them, and a lack of information about the trusted means that a person's seeming acceptance of my trust gives me little assurance that she is really resolved to be trustworthy. So too for deceit and manipulation. In its own terms, trust is unstable just when the relationship between trustor and trusted is such that betrayal may be an attractive and relatively low-cost option for the trusted, or the person trusted is of weak moral character, or is susceptible to temptation—hardly a marginal possibility. Yet even when trust is comparatively unstable or inefficient, compared to these other non-trusting ways of relying on others, something is lost if those other ways are chosen. The loss is something to do with how I treat the person whose agency I bend to my will. The person is treated as a tool, and the relationship between us is corrupted. Deciding whether to rely on others on the basis of threats of force rather than trust is thus hardly a matter of just calculating which is likely to be more effective. These other strategies for promoting reliability cannot be substituted for trust without some loss of value. Because there is some loss of value when reliance is based on threats, deceit, and so on, rather than trust, there is some kind of general presumption in favour of basing one's reliance on trust. The value of trust, then, is not merely instrumental.

Second, consider the situation of the egoist. For the egoist, her own good is the ultimate basis for all her reasons for action. If others' welfare affects her, then she will care about them, but not otherwise. When she seeks to do good for others, it is only because doing so is beneficial for her in some way. As such, the egoist can recognise the agency-enhancing value of trust. She knows that it is not viable to do everything herself; she must rely on others, and sometimes cooperate with them. And she can rely on others in a trusting way—eschewing force, deceit, or manipulation, and instead depending on their word—because she knows that, by doing so, she may get better results than if she were to try these alternative strategies. Indeed, the egoist may be especially interested in trust, precisely because trust has an agency-enhancing effect, and she cares about her own good and wants to increase her welfare. Living among other people, she takes the material benefits that cooperation brings. But there is a clear sense in which, while living among other people, the egoist does not live 'with' them. Egoism is a pathology, in which the social dimension of a person is malformed. The egoist is affectively disconnected from others. As trusting relationships are a central way in which people are affectively connected to each other, an account of trust is thus inadequate if, by its lights, the egoist can endorse its full value. That the egoist endorses

the instrumental value of trust strongly indicates, again, that trust matters to us in other, non-instrumental ways too.

What is that non-instrumental value? An element of its value, and which is morally significant, is its expressive capacity. In trusting another person, I show my respect for them. To see this, consider again how trust contrasts with other bases for relying on others. When I threaten, deceive, or manipulate another, I treat her as an object to be used for my purposes. In contrast, when I trust her, I respect her as a person, able to choose of her own accord and in a way that is freely responsive to reason, and with the capacity to act on the basis of her commitments. I recognise that she is worthy of engaging with on that basis. I invite her to view herself, and me, in the same way. So, trust is both open and involves a moral appeal. It is open, in that both parties can say to each other that that is the reason for their relation of dependence; as Victoria McGeer and Philip Pettit term it, trust is 'publicly avowable' (2017: 24). Indeed, if we both recognise that trust is the basis for our cooperation, our trust may not just survive, but flourish. This is unlike deceit and manipulation, which must be concealed if they are to be effective. They cannot be mutually recognised as the basis for reliance without ceasing to be so; if both parties recognise deceit or manipulation to have been the basis of the relationship, then it either breaks down or comes to be based on one person's power over another. While threats of force are also open, in that they remain effective while both parties recognise that they are the basis for the dependence, they are nonetheless amoral. That is, in merely threatening someone else, without an accompanying justification for the force, the reason for action that you give to her is unrelated to the demands of morality; you have, in effect, tried to step outside of the constraints of morality. In contrast, trust involves a moral appeal, in the sense that its possibility depends on the demands of morality being recognised. In deceiving, threatening, or manipulating another, I act in a way that is contrary to the respect that my target properly deserves as a moral agent or patient. I deny her the respect that she is owed.

Further, in trusting someone, not only do I respect her, but I also express my respect to her. The act of trusting, grounded by my attitude of trust, is motivated by respect for the other. But the act of trusting is, in general, also evident to the person trusted. So, my act of trusting, as well as being motivated by respect, also shows her that I respect her. The respect is not hidden. Trust thus involves a kind of double respect: it is itself an appreciation of another's rational agency and her responsiveness to the demands of morality, and when I act in a trusting way, it also shows her that she is viewed in that light.

The expressive capacity of trust may be recognised by the egoist, even if she does not herself acknowledge the distinctive value that it thereby realises. Because it has this expressive capacity, and others recognise this value, the egoist may use trust for her own purposes. She may engage trustingly with others, predicting that her faked expression of esteem will lead them to be more cooperative than

would otherwise be the case, and this may be useful to her. So, trust may play a role in the covert exercise of power, perhaps by helping me to deceive or manipulate others. These are second-order possibilities, however. They arise only because trust expresses respect, and expressions of respect may be abused. The non-egoist recognises that, because trust is a basis for relying on others in a way that respects their moral agency, it is also to that extent valuable. Its capacity to express respect means that trusting relations matter both to those trusted and those who trust. Not only do I want to be treated with respect, it matters to me that I treat others with respect. I am not neutral regarding the means by which I come to rely on others, even though the egoist is. Again, the value of trust is not solely instrumental.

That trust expresses respect does more than give me reason to trust someone, however. It also creates a normative pressure on me to trust them. I walk up to you and ask you for directions to the train station, which you give me, sincerely and knowledgeably. I thank you for that, and then turn to the person next to you, asking her for directions to the train station. It is difficult to make sense of that second question without seeing it as a direct and emphatic slight on you. My action demonstrates that, for whatever reason, I have not trusted the directions you first gave me, thinking you insincere, or incompetent, or someone of severely diminished agency, to the point of madness. It is not just that my trusting you—which I have not done—would have expressed respect. It is also that my not trusting you has expressed disrespect, regardless of whether that absence of trust is characterised as distrust, or merely withheld trust. Because trust expresses respect, there is a normative pressure on me, as the trustor, to trust you. There is an experienced sense that I should trust you, that failing to do so is a slight or an insult.

The expressive capacity of trust does not exhaust its non-instrumental value, however. Not only does trust signify the attitude that I bear towards you, it may also constitute a particular kind of relationship between us, namely a trusting relationship. In trusting someone else, I share myself with her. I am vulnerable, committing myself affectively to another. Knud Ejler Løgstrup remarks that, in trusting another, one 'lays oneself open'. In encountering someone, it is open to him to affirm and to respect me; equally, he may denigrate and reject me. So, I may choose to encounter everyone with the openness of trust (1956 [1997]: 9). In this sense, trust is similar to love: it is an attitude of receptivity, of not being defensive. Stephen Darwall refers to trust accordingly as an 'attitude of the heart'. By this he refers to the affective part of the human psyche in which we are 'filled with hope and joy or deflated with despair, emptiness and sadness' (2017: 46–7). When two people each have this attitude towards the other, they stand with each other in a distinctive way, realising together a particular kind of good. Trust is part of what it is for two friends to be united in fellowship, where each is 'for' the other: each cares for the other, regarding his pleasures and his sorrows as his own.

It can be realised to greater or lesser degrees. Indeed, it may be that the trust that realises the good of friendship is not just like love but is itself an instance of love. While the egoist can misleadingly express respect for another, through acts of trust, which respect she nonetheless does not endorse, she cannot realise the good of trusting friendship. A constitutive part of that good is care for the other with whom one stands in fellowship.

To distinguish between these different kinds of non-instrumental value, consider the often-made claim that trust is a three-place relation (e.g., Horsburgh 1960; Baier 1986: 236; Good 1988: 33; Potter 2002: 10; Blackburn 2010: 92). When trust is a three-place relation, it relates two people, A and B, regarding some X. Paradigmatically, the X is some action in the future, but it includes testimonial cases, where the X is testimony to some fact about the world. Trust certainly takes this form, and this allows, rightly, that trust is often highly specific. I may trust the bus driver to drop me in the centre of town but not for financial advice. When A trusts B regarding some X, A expresses her respect for B; trust as a three-place relation successfully realises this form of non-instrumental value. But three-place trust does not exhaust the kinds of trust there are. 'Any adequate account of trust must cover those cases where there is no particular task we allocate to the trusted, but merely feel safe when moving among them' (Baier 2013: 179). In particular, three-place trust may be quite distinct from the richer good of a trusting relationship, in which two people are affectively united, care for each other, and are open and vulnerable. Aisha trusts Brett to pick up the children from school. Aisha's trust expresses a form of respect for Brett; she could have done the task herself, but she trusted him to. This does not imply that there is a thicker trusting relationship, however, in which, more simply, Aisha and Brett trust each other. The good of a trusting relationship arises in the latter situation, in cases of two-place trust, when A trusts B. In a trusting relationship, no doubt there are a great many three-place trust relations that may arise. But it mischaracterises the relationship to suppose that it consists in a series of three-place relations. Rather, there is something distinct from those three-place relations, which may also ground them; and, it may be that I wish to engage in a three-place trusting interaction with someone because I wish to establish, or build, a two-place trusting relationship.

There are other ways of marking the distinction which is tracked by that between two-place and three-place trust. The objects characteristically taken by the two-place and three-place attitudes of trust are different. In two-place trust, one trusts a person, and there is nothing further to it. In three-place trust, one trusts another person to carry out an action. So, we could also call the two-place relation 'person-centred' trust, and the three-place relation 'action-centred' trust (the terms are from Domenicucci 2021; others who have observed that trust may properly take a person as its object include Lars Hertzberg (1988: 315); Trudy Govier (1997: 57–9); Olli Lagerspetz (1998: 80); Linda Zagzebski (2012: 99ff.); and Christian Budnik (2018)).

That two-place trust is distinct from three-place trust relations has been argued for with a battery of overlapping arguments marshalled by Paul Faulkner (2015, 2017) and Jacopo Domenicucci and Richard Holton (2017), including the following. One is some linguistic data. 'A trusts B' is permissible, and need have no incompleteness about it. This is unlike 'A relies on B', where one wants to ask '…for what?' While English has the three-place construction—'A trusts B to X'—Latin, Italian, and French do not. Further, as the antinomy of trust, distrust surely has analytic relations to trust. Katherine Hawley has suggested that an account of trust must be unified with an account of distrust, making sense of how they are contraries, even if not contradictories (2014a, 2019). Yet there is no three-place construction for distrust in English. 'A distrusts B to X' is unnatural, where 'A distrusts B' is not. If there is a need for a unified account between trust and distrust, this suggests that there is a distinct two-place form of trust.

A further argument for the distinctness of two-place and three-place trust relations derives from the reciprocal nature of trust. Darwall identifies trust as an attitude that 'seeks reciprocation', and I take him to be correct in this. He fills out the details of this as follows. In trust, we 'invite the person we are trusting to accept our trust and trust in it, to trust that we are indeed trusting him'. This is especially needed when he is 'insecure and suspicious, both of his own abilities and character but also that others could be well disposed toward him'. By trusting him, we communicate that we regard him as trustworthy, and invite him similarly to view himself as trustworthy too. The same is true of love. It invites the beloved to view himself as lovable (2017: 48).

No doubt this is a fitting response to another's trust in me. But it omits what is, I think, the more obvious and important sense in which trust seeks reciprocation. In trusting you, I invite you to trust me. This is also the primary way in which love seeks reciprocation. In loving you, I invite you to love me. It is not as if, at the climactic moment of a courtship that both have delighted in, the hard-won declaration 'I love you!' is met with 'Now I realise that I am lovable!' It is met with the same declaration reciprocated: 'And I love you!' This primary way to reciprocate trust does not make much sense in terms of the three-place form, however. Perhaps you offer to replace the tyre on my car, and I trust you to do so. The fitting response to my trust is that you accept it, and then replace the tyre. While it is possible that you should reciprocally trust me to replace the tyres on your car, there is something quixotic about the arrangement. On the two-place form, however, the primary way to reciprocate trust is plain. A trusts B, and B reciprocates by trusting A. When trust is reciprocated in this way, the relationship realises a particular kind of value, that of a trusting relationship. I take this, with the above arguments, to establish successfully that two-place trust is distinct from three-place trust. This further suggests that trust realises a distinctive kind of value.

Trust matters, then, because of the kinds of relationship it makes possible. It matters expressively because, when I trust you rather than try to bend you to my

will, I respect you, and communicate that respect to you. Trust also matters because it constitutes a distinctive kind of relationship, akin to (and often part of) friendship; in trusting you, I play my part in enabling our relationship to realise that great good. Both of these are forms of interpersonal value—the kind of value that is characteristic of relations between people—and trust is a way of realising it.

1.3 Dilemmas of Trust

I noted at the start of this chapter that trust gives rise to practical dilemmas. It can now be noted that these dilemmas arise because of the two ways in which trust is valuable. That my trust both matters instrumentally, and demonstrates my respect or otherwise for other people, may put me in a quandary, as there are numerous situations where these considerations come apart and may conflict. The dilemmas of trust arise because I must decide what I should do in such a situation.

This is illustrated at its most vivid by situations in which one may trust someone who is clearly untrustworthy. The catalytic scene towards the start of Victor Hugo's *Les Misérables* is an example of this. On his release from the prison hulks, the destitute former convict Jean Valjean is welcomed by Bishop Myriel into the latter's home overnight. Valjean repays the kindness by stealing the silver and fleeing. He is captured and brought back by the gendarmes, with more years of hard labour the certain prospect. Yet Myriel denies that any crime was committed; adds his candlesticks to the rest of the silver, leaving Valjean stunned; and instructs him, 'Do not forget, do not ever forget, that you have promised me to use the money to make yourself an honest man' (Hugo 1862 [1982]: 111). Myriel's act is, most importantly, one of mercy and grace. But it is also one of trust, as Myriel gives Valjean the silver on the promise that he will become an honest man. Myriel's trust is supererogatory, beyond what anyone could expect, and it is heedless of how likely it is that Valjean will actually become honest. It is not that Valjean might be honest or might not be; the evidence is still out; perhaps Myriel could will himself into believing that Valjean is honest; and so on. It is plain that Valjean is not honest. Having been imprisoned previously for theft, he has now stolen from Myriel. But not only does Myriel entrust the silver to Valjean on the basis of the promise, there is a further act of trust on Myriel's part. Valjean has made no such promise. In offering Valjean the candlesticks and ascribing the promise to him, Myriel is also trusting that Valjean undertakes the promise for himself. Myriel has no grounds for thinking that he has done so, other than that Valjean says nothing, takes all the silver, and goes. Myriel trusts Valjean, in an act of double trust, even though he is clearly untrustworthy, and so Valjean's redemption becomes possible.

While the Hugo example is dramatic, instances abound in which it seems that trust is possible, without any strong view about how likely it is that someone will

be trustworthy. Richard Holton gives a similar scenario. He invites us to consider the case of someone convicted of petty theft, whom a shopkeeper decides to put on the till (1994: 63). The person's past gives the shopkeeper good reason to believe that he may be light-fingered. But, if everyone draws the conclusion that he should not be trusted, that past will exclude him from society. The shopkeeper is committed to the possibility of his restoration, so she views him as worthy of trust, even if it is not clear that he will not steal. Trust can also be used as a kind of moral training, as most parents recognise. Whether trusting a child with a letter that needs to be dropped into the postbox, or at a sleepover with some friends, in trusting one's child—in a way appropriate for their age, so at the right time and over the right kind of stakes—one gives them the opportunity to prove themselves trustworthy.

While Myriel's trust of Valjean is wholly supererogatory, there are other situations where there is a clear sense of pressure, pushing one to trust, even though the person trusted is not clearly trustworthy. Victoria McGeer's central example of trust is the friendship between Dorothea Casaubon and the young doctor Lydgate, in George Eliot's *Middlemarch*. Plausible rumours circulate in the town that Lydgate has taken a bribe. Two friends of Dorothea decide to wait for Lydgate to provide the evidence that will clear him before taking a view on the controversy. But Dorothea refuses to withhold judgment. 'Would you not like to be the one person who believed in that man's innocence, if the rest of the world belied him?' (quoted in McGeer 2008: 239). If friendship itself is partially constituted by trust, so distrust threatens that friendship (also Govier 1998: 21). For Dorothea not to believe that Lydgate was innocent would be for her to repudiate the friendship. Unlike Myriel, or Holton's shopkeeper, it is fraught with complexity for Dorothea whether or not she trusts. Whatever she decides, she sacrifices something, whether her friendship, or exposing herself to the risk of being wrong about Lydgate. In these situations, the interpersonal value of trust conflicts with its instrumental disvalue.

It is also possible that the instrumental and the interpersonal value of trust may come apart, but in the reverse way. Someone may realise interpersonal disvalue through trust, for the sake of realising its instrumental value. Scenarios in this latter category are perhaps less vivid than the former practical scenarios, of supererogatory trust, which readily spring to mind, but are nonetheless interesting. Think of Claire Underwood from the Netflix series *House of Cards*—a savvy, Machiavellian political operator who rises to become the president. Suppose it was to my advantage to trust her, in building a political alliance towards a policy goal that is worthwhile. I may feel a corollary pressure not to, because I do not want to dignify her by entering into the kind of moral relation that signifies all is well. To do so may be to make me complicit in her habitual wrongdoing by, in effect, brushing it under the carpet. But that complicity may be the price I need to pay for the sake of the policy goal. Decisions to prioritise instrumentally valuable

outcomes over any relational or moral consideration may easily arise in the commercial world. A financial adviser has mis-sold a product to a client, 'ripping their face off' for the sake of the commission, to the client's anger when this comes to light. But the client may still stick with that adviser in future because the adviser has an excellent network and is able to offer advantageous deals that the client would not otherwise come by. Is this trust? The relationship may not be exactly a trusting one, but there may well be instances of trust within it, in which the adviser's assurance on a phone call that this is a trade the client really wants to make is taken at face value and trusted. The adviser would be trusted for the sake of the instrumental benefits in doing so, and despite the fact that it expresses relational respect.

As well as explaining why some of the deeper dilemmas of trust arise, the different and sometimes competing ways in which trust is valuable have driven what is arguably the central debate in prior philosophical work on trust, namely the attempt to identify trust as either a cognitive or a non-cognitive attitude. Each prioritises one of the competing intuitions. The next sections describe this debate, and how it relates to this fundamental tension.

1.4 Cognitivism about Trust

As a first pass, on a cognitivist account of trust, it is an attitude that is solely or primarily evaluable in terms of truth, similar to mental states such as belief, prediction, hypothesis, and knowledge (taking this last to be a mental state. This statement of cognitivism about trust is qualified later). The simplest such account is offered by Russell Hardin. According to his 'encapsulated interest' account, trust is 'a cognitive notion, in the family of such notions as knowledge, belief, and the kind of judgment that might be called assessment. All of these are cognitive in that they are grounded in some sense of what is true. . . . The declarations "I believe you are trustworthy" and "I trust you" are equivalent' (Hardin 2002: 7, 10). Relatedly, for Partha Dasgupta, trust involves 'correct expectations about the actions of other people that have a bearing on one's own choice of action when that action must be chosen before one can monitor the actions of those others' (Dasgupta 1988: 51; other cognitive accounts include Good 1988: 33; Gambetta 1988: 217; Adler 1994; Fukuyama 1995: 26; Keren 2014, 2019; Reierson 2017; and Vallier 2019: 30, 36, 40). The shared claim, which makes each of these a cognitive account, is that trust is based on a truth-evaluable assessment that the person trusted will, in fact, do what they are trusted to do. Accordingly, trust can be criticised if that assessment turns out to be false or based on a poor assessment of the evidence. That may not be the only ground on which to criticise trust, but it is the central ground.

Cognitive accounts need not be committed to the idea that trust is a belief. On some of these accounts, such as Hardin's, trust just is a specific type of belief. Not

all beliefs are instances of trust, however; you may predict that your worst enemy will take every opportunity to expose your flaws, and there is no 'trust' about this, except in the most metaphorical sense. So, something further must distinguish trusting beliefs from other kinds of belief. A natural proposal is that what distinguishes trust from prediction, for instance, is that trust has some distinctive propositional content, which for Hardin concerns the trustworthiness of the trusted. But other accounts, such as Dasgupta's, do not require belief. Dasgupta takes trust to require only 'correct expectations'. You may have a correct expectation about how I will act without actually believing that I will act in that way. Whether my expectations are held with the strength or the evidential threshold characteristic of belief is a separate issue, although there is clearly a proximity between these possibilities. Nonetheless, cognitivism about trust does not directly entail doxasticism about trust—the view that trust is a belief.

Cognitivism about trust is motivated by a cluster of related concerns. Most obviously, it makes sense of the seeming platitude that you want to trust the trustworthy and not the untrustworthy. In addition, by making trust akin to or identified with belief, cognitive accounts explain cooperation in a way that is consistent with the Humean belief-desire model of action (see Hardin 2002: 7). Trust simply is the belief that another is likely to cooperate, or is functionally equivalent. Conjoined with the desire for its results, cooperation ensues. Last, and in more broad-brush terms, cognitivism about trust brings a concept that is central to our social lives, and is seemingly in tension with the idea that rational agents pursue their self-interest, within the orbit of utility-maximising decision theory. No odd moral notions need to be invoked.

This last indicates how this cluster of concerns are related. They are united by a background sense that, outside the family and intimate friendships, human interaction generally involves individuals trying to do as well for themselves as possible, and that it is naïve to expect anything else. As Lawrence Becker describes the instinct, 'Saintly souls who are persistently trustful and serenely indifferent to the treachery around them may have a few temporary successes at the margins…but more often they are exploited and pose a danger to themselves and others' (1996: 47; as I discuss below, Becker goes on to reject this thought). To the extent that one is drawn to this view of human nature, that it consists in individuals seeking their self-interest to the greatest extent possible, trust poses a problem. How can trust not be mere gullibility? Yet, in folk theorising about the world, trust is not only an important goal, but also a realistic possibility. This tension must be resolved, and cognitivism about trust is a way of doing so.

Simple accounts of cognitive trust, which propose that trust is nothing more than, or equivalent to, the belief that someone is trustworthy, are subject to some established objections. The examples of trust given above—of Bishop Myriel, the ex-con on the till, parental training of children in trustworthiness, and Lydgate (§1.3)—all serve as counterexamples to a solely cognitive account of trust. In all

of these, someone trusts another while lacking adequate reason to believe that someone is trustworthy, or even despite strong reason to believe that they are untrustworthy, and the trust seems to be perfectly defensible, perhaps even praiseworthy. In these instances, trust is responsive solely or primarily to interpersonal considerations. Further reasons for taking trust to be responsive primarily to interpersonal considerations are not hard to find. H. J. N. Horsburgh thought there is a general *prima facie* duty to trust because the discouraging effects of distrust are so severe (Horsburgh 1960; for scepticism, see Harding 2011). Similarly, D. O. Thomas thought that one ought to trust friends and intimates simply because they are friends and intimates (Thomas 1979, so anticipating McGeer 2008; relatedly, see Stroud 2006; and for qualification, see Hawley 2014b).

The same scenarios are often taken to be counterexamples to a purely cognitive account of trust on the grounds that they show trust sometimes to be voluntary, whereas belief is not voluntary. It seems like Bishop Myriel can decide whether to trust Valjean with the candlesticks, and that the shopkeeper can decide whether to trust the ex-con with the till. If one can decide to trust in a way that one cannot decide to believe, doing so in situations where the reasons one has for taking the person trusted to be trustworthy are some way short of decisive, or sufficient to justify reliance, then a cognitive account of trust must be rejected (Holton 1994; Faulkner 2011: 149, 2014a: 1979; Hawley 2014b: 2030).

Not only do all these cases seem to be defensible examples of trust, they also seem to be important examples of trust. Where trustworthiness is uncertain, it seems trust may weave its magic, helping to create flourishing relationships. As well as a cognitive account of trust having to explain—or, the critic objects, to explain away—these cases, it has to explain why these cases seem to matter to us.

Objections to cognitivism about trust thus do not depend on putative counterexamples alone. There is a persistent suspicion that cognitive accounts of trust, in some important sense, do not take seriously the ways that trust is distinctive. Laurence Becker has expressed this suspicion forcefully. He suggests that trust can be characterised as involving a disposition to believe what others say, a disposition to rely on them, and a sense of security about their motives (in effect, that they are acting in good faith). He terms these characteristics, respectively, credulity, reliance, and security about motives. It seems also that trust is relevant when I have some deficit in knowledge or power, and so am unable to pursue my goals effectively. But, when it comes to pursuing my goals effectively, it is wholly unclear that credulity, reliance, and security have any place. Each seems 'superfluous at best, and at worst a recipe for undermining rational decision making'. In each case, I should believe what others say, or rely on them, or make any hypothesis about their motivation, only to the extent that I have reason to believe that they are truthful, or reliable, or acting in good faith. Thus, cognitive accounts of trust have a 'disturbing peculiarity', in that they 'appear to eliminate what they say

they describe'. Trust 'is either made synonymous with the knowledge and power we can gain through strategically defensible thinking or dismissed' (Becker 1996: 47–50). Call this the 'disappearance of trust' dilemma. The concern that trust should not be mere gullibility seems to push one onto either horn of the dilemma.

Oliver Williamson's view of trust, which radically delimits its scope, illustrates well the latter horn. On his view, outside of intimate relationships, human behaviour should be assumed to be calculative, in which people pursue their self-interest with guile and with limited cognitive capacity. Yet calculativeness drives out trust. 'I subscribe to the notion of economizing on trust.... [T]rust, if it obtains at all, is reserved for very special relations between family, friends, and lovers' (Williamson 1993: 483, 484). This is a highly revisionary claim, however, denying that much human behaviour that seems trusting is in fact so. It is so revisionary that many will take the position to be its own *reductio*.

Recognising that this is highly revisionary, however, the alternative is to account for trust in a way which allows for trustworthy behaviour to be motivated by self-interest and calculation. As Becker has it, this is to make trust synonymous with strategically defensible thinking. A solely cognitive account permits this, as it may be neutral on the reasons for which someone may do what they are trusted to do. Such accounts face a challenge, however. It seems that accounting for trust in this way makes it, in effect, an attitude generically directed towards others' cooperation. Encompassing such a wide array of phenomena, it then becomes difficult to identify what makes trust interesting. Further, if 'trust' denotes merely a general attitude towards cooperation, then '[t]he claim, commonly made by social scientists, that trust is a form of social capital that reduces transaction costs, is false' (Jones 2004: 17).

Given the role that a background conception of human nature may play in motivating cognitivism about trust, it is no surprise that such accounts have often been espoused by those working in academic disciplines which take *homo economicus* to describe the bulk of human behaviour, with the corollary implication that defences of ethical principles are failed or failing attempts at moralising. Accounts of trust in such disciplines are readily recognised as having adopted either horn of the 'disappearance of trust' dilemma. By contrast, the great majority of philosophers writing on the topic have, in the main, been more impressed by scenarios in which the interpersonal dimensions of trust loom large. In this they tend to be joined by theologians, anthropologists, poets, and story-tellers—those, perhaps, who are more impressed by the idiosyncrasies of human nature, by our capacity for love and attachment, and who take moral commitment to be a genuine possibility. Returning to Becker, he concluded that, because cognitive accounts of trust seem to lead to its disappearance, there must be something else which characterises it, and accordingly identified trust as a non-cognitive disposition.

1.5 Non-Cognitivism about Trust

On a non-cognitivist account, trust is an attitude that is not solely or primarily evaluable in terms of truth, but properly responds to considerations other than whether someone is trustworthy. Non-cognitivism thus takes trust to be evaluable principally in practical terms, in the way that attitudes such as intention, resolution, and desire are. A non-cognitive account need not require trust to be an attitude, but may be a disposition, as per Becker's proposal, or be realised in some other way. The principal motivation for non-cognitive accounts lies in the problems that cognitive accounts face. In rejecting the background conception of human nature that often motivates cognitivism about trust—and taking humans to be capable of loyalty, mercy, and acting on principle; of often seeking the common good as well as individual interest; and of forming wide attachments as well as a few intimate ones—so the interpersonal dimensions of trust loom larger.

Proposals then vary as to what non-cognitive trust consists in. Start with Becker's view, for instance, that trust consists in a 'disposition to be trustful' towards another, where this general disposition consists variously of credulity, reliance, and security about others' motives. This disposition may be held independently of whether, on a given occasion, I actually believe what someone else says, or rely on them, or make any assumption about their motives. Further, the disposition may be independent of our beliefs or expectations about their trustworthiness, or whether we believe they deserve it (1996: 44–5, 50). Becker's account usefully brings out some structural features of non-cognitive accounts which should be noted; here, I focus on two.

A first, structural feature of Becker's view of trust is that he takes it not to be a belief. This reflects a more general point that non-cognitivism about trust is readily conjoined with, and indeed entails, non-doxasticism about trust. Richard Holton makes the connection. He takes a game from drama classes as an example, where someone is invited to fall backwards with her eyes closed. She presumes that the others will catch her but does so in the absence of any formed belief that they will. It feels as if she is at that moment deciding whether or not to fall. In deciding whether to fall, she decides whether to trust. Although Holton does not state the reasons for which one decides to fall, and thereby decides to trust, they clearly go beyond just whether those behind will catch. Holton proposes that this phenomenology of decision should be taken at face value and concludes that some instances of trust are voluntary. Because belief is not voluntary, so trust does not involve belief (Holton 1994: 63).

A second feature of Becker's account is that, on his view, non-cognitive trust involves the trustor ascribing, to the person trusted, a specific motive towards herself. The motives to which Becker takes trust to be properly responsive are a capacious group; they comprise benevolence, conscientiousness, and a commitment to justice, in the form of reciprocity (1996: 53). In sharing this structural

feature, Becker's account follows Baier's, which likewise ascribes specific motives towards the person trusted. On Baier's view, that motive is one person's good will towards another, where this contrasts with their ill will or indifference. Further, Becker follows Baier's account in a more specific way, making explicit what is only implicit there, namely that the motives ascribed to the person trusted encompass both the affective and the deontic, whereby some kind of duty is recognised, or a related normative standard. The concise statement of Baier's view—'dependence on good will'—seems to ascribe only an affective motivation to the trusted. The longer discussion reveals a more complex picture, however. In trusting, I take the other's good will to include her understanding what is involved in expressing it appropriately. Creeping up and seizing me from behind in the library's book stacks, out of the sincere concern that I should learn to be more cautious in the future, would breach that trust (1986: 237–8). This deontic dimension, only implicit in Baier's account, is made explicit in Becker's, in which, in trusting, I ascribe a commitment to justice to the trusted.

By ascribing specific motives to the person trusted, such accounts successfully distinguish trust from situations in which another's cooperation is based solely on fear. But they do not distinguish it from situations in which someone's willingness to cooperate is based on manipulation. Richard Holton observed that the confidence trickster might rely on your good will, yet not trust you (1994: 65). So, not only is it not sufficient for trust that the person trusted should bear good will, it also seems that trust involves some further, distinctive attitude on the part of the trustor. Holton's own account picks up on the deontic dimension of trust, broadly construed, and takes this to characterise the trustor. He proposes that trust involves adopting the 'participant stance' towards the other, being disposed to feel gratitude if they uphold my trust, and resentful if they should betray it. These attitudes characteristically arise in viewing the other as a person, not a machine—such 'reactive attitudes' are not appropriate towards cars (Holton 1994: 65, 67; he draws on Strawson 1974).

Non-cognitive accounts have then variously combined these elements: affective and deontic attitudes, on the part of both the person trusted and the trustor. On Karen Jones' early account of trust, it is an attitude of optimism about the other's competence and good will, with an expectation that the one trusted will be directly and favourably moved by my dependence (1996. In her later, revised account of three-place trust, this expectation is explicitly a normative attitude; see Jones 2004: 18). Jones' early account builds on Holton's, in taking trust to involve a kind of expectation on the part of the trustor, as well as the attitude of optimism, directed towards specific features of the person trusted, namely their competence and good will (relatedly, see Helm 2014; Marušić 2017; and Thompson 2017). For Bernd Lähno, trust is a specifically emotional attitude on the part of the trustor, not just optimism, but also involves normative attitudes (2001, 2020); Carolyn McLeod's account is structurally similar, taking trust to be an emotional attitude,

encompassing perceptual, behavioural, and cognitive dimensions, and which includes optimism about the trusted's moral integrity (2002: 9, 85). That trust involves some kind of deontic attitude similarly features in Paul Faulkner's account, who takes it to involve normatively expecting the person trusted to be motivated by the fact of my dependence on her (2011: 143–50). On views which are closely related, but nonetheless distinct, the attitude of trust may not be itself deontic, but may be an attitude towards a deontic fact. Philip Nickel takes trust to involve the ascription of an obligation to the person trusted (2007; Tallant 2022 is closely related), and Katherine Hawley argues that trust involves my believing that the other has committed to do something, and relying on her to do so (2014a, 2019: 9; in his 2021 article, Matthew Bennett expands the category of commitments, so that it may be either a normative or a psychological commitment). Others focus on the affective dimension, with Simone Belli and Fernando Broncano taking trust to be a meta-emotion (2017; see also Evan Simpson 2013). Non-cognitive accounts need not require trust to be a mental state; one proposal is that trust is a non-inferential, discriminating disposition to rely on trustworthy parties (Kappel 2014).

1.6 The Pluralist Challenge

In the rivalry between cognitive and non-cognitive accounts, and the different variants and sub-variants that each may take, the central philosophical debate about trust has been an attempt to answer the question, 'what is trust?' Given the long history of philosophical attempts to probe puzzling and important features of the world by asking questions of the form, 'what is X?'—what is knowledge? What is free will? And so on—this project has a natural philosophical gravity. Not only so, but there is a readily available method for it, with philosophical questions of the form, 'what is X?', being habitually answered by conceptual analysis, assuming that the rules for the correct use of the term 'X' are determinate and therefore can be stated, because there is some identifiable X which is being referred to. At its zenith, the project issues in a definition in the form of necessary and sufficient conditions, using only terms understood more clearly than the definiendum. Insofar as the dispute between the various cognitive and non-cognitive accounts of trust is construed as rival claims of the form, 'trust is *this*', where there is a unique, correct answer, the controversy will be settled by taking as data points our judgments about whether trust occurs in each of a number of cases; identifying a candidate attitude that might yield that pattern of verdicts; and then subjecting the answer to trial by counterexample, refining the proposal until we reach reflective equilibrium.

This seems to me to be the wrong way to go, however, and there are two considerations which jointly motivate a different approach. First, conceptual analysis

is not apt for identifying what trust is, because there are strong grounds for supposing that there is no single attitude to which 'trust' refers, with a diverse range of phenomena all being naturally identified as instances of trust. Call this *pluralism about trust*. Pluralism about trust is not primarily a claim about the concept of trust. It is a claim about the nature of trust itself, namely, that it takes plural forms, and in particular, that it is sometimes cognitive and sometimes non-cognitive.

Initial evidence for pluralism about trust is given by the vulnerability of existing analyses of trust to counterexample. I argued above that cognitive accounts are subject to simple counterexamples (§1.5). However, this point is as true for non-cognitive accounts as it is for cognitive accounts. Within the family of non-cognitive accounts, affective approaches to trust are vulnerable to counterexamples where trust seems to have little to do with good will or any emotion. Against Baier's account, Onora O'Neill makes the point that a patient may trust a doctor to exercise proper professional judgment in their case, while knowing full well that the doctor finds them particularly irritating (O'Neill 2002b: 14). As well as O'Neill's objection, affective accounts are vulnerable to counterexamples where trust seems to be a matter of cold calculation on the part of both parties. Think of a group of oligarchs arranging to price-fix. There is no love lost between the ruthless competitors. But they may still successfully manage to collude on raising prices over a staggered period, to avoid the suspicion of coordinated action, with at least the initiating party having to trust that the others will follow her lead and not take advantage of the situation by maintaining the new price differential to increase market share. The initiating oligarch does so only because she has good reasons to believe that the others will follow; optimism does not come into it. Trust is not always an affective matter.

Deontic accounts, part of the family of non-cognitive approaches, are vulnerable to counterexamples from instances of trust between intimates. A father certainly could be relied on by his daughter while she adopts the participant stance towards him, for instance. But it is odd to suppose that the participant stance is doing any distinctive work here. The daughter trusts her father because he loves her, and she knows that. She need not even have the concepts required to adopt attitudes such as those mandated by the participant stance; take a two-year-old looking forward to her breakfast, for instance. Matthew Bennett gives the example of a friendship which dissolves as one party changes and interests drift apart. That a friendship built on trust can dissolve without wrong being done shows that the trust which constituted that friendship was not, fundamentally, deontic (Bennett 2021: 520–3).

Reasons of space prohibit surveying all claims of the form 'trust is this', so I can do no more here than report that counterexamples can be similarly easily generated for all those I have found. I suggest the reader will find the same. As the cognitive, affective, and deontic approaches to trust represent the major ways of

thinking about trust in the literature, this is inductive reason to think that other such attempts will be similarly vulnerable to counterexample. The advice is then: when you are in a hole, stop digging. The defender of the possibility of an analysis of trust may reply, however, that trust has received relatively little attention from philosophers, especially when compared to endeavours like the analysis of the concept of knowledge. That existing definitions are vulnerable to counterexample does not show that trust is not amenable to such a treatment. Perhaps no one has given it a really decent try yet. (Digging over here may get us out.)

I am dubious. The ways that the word 'trust' is used are simply too various to be regimented into one definition. This provides further evidence for pluralism about trust. Start with the range of attitude that it is used to refer to. Sometimes 'trust' is naturally understood as referring to a sort of affective attitude: 'I will trust my wife; I will not be jealous.' Sometimes it is naturally understood as referring to a deontic attitude of expectation: 'I trust you to return the car in time; I'm sure you won't let me down.' At other times it refers to a conative attitude, more akin to a decision about how one will act: 'Come what may, I will trust you to the end.' And at yet others it refers to cognitive ones: 'I know you are honourable; I trust you.' Sometimes it is not a mental state but action itself which is described as trust: 'The patrol followed the scout, trusting him to spot an ambush before it was too late.' Or, it may be not a specific action but a state of affairs, characterising a relationship, rather than an attitude: 'The neighbours lived in harmony and trust' (attention has been drawn to this by Fay Niker and Laura Specker Sullivan 2018). Klemens Kappel remarks that it 'seems to be a fairly heterogeneous family of related phenomena that may be variously picked out when we talk about trust' (2014: 2010, also 2025).

As well as natural uses of 'trust' suggesting that the term may refer to a range of different types of attitudes, there is a variety of locutions in which the term may be used which are identifiably distinct. I noted earlier that trust may take a two-place and a three-place form, in which, schematically, A trusts B, and A trusts B to do X, and adopted the terms 'person-centred' and 'action-centred' trust respectively. This oversimplifies a more complex reality. There is a distinct three-place trust locution which highlights that one's trust may involve entrusting something to another, such as trusting a friend with one's pet tortoise during a holiday—schematically, A trusts B with V, where V is a valued object. Another locution highlights that one person may trust another in relation to a specific role, such as trusting Jenny as an electrician—A trusts B as an R, where R is a role. Another locution highlights that one person may trust another regarding what they say—A trusts B that P, where P is a proposition. As well as the basic 'A trusts B' locution, two-place trust also has the important and distinct 'trusting in' locution, in which A trusts in B. Further, all of these trust locutions can be contrasted with propositional trust, in which the object of one's trust is a propositional clause; schematically, A trusts that B will do/has done X. One's propositional trust

may then be directed at versions of the above, three-place trust relations, but doing so seems to demand that the nature of the trust be spelled out, and there is no uniquely correct way of doing so. For instance, A may trust that B will look after V, or A may trust that B will fulfil her role as an R, or A may trust that B has spoken truthfully about P. Interestingly, there is no parallel, propositional construction for two-place trust: it is nonsensical to say, 'A trusts that B'. But propositional trust need not be directed at a three-place trust relation; I may just trust that the fog will clear. (I am grateful to Michael Pace for highlighting these points.)

Ordinary use of the term 'trust' thus indicates that there is a diversity of phenomena—encompassing a range of attitudes, action, and states of affairs, which are sometimes marked by distinct two-place and three-place locutions—which it is felicitous to describe as 'trust'. As the examples show, at minimum, any successful set of necessary and sufficient conditions for trust will not require it to be always the same kind of mental state. These all support the inductive argument against the plausibility of analysing trust as a particular kind of mental attitude. More widely, counterexamples can be given so easily because there are so many ways the word may permissibly be used. It would be foolish to seek a single definition because diverse attitudes, acts, and states of affairs are all instances of trust. Pluralism about trust is not undermined by the fact that advocates of claims of the form 'trust is this' may have replies to proposed counterexamples. The replies will tend to consist in argument over the cases—'is that really an instance of trust?'—and ultimately a discovery that different people use the word 'trust' to describe different things on different occasions. This reinforces the case.

Why does trust take plural forms? The equivocity that 'trust' exhibits is not the same as that of 'bank'. It is a linguistic accident that 'bank' is ambiguous between a financial institution and the side of a river. It is no accident that 'trust' functions as an umbrella term that may refer to a variety of phenomena that, while non-identical, nonetheless share a range of similar features. The explanation for why these are referred to by the same term is partly analogical, due to resemblance, and partly genealogical, by which a root notion is linguistically salient in divergent but related contexts of use, which in turn may change its meaning. The task of the genealogist of trust is to identify that root notion, and the social purpose (or purposes) which that root notion serves. I give a genealogy of trust elsewhere, proposing that the root notion is reliance on others' free cooperation (Simpson 2012a. Carolyn McLeod's focus on prototypical instances of trust, which likewise eschews an analysis in terms of necessary and sufficient conditions, has some overlap with a genealogical approach; see McLeod 2002: 11–34). For present purposes, however, nothing depends on what the correct genealogy of trust is. What matters is that trust is susceptible to genealogical enquiry. That it is bolsters the case for pluralism about trust and gives principled reasons against attempting to refine and defend a definition of trust.

More specifically, in allowing trust to take both cognitive and non-cognitive forms, pluralism about trust deflates—or, at least, reframes—the rivalry between cognitive and non-cognitive accounts. If trust takes plural forms, then there is no problem in recognising instances of trust which are primarily affective or deontic attitudes, and so trust is non-cognitive, while recognising other instances in trust is primarily a matter of believing that someone is trustworthy, and so trust is cognitive. Each consists in mental states closely related to whatever the root notion of trust is, and which I have proposed to be reliance on others' free cooperation. That any one form of trust exists in no way suggests or indicates that others do not. Adjudicating between the various cognitive and non-cognitive accounts on the basis of trial by counterexample makes sense only if a supplementary, suppressed premise is granted, namely that 'trust' is univocal. Pluralism about trust denies this and indicates that an understanding of the reasons of trust should be sought through some means other than conceptual analysis.

A second consideration, which also counts against attempting to adjudicate on the various cognitive and non-cognitive proposals, and decisive for present purposes, is that an account of trust is unlikely satisfactorily to answer the focal and practical question of when and why one should trust. Karen Jones's early discussion of trust starts by noting that an account of trust 'sets constraints on what can be said about the justification conditions of trust'. Proceeding on the assumption that there is a correct account of trust, she takes it to be a merit of her non-cognitive, affective account that 'it is able to view a wide range of our trustings—including many of those undertaken for instrumental reasons—as justified', whereas a cognitive account is more restrictive (1996: 4, 5). But, as argued, this assumption is suspect, and a reason for taking it to be suspect is precisely the consequence that Jones identifies, namely that it must rule as unjustified some cases in which the conditions for her affectively based attitude are not met, yet which an observer would nonetheless take to be instances of justified trust. To avoid such consequences, an account may instead take seriously the divergent contexts in which trust seems appropriate. Karen Frost-Arnold's (2014) proposal is illustrative, in which trust is either the belief or the acceptance that B will be trustworthy. The disjunction of these two attitudes, belief and acceptance, thus allows that trust is sometimes undertaken in response to another's trustworthiness, as a belief, while at other times it is undertaken for non-trustworthiness-related reasons, as an acceptance may, being a supposition for the purposes of practical reasoning (on the attitude of acceptance see, canonically, Bratman 1992; Cohen 1989, 1992. Dimock 2020 is another disjunctive account, allowing trust to be either cognitive or non-cognitive). But, while giving a more adequate account of what trust consists in, a disjunctive proposal then provides less by way of guidance as to why one should trust. (It also faces the charge of being ad hoc.) The more general point is that, while the two are related, there is no simple entailment from an account of what trust consists in to what those reasons are to which trust responds.

1.7 Axiology First

Rather than attempting to define trust, then, the goal is to get clear on the reasons of trust. To do so, start with a simple conceptual connection between value and reason: there is reason to bring about or preserve value. So, there are reasons to act in ways that bring about or preserve value, or to hold the attitudes that appropriately do so. The conceptual connection justifies a *pro tanto* reason only. Other considerations could show that reasons are overridden by competing reasons which are of greater weight, for instance. The conceptual connection is a 'minimal form of engagement with value' (Raz 1999: 164). It is not committed on the relative priority of reason or value (so it is compatible with, in particular, the 'buck-passing' account of value, proposed by Scanlon 1998; see Heuer 2006 for analysis of different claims between the relative priorities of reason and value). As there are different kinds of value, so there are corresponding kinds of reasons. There are moral reasons to bring about or preserve moral value; aesthetic reasons to bring about or preserve aesthetic value; and so on.

Recall the forms of value that trust may realise or help to promote, both instrumental and interpersonal. As trust has instrumental and interpersonal value, so there are corresponding reasons to trust. There are instrumental reasons to trust, insofar as trust will be an effective agency multiplier, and there are interpersonal reasons to trust, insofar as trust will realise the interpersonal value of expressing respect and partly constituting a friendship.

Take its instrumental value first. Being a more effective agent is valuable for me due to the projects I can thereby complete. This gives me instrumental reason to trust. But it does not give me instrumental reason simply to trust. It gives me instrumental reason to trust well. Trust is an agency multiplier only to the extent that I trust the trustworthy and do not trust the untrustworthy. The consequences of trusting the untrustworthy are more dramatic, at least in psychological terms; they consist in betrayal and unreliability. The consequences of not trusting the trustworthy tend to be less psychologically salient, as they consist in opportunities foregone, even though in the long run these may be equally or more significant. In seeking to trust well, then, I want to trust the trustworthy. In the first instance, the trustworthiness that instrumental reason seeks to identify is thin. It consists in reliability: a person fulfilling what they have been trusted to do, in paradigm instances because they have undertaken a commitment to do so, whether fulfilling that commitment involves them keeping their promise, fulfilling their commitment, or speaking truthfully. (In taking this as the paradigm I follow Hawley 2014a, 2019.) It does not require that they are reliable for any particular reason—say, because they are virtuous, or morally committed, or whatever. This does not imply an all-things-considered judgment that trust is compatible with any kind of motivation on the part of the person trusted. That threats, deception, and manipulation can be contrasted with trust, as a basis for

reliance, suggests that some motives are incompatible with trust, and the genealogical account of trust I favour, on which trust is reliance on others' free cooperation, explains which forms of motivation are compatible, namely those which are consistent with one's free cooperation. Those who are threatened, deceived, or manipulated do not cooperate freely.

Because of the instrumental value of trust, I should be sensitive to the reasons there are for believing someone to be trustworthy. To trust well, in instrumental terms, I must successfully discriminate between the trustworthy and the untrustworthy, and the reasons by which I do so are theoretical. They exhibit a 'mind-to-world' direction of fit; that is, by them I conform my mind to the world. This is so regardless of whether the outcome of my trust is practical or theoretical. Suppose the desired outcome of my trust is theoretical—that is, I wish to learn about the world from a testifier. Then I want my trust to be sensitive to the reasons there are for thinking that she knows what she is talking about. Or suppose the desired outcome of my trust is practical—I wish to complete a project by collaborating with someone. Then I want my trust to be sensitive to the reasons there are for thinking that she is competent to undertake this kind of project and will be diligent in doing so with me. Successfully discriminating the trustworthy from the untrustworthy is not all there is to trust but, in order to realise its instrumental value, it is a necessary part, and likely to be the dispositive part. In sum, when I have reason to trust the trustworthy and not the untrustworthy, my trust should be based on theoretical reasons.

This is not to imply that trust should always be held for reasons which derive from its instrumental value. It asserts that, insofar as trust matters instrumentally, I must be sensitive to the theoretical reasons I have for taking the trusted to be trustworthy. I should evaluate how significant they are according to the demands of epistemic rationality. Once I have done so, those theoretical reasons, taken together, to that extent have significance in judging whether I ought to trust. They are then significant because of the instrumental value there may be in trusting.

Now consider the interpersonal value of trust. There are interpersonal reasons to trust because of the interpersonal value that it realises: the respect it expresses, and the trusting relationships it enables. The reasons that I have for bringing about or preserving this value are practical. In acting in a trusting way, or having the attitude of trust, I bring about something that matters to you and me. The reasons I have for doing so exhibit a 'world-to-mind' direction of fit; by them, I conform the world to my mind. (This includes those cases in which my mental attitude is the relevant new worldly fact that I have successfully created, in conformity with my will; that is, simply by having a valuable attitude, such as that trust which in part constitutes my friendship with another, I have created new value in the world, which is what I intended to do.) The reasons I have to do this are practical, of which interpersonal reasons comprise a subset. In sum, when

trust has interpersonal value, expressing respect for others and enabling trusting relationships, I thereby have practical reasons to trust.

As before, this applies only when my relationship with you could bring about interpersonal value. Yet trust nearly always matters interpersonally, at least because it expresses respect for the trusted, and expressing respect generally contributes to the goodness in the world. Less widely, but nonetheless importantly, I also often have reason to enable those friendly and intimate relationships marked by two-place trust.

I stated earlier that cognitive and non-cognitive accounts are well understood as being motivated by competing intuitions about the value of trust, whether it matters in terms of its outcomes or in terms of interpersonal relationships. While this is generally true as a description of prior work, it is not fully accurate, as there are accounts which are formally cognitive, but are 'non-cognitive in spirit', and there are accounts which are formally non-cognitive, but are 'cognitive in spirit'. This suggests that the reasons to which trust is responsive reflect a deeper divergence in the different forms of trust, than whether it is identified as a cognitive or non-cognitive attitude. Illustrating the former, there are doxastic accounts of trust—which are thereby cognitive—which nonetheless allow that trust may appropriately respond to considerations other than the trustworthiness of the person trusted. On Hawley's conjunctive account of trust, trust involves believing that the person trusted has made a commitment, and relying on them to fulfil that commitment. So, for Hawley, trust involves a belief. This belief is relatively easily acquired, however, on the basis of the facts as to whether someone has made a commitment, whether explicitly or implicitly, and clearly enough, whether someone has made a commitment is a very different question than whether they are likely to fulfil that commitment. The more important factor in determining whether one trusts is whether one decides to rely on the other person; practical, interpersonal reasons for reliance are thereby included when one deliberates as to whether to trust. Thus, Hawley's account is formally cognitive, but it does not endorse the thought which motivates accounts like those proposed by Hardin and Dasgupta—that when you trust, you want to trust those who are trustworthy. (While Hawley states that trust is 'usually … conditional on the trustworthiness of the trusted person' (2019: 73), this claim is not developed further, and her commitment-based account does not require or imply this.) Also in this category is Anthony Booth's thesis that trust takes 'the guise of belief', fulfilling the role of belief but responding to pragmatist concerns (Booth 2018), as well as, importantly, the family of 'second-personal' accounts of trust, which take it to be a belief about another's trustworthiness, but with that belief adopted for reasons that are second-personal in nature—deriving their justificatory force from the interpersonal relationship between the trustor and the trusted—also fall into this category. (I return to these views below, and consider them in detail later, in Chapter 4.)

Contrariwise, and illustrating the way that a non-cognitive account may be 'cognitive in spirit', some ways of thinking about trust construe it not as a mental attitude but as an action. According to Piotr Sztompka, 'Trust is a bet about the future contingent actions of others' (1999: 25; see also Coleman 1990: 99; Tutić and Voss 2020: 180. Pettit 1995 broadly adopts this approach too). While this approach is formally non-cognitivist, in claiming that trust is not an attitude, it is motivated by the same concerns that underlie most cognitivist accounts. The distinction between cognitive and non-cognitive accounts, then, does not reliably correlate with the deeper disagreement about the reasons to which trust responds.

That disagreement arises because of the different ways in which trust is valuable, and the conceptual connection between reason and value. When the forms of value which trust may realise conflict, so the reasons to which my trust should respond likewise conflict. It is this which gives rise to the dilemmas of trust. When these forms of value conflict, I must assign a priority to one form over the other, whether instrumental value over interpersonal, or the reverse. This may mean that I am forced to trade off the realisation of one form of value against the other. In other situations, I may refuse any trade-off, because my ascribing or recognising one form of value as having lexical priority over the other means that the former form of value must be realised regardless of the loss to the latter. So, we may identify two distinct forms of trust, according to the reasons to which trust responds, and the forms of value that it may promote. Recognising that the distinction between cognitive and non-cognitive trust does not accurately track the distinction between the reasons to which trust responds—which is the deeper—some different terms are needed. Anticipating the position developed in the next chapter, and on a suitably capacious and neutral understanding of the term, we may identify trust as being *evidence-constrained*, when it responds solely or primarily to the theoretical reasons one has to believe the person trusted to be trustworthy, and as being *beyond the evidence* when it responds solely or primarily to the interpersonal reasons one has to trust.

This yields an initial 'map' of the conceptual territory, illustrated by Figure 1, below. The map locates the two kinds of trust: that which is evidence-constrained and that which goes beyond the evidence, according to the kinds of value that they promote and the kinds of reasons to which they are responsive. This reflects the simple conceptual connection between reason and value; that there is reason to bring about or preserve value.

The map represents some normative claims about the conditions under which evidence-constrained trust, and trust beyond the evidence, is appropriate. When one is concerned about the interpersonal value of trust then, if one trusts another, one's trust may go beyond the evidence. Likewise, when one is concerned about the instrumental value of trust then, if one trusts another, one's trust should be constrained by the evidence. The blocked-out boxes exclude the converse possibilities. If one is concerned about the instrumental value of trust,

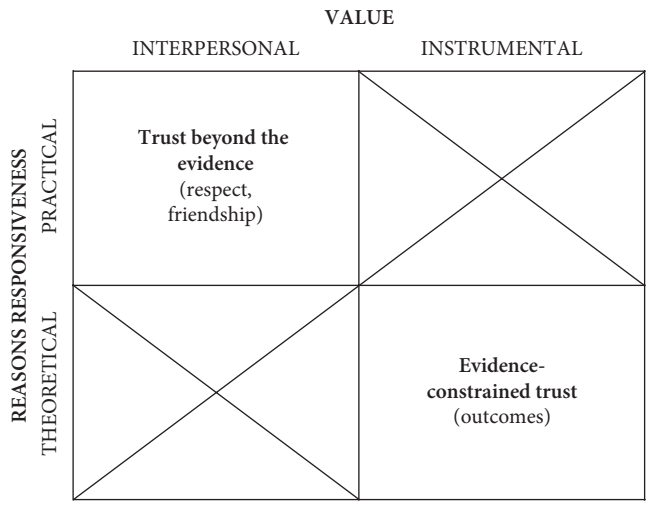

Figure 1 Trust's reasons and value—initial conception

do not trust solely or primarily for practical, interpersonal reasons. If one is concerned about the practical, interpersonal value of trust, do not trust solely or primarily for theoretical reasons. Pluralism about trust thereby reorients the enquiry away from analysis of the nature of trust, to the axiological question of what kinds of trust matter. The kind of trust that matters determines the reasons to which one's trust should be responsive, and these reasons depend on the forms of value that one's trust can realise. This is why axiology has been the starting point for this enquiry. (I have been helped in this by Jon Kvanvig, who helpfully developed some points from my earlier work (2012a); see his (2018: 24ff). Arnon Keren also identifies the conceptual connection between reason and value, seeing at as posing an 'unappreciated challenge' (2020b: 117–18), while the 'paradigm-based' methodological approach advocated by Friedemann Bieber and Juri Viehoff (2023) also has some affinities with that which starts from the axiology. The challenge that Bieber and Viehoff do not confront is whether one of the functions that trust may perform has some kind of priority or fundamentality over the other.) The axiology of trust thus grounds a moral enquiry, for it concerns who may or should aim to realise value for whom, by entering into which trusting relationships.

The central, substantive thesis of this book is that trust is normally evidence-constrained. The kind of trust that generally matters is that which is constrained by the evidence: trust that is grounded in the theoretical reasons that the trustor has to believe that the trusted will be trustworthy. It is not exclusively the kind of trust that matters, and I shall suggest that the principal exception is contexts when trust must be repaired after betrayal. But it is normally so. That trust is normally evidence-constrained comprises three sequential but separate claims.

First, it is a claim about what our practices of trust and distrust are functionally fitted for. Our practices are functionally fitted to ensure that we trust the trustworthy, and they do so in part by inducing people to be trustworthy. When our practices of trust are functioning properly, we trust according to the evidence. Enabling rational, evidence-constrained trust is the goal of our practices of trust.

Second, because these practices persist, evidence-constrained trust is a rational, default response to another's invitation to trust. This default response is defeasible, but, other things equal, one is rationally justified in starting from an assumption of trust.

Third, and applying in some situations but not in others, depending on how embedded norms of trustworthiness are in a given culture or sub-culture, that evidence-constrained trust is normal is a claim about how frequently one's default assumption that another is trustworthy is unrevised, and rationally so.

For brevity, I call the conjunction of these claims, which is summed up as the thesis that trust is normally evidence-constrained, *evidentialism about trust*. Evidentialism about trust does not imply that there are no forms of trust which respond to non-evidential considerations, or which go beyond the evidence. In the way I use the term here, stipulatively, evidentialism about trust is a claim about the conditions under which trust is normally rational. A better way of understanding the cognitive/non-cognitive dispute about trust is that it is really a dispute over evidentialism about trust: nearly all those who have advocated cognitivism have been inspired by the instinct that justifies evidentialism about trust; those who reject this instinct have usually, but not exclusively, been anti-evidentialist. A focus on the 'substrate' that realises trust is, however, misleading, as it omits the more important tie between reason and value. It is that tie which evidentialism about trust focuses on.

In arguing that trust is normal—and evidence-constrained trust at that—I aim to make sense of the superlative analogies that trust has often attracted. It is like the air we breathe (Bok 1978: 31; Hardwig 1991: 693; Baier 1994: 98); the cement or glue that holds society together (Acton 1974: 14; Govier 1998: 6; D'Cruz 2019: 935); 'part of the deep grammar of any society' (Hosking 2014: 22); or 'the bond of society' (Locke 1663 [1954]: 213). 'Without the general trust that people have in each other, society itself would disintegrate' (Simmel 1900 [2004]: 177–8; similarly, Reid 1788 [2010]: 334).

But the picture presented so far is too basic. The simple conceptual connection between reason and value asserts that there is reason to promote value: if acting in a given way will realise a particular form of value, to that extent there is reason so to act. It does not follow from this, however, that in order to realise a particular form of value, one must act for the reason that by doing so one will realise that value. The glory of being the fastest man in the world over 400m may give reason to run in the 400m race at the Olympics. But, while Eric Liddell achieved that glory, he did not run the race in order to achieve it. He ran because

in doing so he felt God's pleasure. Indeed, it may be that, for some forms of value, they can be realised by a given action or course of action only if the reason for which one acts is distinct from that of aiming to realise that value. For instance, the paradox of happiness claims that, if I start cooking seriously because I think it will make me happy, I will not get any happier. But if I start cooking seriously because I want to develop an excellence and explore new cuisine, I will find that doing so makes me happy. So, it should not be assumed without argument that the instrumental value of trust can be realised only by those trusting attitudes that are responsive to theoretical reasons, nor that the interpersonal value of trust can be realised only by those trusting attitudes that are responsive to practical reasons.

Precisely these possibilities have been proposed. I have noted the family of 'second-personal' accounts of trust, which take trust to be a belief about another's trustworthiness, but with that belief adopted for reasons deriving from the interpersonal relationship between the trustor and the trusted. It is not a belief about the practical reasons that obtain, with its justification being the evidence that one has for the existence of the practical reasons. Rather, respect for another's agency gives rise to reasons that justify the belief; what it is that the reasons count-in-favour-of is the belief. Interpersonal facts give these reasons their justificatory force, and what they justify is a belief. Nor is this for the William James-style reason that the belief would be useful to have, but because it is true. Interpersonal reasons are thus essentially second-personal: they arise in the context of people who relate to each other as subjects, addressing each other as 'you'. (A third-personal reason or attitude does not involve one person addressing another as 'you', but as 'he' or 'she'.) In proposing that trust should properly be based on interpersonal reasons only, and not on the evidence one has for believing that someone is trustworthy, so second-personal accounts of trust propose that the instrumental value of trust should be realised only by those trusting attitudes that are responsive to interpersonal reasons. Second-personal accounts thus break the connection, proposed in Figure 2, between the reasons to which trust is responsive, and the value that it realises.

I shall reject this and argue for a version of the contrary possibility. Evidence-constrained trust realises a kind of interpersonal value which is inaccessible for that which is beyond the evidence, and which is often, and indeed normally, what those trusted expect of trustors. It is the trust which is responsive to the theoretical reasons one has for believing that someone is trustworthy which properly promotes interpersonal value. This likewise breaks the connection between the reasons to which trust is responsive and the value that it realises, but in the opposite direction. I shall argue that there is a variety of grounds—deriving both from morality and from the obligation to promote truth, and which apply in important practical circumstances—for taking the kind of trust that matters to be one in which there is an overriding concern for outcomes, and so my concern is to

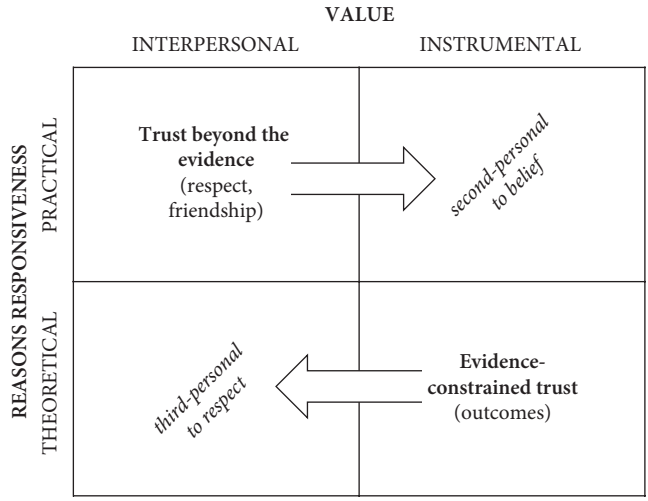

Figure 2 Trust's reasons and value—revised conception

ensure that I trust the trustworthy. Surprisingly, one of those concerns is the interpersonal demand to maximise interpersonal value.

These positions modify the map of the conceptual possibilities, as follows. Second-personal accounts propose that a trusting belief may be based on practical reasons, and yield instrumentally valuable outcomes. The account I defend proposes that trusting belief should be based on theoretical reasons, and that this also yields interpersonal value.

Pluralism about trust, then, does not eliminate the rivalry between different accounts of trust, whether cognitive, non-cognitive, or some other. Rather, it changes its terms. The advocate of a given account of trust ought to acknowledge that there are cases of trust, which are not amenable to their view, and which are genuine trusting attitudes. But this does not eliminate the possibility that some categories of cases have a kind of primacy, and so a different set of considerations must be appealed to. Cases are important not because they are data points which must be taken into account in formulating a definition of trust, counting otherwise as counterexamples. That method was always vulnerable to the charge that the remaining dispute is purely verbal, in which advocates of each may agree on the non-linguistic facts of the matter, and just disagree about whether the term 'trust' is appropriately used in a given context (for this test for verbal disputes, see Chalmers 2011). Rather, cases are important insofar as they help to elucidate the kinds of trust that matter. The philosophical task of elucidating trust thus turns on the kinds of value which trust may realise, when those kinds of trust matter, how that value is realised, and therefore what those reasons are to which trust responds.

To make progress on this is also to make progress on resolving the practical dilemmas of trust. Accordingly, I turn in the next chapter to making the case for

the practical importance of evidence-constrained trust. Before doing so, however, I address a final methodological question. What is the relation of 'axiology first' to conceptual engineering?

1.8 Is this Conceptual Engineering?

The case against conceptual analysis must still be made, in part because old habits die hard and, presumably for related reasons, in part because philosophical work on trust has not taken seriously the possibility that it takes plural forms. (As Carolyn McLeod observes, 'Rather than go in the direction of pluralism, however, most philosophers continue to debate what unifies trust in a way that makes it different from mere reliance'; 2020: §1.2.) But, in the context of more general discussions of philosophical methodology, many will be sympathetic to the claim that an alternative approach should be tried. At present, the most prominent alternative to conceptual analysis is the proposal that philosophers should engage in 'conceptual engineering'.

Conceptual engineering starts from a diagnosis of conceptual failure. A concept (or lexical term) which is of philosophical interest is identified as defective in some dimension. It may be defective because the principles that guide its use in natural language give rise to logical paradoxes and inconsistency. Truth and freedom are candidate examples (see Scharp 2013 and Van Inwagen 2008 respectively). Or, more weakly, there may be sufficiently divergent views about the correct intension of a term that different speakers arrive at markedly varied conclusions about its correct extension. Knowledge is a candidate (see Jackson 1998, 2005). Or, a concept may fail to pick out a natural kind, so being unhelpful for explanation; Clark and Chalmers make this claim for belief (1998: 12, 14; on concepts 'carving nature at its joints' more generally, see Sider 2011). Such a failure may also have deleterious social and political effects; gender and race are candidates (Haslanger 2000, 2020). Or the concepts belonging to a domain of discourse may assert that some set of facts exist, but may systematically misfire in doing so. Moral and religious concepts are often the target of such critique (Ayer 1946; Carnap 1958; Mackie 1977; Joyce 2006). Concepts may thus fail in a variety of ways.

Given a diagnosis of conceptual failure, the conceptual engineer proposes that the term or concept should remain in use, but that it should be suitably amended. Paradigmatically, the conceptual engineer keeps the lexical item but changes what she means by it; in changing its intension, so its extension is also changed. This is Sally Haslanger's approach to gender and race. Or, in the case of concepts which are inconsistent, the engineer may disambiguate, and propose that the old lexical item be replaced by two (or more) new ones. Kevin Scharp's proposal that truth should be replaced by the successor notions of 'ascending truth' and 'descending

truth' takes this form. Subscripts used to denote a cleaned-up, stipulated definition (e.g., privacy$_1$ and privacy$_2$) achieve the same effect (see Jackson 2017. Note, however, that Jackson takes this to be an example of conceptual analysis, not engineering. This suggests at minimum that the boundary between these activities is vague; perhaps 'conceptual engineering' is just a new term for more revisionary forms of conceptual analysis). While both strategies change the terms of discourse in some way, each still aims to ensure that users of the lexical term continue to talk about the same thing, at least in broad terms; the point is that it is the same topic under discussion, not a new one. While there is lively debate as to the details of what conceptual engineering should consist in (see, especially, the papers collected in Burgess et al. 2020), Herman Cappelen provides an ecumenical summary: 'Some concept is considered defective along some dimension, in some cases that deficiency can be ameliorated, and various proposals are made about how best to ameliorate' (2018: 33; his discussion there highlighted to me most of the examples above).

In amending a concept, or the meaning of a lexical term, the conceptual engineer's results are explicitly revisionary. She prescribes how a lexical term should be used, being thus engaged in 'conceptual activism' (Cappelen and Plunkett 2020: 4). In ambitious instances, the prescription may be aimed at the community of English speakers (or other natural language). Cappelen is cautious about the prospects of success here (2018: 72ff.) and this is surely realistic. Less ambitiously, proposals may be aimed at more specific communities of language users—such as the community of philosophers, logicians, or social scientists. In prescribing how a lexical term ought to be used, the conceptual engineer's results are immune to counterexample. The test for success is then whether the revised lexical item remedies the original defect, gives rise to no worse problems which would also need to be ameliorated, achieves both of these goals optimally, and ultimately whether it is adopted by the relevant linguistic community.

In addressing trust here, am I engaged in conceptual engineering? And, if not, why not? The conceptual engineer will find much of my argument so far to be congenial. I have argued that, if the concept of trust is taken at face value, we reason about it in ways that at minimum give rise to a philosophical problem, and indeed may be incoherent, with the attitude of trust simultaneously being subject to theoretical and practical forms of normativity, in ways that are not obviously coherent. To resolve this, I have disambiguated between forms of trust which are constrained by the evidence, and those which go beyond it. By determining whether one's trust is evidence-constrained or beyond the evidence, we may avoid purely verbal disputes. Disambiguating in this way also promotes clarity in our reasoning about trust, as it is thereby explicit what reasons determine whether trust is rational. Putative counterexamples do not threaten the project. The conceptual engineer may agree with all of this.

Nonetheless, my approach differs from conceptual engineering in some important respects. For one, I am not clear that the equivocity of 'trust' is an instance of conceptual failure. That trust takes plural forms, some of which are appropriate on some occasions but not on others, shows merely that natural language is flexible and adaptive. It is not confusion but the richness of English which ensures that 'trust' can be naturally applied in contexts as various as oligopolists agreeing to price-fix, parent-infant intimacy, testimony in a courtroom, and promises to a friend. While this may sometimes make it opaque as to how to reason correctly about trust, it also makes rhetoric and persuasion possible. Opacity in how to reason correctly about trust is the price that must be paid for that richness. Just because the term 'trust' may offend an analytic philosopher's sensibility, who prizes clarity, does not mean that her priorities should be legislated for, and so it is not my goal to prescribe how the term 'trust' should generally be used. I do not share the conceptual engineer's judgment on natural language, at least as regards to trust.

More significantly, my enquiry here goes beyond that of the conceptual engineer. Her goal is to refurbish a concept or lexical term, successfully ameliorating some deficiency, with the refurbished concept then widely adopted. In contrast, I aim to get clear on the concept of trust because doing so helps me to address the question, why should I trust? My concern is with when we should place trust, and why. Some clarificatory work is required to address this, but the clarificatory work is not itself the point. The goal is to understand what trust is, and why we should place it. Addressing this principally involves first-order questions of morality and agency, not meta-questions of how better to define trust. And the starting point for these first-order questions is the ways in which trust is valuable.