A Telic Theory of Trust

J. Adam Carter

Contents

Preface

This book uses a telic normative framework in order to explain a range of phenomena related to *trust*, including its nature and varieties, the evaluative norms that govern good trusting and distrusting (both implicit and deliberative), how trust relates to vulnerability, risk, and monitoring, as well as to trustworthiness and, more generally, to our practices of cooperation.

The overarching theory of trust is centred around a simple idea: that trust is a *performative kind* and that the evaluative normativity of trust is a special case of the evaluative normativity of *aimed* attempts generally. Chapter One motivates the need for this kind of proposal by showing why traditional views of trust (e.g., doxastic, non-doxastic, and conative accounts of trust) lack the resources to satisfactorily explain what *good trusting* consists in.

Chapter Two lays out the key elements of the 'trust as performance' view, by (i) explaining the sense in which trusting is a constitutively aimed performance; and then (ii) showing how we can fruitfully assimilate the evaluative norms of trusting to telic norms of success, competence, and aptness which are applicable to aimed attempts as such. The core framework developed in this chapter – which will be extended further and precisified as we go on – supplies us with a fresh lens to view traditional philosophical problems about trust, and also to explore entirely new avenues that will be taken up in subsequent chapters.

Chapter Three distinguishes between two fundamental species of mature

human trust – *implicit trust* and *deliberative trust* – and then shows that both varieties have their own performative analogues in the case of *distrust* – viz., implicit and deliberative *distrust*. As we'll see, key to the theory of distrust that is then developed in this chapter is the distinction between wide-scoped and narrow-scoped (i.e., aimed) forbearance from trusting, with reference to which evaluative norms of success, competence and aptness are applicable to narrow-scoped (but not wide-scoped) distrust.

Chapter Four takes as a starting point the distinction between implicit and deliberative trust (from Chapter 3) and asks: under what conditions is one kind species of trust more appropriate than the other, and what kinds of considerations determine this? In answering this question, a broader normative distinction is drawn – one that delineates the kind of (telic) normativity applicable to trusting and distrusting and shows how it contrasts with a broader normativity applicable to the ethics of cooperation. In framing this distinction, an analogy is drawn to how norms that apply to believing stand in relation to norms that govern which inquiries to take up.

Chapter Five zeroes in on deliberative trusting, which involves on the trustor's part (intentionally) aiming not just at trusting *successfully* (which is the constitutive aim of *implicit* trust, viz., that the trustee take care of things as entrusted), but at trusting *aptly*. The deliberative truster's attainment of *this* more demanding aim may then itself be (or fail to be) successful, competent and apt; when it *is* apt, the resulting trust is then *convictively apt*, of the highest telic quality. This chapter explores, in detail, what is involved on the part of the trustor in the *apt attainment* of apt trust. An answer is developed in two parts, through (i) the development of a substantive view of first- and second- order trusting *competences*, as well as a structural theory of how (when deliberative trust is apt) first-order and second-order aptness is 'connected', which will be, on the view proposed, only when the former is appropriately *based* on the latter.

Chapter Six asks turns to the question of what kinds of risks of *inaptness* to trust the convictively apt trustor can *non-negligently* ignore. An initial answer to this question is suggested in Ernest Sosa's (2021) recent work

on telic normativity, an answer that to our question that would be framed in terms of what Sosa calls 'background conditions'. While *prima facie* promising, such a strategy is shown to come up short. A very different answer is then motivated in developed, one for which the irreducibly normative notion of 'de minimis' risk plays a central role. The conclusion that is ultimately reached and then defended is that the convictively apt truster can't non-negligently ignore what are termed *cooperative-relative risks* to the inaptness of trust, except when these risks count as *de minimis* with reference to cooperation-sustaining rules.

Chapter Seven takes up two foundational questions in the philosophy of trust: (i) In what sense does trusting essentially involve subjecting oneself to risk of betrayal?; and (ii) in what sense is monitoring for risks of betrayal incompatible with trusting? These questions have traditionally been pursued independently from one another. It will be shown that they are much more closely connected than has been appreciated. The central objective will be to demonstrate how the telic-normative approach to theorising about trust (and its theoretical cognates) developed so far can be used to answer both questions in a principled way, one that reveals a deep connection between not just the questions themselves, but also between the concepts of vulnerability, monitoring, and *de minimis* risk.

Chapter Eight shows how *therapeutic trust* – roughly: trust that aims at *trust-building* – fits into the picture developed so far. Therapeutic trust is a vexed topic in the philosophy of trust, not least due to the fact that, in cases of therapeutic trust, the trustor's attitude towards the trustee's reliability is characteristically much *less* optimistic than in standard non-therapeutic cases, with risk of betrayal typically but not always higher. Several strategies that philosophers of trust have appealed to in order to make sense of therapeutic trust and its relationship with ordinary non-therapeutic trust are considered and shown to be problematic. A new way of thinking about therapeutic trust is then developed, one that avoids the problems facing the other three views while at the same time offering its own additional advantages (especially, that of explaining why therapeutic trust is *good* when it is). Key to the positive proposal is a recognition of two very different species of therapeutic trust: what I call *default* thera-

peutic trust and the more philosophically interesting *overriding* therapeutic trust. In the latter case, one's trusting constitutively aims not at mere successful trust, nor at the mere building of trust, but at building trust *through* successful trust.

Chapter Nine concludes by reorienting our view of the relationship between trust and trustworthiness, by locating both within a broader picture that captures largely overlooked symmetries on both the trustor's and trustee's side of a cooperative exchange. The view I defend here takes good cooperation as a theoretical starting point; on the view proposed, cooperation between trustor and trustee is working well when achievements in trust and responding to trust are matched on both sides of the trust exchange. In a bit more detail, the trustor 'matches' her achievement in trusting (an achievement in fitting reliance to reciprocity) with the trustee's achievement in responding to trust (an achievement in fitting reciprocity to reliance). From this starting point, we can then appreciate *symmetrical* ways that the trustor and trustee can (respectively) fall short, by violating what are shown to be symmetrical evaluative norms – of success, competence and aptness – that regulate the attempts made by both trustor and trustee. The overall picture is shown to have important advantages over the received way of theorising about how trust stands to trustworthiness, and it clears the way – by identifying key questions that have been obscured – to making further progress.

The book is written largely from scratch, although some chapters are expansions of ideas developed in some recent articles. Material from Chapters 1-3 draws in various places from my paper 'Trust as Performance', (2022, *Philosophical Issues: A Supplement to Noûs*), and Chapter 2 in particular is based around some ideas that I first developed in 'On Behalf of a Bi-Level Account of Trust', *Philosophical Studies* (2020b). Finally, the account of the relationship between trust, risk, and negligence advanced in Chapter Six draws from the more general theory of *de minimis* risk that first appeared in a paper principally about virtue epistemology, 'De Minimis Normativism: a New Theory of Full Aptness', The *Philosophical Quarterly* (2020a) 71:1: 16–36.

This book was written as part of the Leverhulme-funded 'A Virtue Epistemology of Trust' (#RPG-2019-302) project, which is hosted by the University of Glasgow's COGITO Epistemology Research Centre, and I'm grateful to the Leverhulme Trust for supporting this research. I also want to thank my collaborators on this project, Christoph Kelp, Mona Simion, Emma C. Gordon, and Ísak Andri Ólafsson, as well as my other friends and colleagues at COGITO and in Philosophy at the University of Glasgow.

Chapter 1

What is Good Trusting?

1.1 Introduction

Trust is indispensable to the success of almost every kind of coordinated human activity, from politics and business to sport and scientific research. It is accordingly important that we know how to do it *well* – and how to avoid doing it badly.

But the question of what it is to trust well is not easily separable from the question of what kind of thing trust is. And the matter of what kind of thing trust is — viz., whether it is a belief[1], or some non-doxastic attitude or stance (of one kind or another, perhaps *affective*[2], perhaps *conative*[3]) is divisive.[4] It is also, as Jon Kvanvig (2016) puts it, somewhat 'stupefying' (2016, 8) given that the various distinct things trust has been identified with are also very different from *each other*. In response to considering the menu of disparate options about the nature of trust in the literature,

[1] See, e.g., Hardin (2002), Hieronymi (2008), and McMyler (2011). For a more moderate doxastic account, see Keren (2014, 2020).

[2] Some notable examples here include Baier (1986) and Jones (1996).

[3] See Holton (1994).

[4] For a representative picture of this divisiveness, see, e.g., Carter and Simon (2020) and Faulkner and Simpson (2017).

Bernd Lahno (2004) writes, 'any adequate theory of trust must include behavioral, cognitive and affective dimensions or aspects' (2004, 30) by contrast, and more pessimistically, Thomas W. Simpson (2012) suggests that 'There is a strong *prima facie* case for supposing that there is no single phenomenon that 'trust' refers to, nor that our folk concept has determinate rules of use' (2012, 551).

Unsurprisingly, given the disparity of opinion about the ontology of trust, what we find in the philosophy of trust are various incompatible (and starkly different) pictures of the *evaluative normativity* of trust – viz., of what the relevant standards are that good trusting, *as such*, should be expected to meet.[5]

Generally speaking, evaluative norms – unlike prescriptive norms, which prescribe conduct – regulate what it takes for a token of a particular type of thing to be good or bad with regard to its type, where the 'goodness' here is *attributive* in Geach's (1956) sense – viz., the sense in which a sharp knife is a good knife, *qua* knife, regardless of whether it is good or bad *simpliciter*. (Likewise, in this sense, a known belief is a good belief, regardless of whether it would be good or bad *simpliciter* – viz., as it would be were the content of the knowledge instructions for igniting a terrible bomb.)[6]

Without a defensible picture of plausible evaluative norms for trusting, we're poorly positioned to say when trusting is good or skilled *as an instance of trusting*; we would be relegated – at least in our evaluative theorising – to seek out conditions under which trust is good or bad *simpliciter* – e.g., by investigating, like we might with anything else, when it paradigmatically leads to good (and bad) consequences, and if so what they happen to be.[7] It should be no surprise that philosophers of trust

[5]For overviews, see Carter and Simon (2020, sec. 2) , McLeod (2020, sec. 3); see also Faulkner (2010, 2014), Frost-Arnold (2014), Hieronymi (2008), and Hinchman (2020).

[6]For a helpful overview of the prescriptive/evaluative norm distinction, with reference to attributive as opposed to predicative goodness, see McHugh (2012, 22) and, as this distinction applies to belief specifically, Simion, Kelp, and Ghijsen (2016, 384–6).

[7]We would be in poor shape philosophically if we asked only how belief is good *simpliciter* and not what makes something good *qua* belief (then there would be very lit-

have attempted to go further than this – i.e., further than exploring (e.g., as social psychologists have[8]) how trust might be good because, e.g., it enhances cooperation[9] – typically by first taking some kind of stand on the ontology of trust – a stance on what kind of thing it is – which is then used as a kind of 'blueprint' for thinking about good trusting *as such*.

According to *doxastic accounts of trust*, trust is essentially a kind of *belief*, a belief about the object of trust, e.g., that the trustee will take care of things as entrusted.[10] Good trusting, on simple doxastic accounts of trust on which belief of a particular sort is necessary and sufficient for trust, will just *be* a kind of good believing; that is, it will be an instance of the very thing – *belief* – whose attributive goodness it is always appropriate to assess by looking at its rationality, asking whether it's true or known, whether it coheres with other justified beliefs, whether its production manifested epistemic virtues, etc.[11] Rationality, reliability, truth, coherence, knowledge, etc., – paradigmatic evaluative norms of belief, as such – are (on these views) also norms that would regulate what counts as good trusting, in so far as trusting is believing.[12] Moreover, if trusting is a kind of believing, then what the proponent of the doxastic account of trust tells us counts as good trusting will need to reflect any constraints on beliefs as such, including that we cannot bring them about via direct control.

---

tle epistemology!). The same goes for emotion; it is of philosophical importance what makes an emotion good or bad as such, not merely simpliciter.

[8] See, for example, Berg, Dickhaut, and McCabe (1995) and Braynov and Sandholm (2002). For a philosophical incorporation of trust's importance to decision-theoretic cooperative behaviour, see Faulkner (2011, Ch. 1).

[9] See Hardin (2002). Cf. Cook, Hardin, and Levi (2005, Ch. 1).

[10] For a recent and helpful overview of doxastic accounts of trust, see Keren (2020).

[11] These are just some representative evaluative norms of belief. For related discussion, including of prescriptive norms of belief, see Simion, Kelp, and Ghijsen (2016), Whiting (2013a), Benton (2014), McHugh (2012), Shah and Velleman (2005), and Gibbons (2013). For criticism, see Glüer and Wikforss (2009) and Papineau (2013).

[12] As Karen Jones (1996, 2–5) captures this idea, what we say about the nature of trust, viz., whether it is a belief, constrains what we say about the rationality of trust, given that belief already has clearly defined standards of rationality. See also Keren (2014) for related discussion.

According to *non-doxastic accounts of trust*, it's false that trust is a kind of belief, even if trusting sometimes accompanies belief. And by extension, it is false that some type of good believing is what it is that the evaluative normativity of trust should be thought of regulating. But here things complicate quickly. Some non-doxastic accounts of trust maintain that trust is essentially an affective attitude, or an emotion – some even hold that whatever affective attitude it is, it *must* not be accompanied with belief. Other non-doxastic accounts of trust maintain that trust is a non-doxastic, non-affective conative attitude – e.g., a kind of moral stance – and so neither a belief nor a non-doxastic affective state.

In light of the above, it looks very much like the endeavour of getting a clear grip on what *good trusting* involves threatens to fall into disarray; after all, *the evaluative norms of belief bear little to no resemblance to the evaluative norms of, e.g., hoping, being optimistic, adopting moral stances, etc*.[13] But perhaps all is not as bad as it seems.

Here is the plan for what follows. §1.2 demonstrates several insuperable problems for accounts of the evaluative normativity of trust that fall out of doxastic accounts of the nature of trust. §1.3 shows that different problems arise for attempting to extract an account of good trusting by looking at the attributive goodness of various non-doxastic attitudes (both affective and conative) that have been identified with trust.

1.2 Good trusting as good believing: the doxastic account

So *is* trust a kind of belief? Let's sharpen this initial question in two ways; first, by bracketing two-place trust (i.e., X trusts Y) and focusing on *three-*

[13] Consider, for example, some fundamental disanalogies that will bear on what the respective evaluative norms will look like. There is a disanalogy between belief, and emotions on the one hand, and the adoption of a moral stance, on the other, when it comes to voluntariness (e.g., Alston 1989). However, adopting a moral stance lines up with emotion and other affective attitudes when it comes to direction of fit (see, e.g., Humberstone 1992).

place-trust: schematically (S trusts X to ϕ).[14] Second, for simplicity, let's consider just cases of *interpersonal* three-place-trust, which have been a central focal point in the philosophy of trust, and which involve one person trusting another person to – in a broad sense – take care of something, ϕ, as entrusted[15], and which further involves (unlike in cases of *mere* reliance) subjecting oneself to the possibility of betrayal.[16]

In the specific case of *testimonial trust* – of special interest in social epistemology – when a hearer forms a testimony-based belief on a speaker's say-so[17], the something she trusts the speaker for is the truth, or perhaps knowledge[18], of what she says. On a simple way of thinking of the relationship between testimonial trust and three-place interpersonal trust generally, the former is just an instance of the latter, an instance where 'the truth' is plugged in for ϕ in the schema, and which betraying the hearer's trust involves misinforming her.

With these caveats aside, let's now consider the *strong doxastic account* of three-place interpersonal trust (hereafter, trust) according to which trust is essentially a belief, viz., a belief that the trustee will take care of things as entrusted.

[14] The distinction between two- and three-place trust was drawn initially by Horsburgh (1960). According to one popular way of thinking about relationship between two- and three-place trust, three-place trust is fundamental in the sense that two-place trust is explained in terms of it. For some representative examples of three-place fundamentalism, see, e.g., Baier (1986), Holton (1994), Jones (1996), Faulkner (2007), Hieronymi (2008), and Hawley (2014). Cf., Faulkner (2015).

[15] The locution 'as entrusted' is meant to encompass views on which the trustee counts as taking care of things as entrusted only if doing so in a particular way, including, e.g., out of goodwill (Baier 1986; Jones 1996) or in conjunction with a belief that one is so committed (e.g., Hawley 2014). The present proposal – which is theoretically neutral on this point – is compatible with opting for either such kind of gloss.

[16] As Annette Baier (1986) puts it, interpersonal trust involves subjecting oneself 'necessarily to the limits' of another's goodwill (1986, 235) and in a way that differs from the kind of reliance we place in mere objects. For related discussion, see McMyler (McMyler 2011, 124).

[17] For an overview of what qualifies as a testimony-based belief, see, e.g., Graham (2016, 172–3).

[18] See, e.g., Kelp (2018) for a view friendly to this suggestion.

This kind of proposal – variations of which have been defended by Russell Hardin (2002), Pamela Hieronymi (2008), and Benjamin McMyler (2011) – is strong because it takes believing something about the object of trust to be type-identical with trusting. And there are some marks in its favour. For one thing, in paradigmatic cases of testimonial trust, the hearer trusts what the speaker says only if the hearer *believes* that the speaker has told the truth. And, more generally, as Hieronymi (2008) notes, if you entrust any kind of task to someone while believing they won't do the thing, it seems you're not *really* trusting them to do it.[19]

Unfortunately, regardless of whatever else we might say for or against a strong doxastic account of trust, there are serious problems for the idea that trusting *well* is principally a matter of believing well, viz., of holding rational, reliable[20], true, coherent, etc., beliefs about the object of trust – and regardless of what features in the content of that belief (i.e., that the trustee will encapsulate the truster's interests[21], or prove trustworthy

[19] In support of this line of thought – viz., that one's trust tracks one's belief that the trustee will prove trustworthy – Hieronymi offers the following case-pair involving the betrayal of a secret. 'SECRETS: Consider two cases. In one, I fully believe you are trustworthy; in the other, I have doubts about your trustworthiness, but, for other reasons (perhaps to build trust in our relationship, perhaps because I think friends should trust one another, or perhaps simply because I have no better alternative), I decide to tell you my secret. Suppose that, in both cases, you spill the beans, and that you do so in the same circumstances, for the same reasons (2008, 230)'. According to Hieronymi, once we hold fixed both (i) the 'importance of the good entrusted'; and (ii) 'the wrongness of the violation,' then: '[...] it seems plausible that one's degree of vulnerability to betrayal tracks one's degree of trusting belief ... further, this seems to be because, in the second case, there was less trust to betray (2008, 230–1)'. If the degree of one's trust is, as Hieronymi thinks, positively correlated with the degree of one's belief the trustee will prove trustworthy, then this counts in favour of the strong doxastic account, which would straightforwardly explain this correlation. See Chapter 6 for additional discussion of this case.

[20] Of course, a norm of reliability will be an example of an 'externalist' norm on good trusting just as reliability is a paradigmatically externalist norm on good believing, in that that in virtue of which one satisfies the norm needn't be reflectively accessible to the truster. See McLeod (2002, 91–100) for discussion.

[21] (Hardin 2002).

out of goodwill[22], etc.) To appreciate this point, consider that any belief, as such, is better than it would be otherwise if it complies with the paradigmatic evaluative epistemic norms of belief, e.g., norms that hold that beliefs ought to be supported by evidence and known.[23]

However, (a) complying with a standard evidence norm[24] (i.e., on which evidential support improves a belief's quality) fails to improve trust for the reason that there is a constitutive tension between trusting and complying with evidential norms; and, second, (b) complying with the knowledge norm, specifically, undermines (or: moots) trust (rather than improves it) because it eliminates vulnerability.

On the first point, consider the following case due to Jeremy Wanderer and Leo Townsend (2013):

> PARANOID PARENT. A paranoid parent [...] organises a babysitter for their child, and then proceeds to spend the evening out monitoring their babysitter's antics remotely, via a 'nanny-cam'. The paranoid parent is not only a lousy date, but also a lousy trustor; in performing the seemingly rational act of broadening the evidential base relevant to her judgments of trustworthiness, she is, precisely, failing to trust the babysitter (2013, 1).

[22] (Jones 1996; Baier 1986).

[23] Because we are discussing evaluative rather than prescriptive norms here, the 'ought' should be read as a kind of 'ought to be' rather than an 'ought to do' – viz., in the sense that a good knife ought to be sharp. For some useful discussion of the difference here, see, along with McHugh (2012, 22) and Simion, Kelp, and Ghijsen (2016, 384–6) and Schroeder (2011, 5–8) for evaluative norms featuring 'ought' claims specifically.

[24] As Wanderer and Townsend (2013) put it, 'No matter how the norms of Evidentialism are construed, trust invariably seems to stand in tension with them' (2013, 7). See also Booth (2007). For a simple expression of an evaluative evidence norm on belief, take the following from Jonathan Adler (1999): 'One's believing that p is proper (i.e. in accord with the concept of belief) if and only if one's evidence establishes that p is true' (1999, 51). Alternatively, see Richard Feldman (2000): 'When adopting (or maintaining) an attitude towards a proposition, p, a person maximises epistemic value by adopting (or maintaining) a rational attitude towards p' (2000, 685).

The kind of belief that the proponent of a strong doxastic account identifies with trusting is such that the paranoid parent improves its quality *qua* belief – by strengthening the evidence basis for the belief – only by *at the same time* doing something that apparently undermines her trust. As Wanderer and Townsend put it, cases like PARANOID PARENT indicate that part of 'what it is to trust' is to *refrain* from complying with evidence norms on belief (2013, 2).[25] So, you can't by complying with such norms thereby trust *better*.

Likewise, consider knowledge as a norm governing what counts as good belief – a position embraced by, e.g., 'knowledge-firsters.'[26] If the kind of belief that the proponent of a strong doxastic account identifies with trusting satisfies the knowledge norm, it arguably ceases *thereby* to qualify as trust. But this is not because of anything to do with trust's relationship to evidence. Rather, it is to do with a constitutive tension between trusting and *securing* an outcome. If you have – put roughly – some kind of 'guarantee' that it is impossible for X to betray your trust, then as the thought goes, you are thereby no longer trusting them to do anything.

This idea, viz., that trusting essentially involves subjecting oneself willingly to non-negligible vulnerability – at the very least, as Baier (1986, 244) notes, to the limits of another's goodwill, though also to the lim-

[25] This idea is sometimes captured in terms of a *prima facie* incompatibility between trusting and monitoring. As Baier (1986) vividly expresses the idea 'Trust is a fragile plant [...] which may not endure inspection of its roots, even when they were, before inspection, quite healthy' (1986, 260). Belief, by contrast, not only withstands, but *improves* through inspection of its roots.

[26] See, e.g., Williamson (2013, 5). A typical way that this view is defended by knowledge-first philosophers involves two steps: first, there is a defence of the view that justification is the primary norm governing belief; and then, there is the further and crucial step that a involves a defence of the thesis that a belief is justified if and only if it is known. See Williamson (2016) and Sutton (2007) for representative statements of this idea. For an overview, see Benton (2014).

its of her competence[27] – is mostly uncontroversial.[28] As Hardin (1992) summarises:

> As virtually all writers on the subject agree, trust involves giving discretion to another to affect one's interests. This move is inherently subject to the risk that the other will abuse the power of discretion (1992, 507).

But then if trust essentially involves rendering oneself vulnerable to betrayal, it is hard to see how – by coming to *know* that the trustee has taken care of things as entrusted – the trustor has improved any trust she might have had prior to acquiring this knowledge as opposed to having simply rendered her trust moot.

Here is another problem for the idea that we can – as the proponent of the strong doxastic account must permit – profitably defend an account of what it is to trust well that is constrained by facts about what it is to believe well. The reasoning behind this second problem – call it the *argument from voluntariness* – goes as follows. Belief is never subject to arbitrary voluntary control. And that means that norms of believing never regulate what it takes for something subject to arbitrary control to be good or bad with regard to belief. If trust is a species of belief, then norms of good believing are *always* applicable to trust. Some cases of trust are subject to arbitrary voluntary control. So, norms of belief sometimes do not regulate trust.

The idea that belief is not subject to arbitrary voluntary control is platitudinous, and is central to marking the difference between 'belief' and

[27] Perhaps also: to the trustee's capacity to remain in conditions conducive to her cooperation, even if we hold fixed *both* good will and competence. For example, you might entrust someone to repay a debt. They are capable and willing, but fail to repay the debt due to an unexpected natural catastrophe.

[28] For various expressions of this idea, see, along with Hardin (1992), e.g., Baier (1986, 244), McLeod (2020, sec. 1), Nickel and Vaesen (Nickel and Vaesen 2012, 861–2), Carter (2020b, 2301, 2318–9), Carter and Simion (2020, sec. 1.a), Becker (1996, 45, 49). Dasgupta (1988, 67–68), Dormandy (2020, 241–2), Kirton (forthcoming), O'Neil (2017, 70–72), Potter (2020, 244), and Hinchman (2017). Cf., Pettit (1995, 208).

'make belief.'[29] However, the idea that trust is, at least sometimes, subject to arbitrary voluntary control is something we could give up only on pain of failing to countenance *therapeutic trust* – viz., where one trusts (e.g., a teenager with no established track record of reliability) with the intended aim of bringing about (or increasing) trustworthiess.[30] To the extent that therapeutic trust is voluntary in exactly the sense in which belief is not, as the worry goes, good believing does not provide us any kind of blueprint for good trusting.[31]

A third and final argument against the assimilation of good trusting to good believing focuses on cases of *trusting through doxastic suspension*. For example, suppose you see your friend holding a bloody knife, standing

[29] For a classic defence of this position, see Williams (1970). See also Scott-Kakures (1994) and, for an overview, Vitz (2008). It is worth noting, as Heil (1983, 355–6) points out, that we often use language to talk about belief – such as duty-based language – that seems to imply a kind of voluntariness. Such language is found in Descartes – who seems in the *Fourth Meditation* to speak of belief through the will (see Cottingham 2002, 352–5) but it is also used widely by contemporary writers (e.g., BonJour 1980, 60–61) who say (in various ways) that affirmation about whether something is so without suitable 'inspection' of what one is affirming violates an epistemic duty, perhaps a duty to be epistemically responsible in one's belief formation (Kornblith 1983, 34–37). It is a mistake though to think that this kind of talk implicates the idea Williams challenges of having direct arbitrary control over belief (though, cf., Vitz 2010). One helpful way to see why is to consider how *judging* whether something is so is both an intentional action but not subject to arbitrary control. For a detailed discussion on this point, see Sosa (2021, 32, 105 n. 59) and Sosa (2017, 88–91).

[30] The first notable discussion of therapeutic trust as a species of trust is due to Horsburgh (1960, 5, 7–8, 12). See also Jones (2004, 5–7) and Frost-Arnold (2014, 1960–3).

[31] Therapeutic trust cases (see Chapter 6) also raise another problem for strong doxastic accounts, which concerns the relationship between trust and expectation. As Peter Railton (2014) puts it, 'Belief that p is a degree of confidence [...] in a representation, p, that gives rise to and regulates a degree of expectation that things are or will be as p portrays them' (2014, 145). Therapeutic trust, however, often does not involve any such expectation. What this means then, in addition to that the strong doxastic account is problematic in its own right, is that an account of good trusting can't be an account of something that *essentially* involves an expectation. Granted, some proponents of doxastic accounts have sidestepped entirely issues to do with therapeutic trust by biting the bullet (Hieronymi 2008) and simply denying that therapeutic trust is genuine trust. See, however, Frost-Arnold (2014) for criticism of this strategy.

next to a body, after which you accuse your friend of the murder. Your friend, appreciating how overwhelming their guilt looks in light of the evidence, implores you to not rush to judgement until you hear the full story. They ask you to *trust them* – and wait until you hear the full explanation before drawing any conclusions whatsoever about what you've just seen.

Suppose you do then trust your friend. In doing so you are explicitly *not* forming a belief about whether they will prove trustworthy in this case, nor are you forming a belief that they will or will not betray your trust in any way. You trust through doxastic suspension, such that the suspension from belief *constitutes* your trusting. Belief, as a kind of affirming, categorically precludes suspension.[32] Thus, the trust you place in your friend here isn't something we could account for as good or bad trust in terms of norms governing belief, e.g., by asking if the belief counts as knowledge, or if it is rational.

It's looking like trusting well doesn't line up very well with believing well. Is it of any use to weaken the account – such that trusting well might be understood as a function of satisfying *at least* (some) norm of belief along with perhaps some other norms? An initial hurdle, of course, is that we've already seen that there are reasons to think good belief (or any kind of belief for that matter) may not be necessary for trusting, well or otherwise. But, even if those problems could be overcome, we'd need to know exactly what those other norms are. After all, the good cracking of an egg, even if necessary for a good cake, provides us little clue for what the standard is for a good cake.[33] Let's now look at some non-doxastic norms, norms on trust whose motivation is sourced in very different, *non-doxastic* approaches to

[32]For discussion, see Turri (2012) and Carter (2018) For an extended treatment of forbearance, see Sosa (Sosa 2021, Ch. 3).

[33]Granted, if belief is necessary for trust, then we will know that when a norm on good believing is not complied with when one trusts, then the quality of that trust is to that extent defective. This is information about the evaluative normativity of trust, but it is not a useful guide to what good trusting involves, unless we know *in addition* what other norms, beyond norms of belief, would need to be complied with to trust well.

the ontology of trust.[34]

1.3 Alternative norms on good trusting: non-doxastic accounts

1.3.1 Good trusting as good affect

Non-doxastic accounts of the nature of trust embrace a negative and a positive thesis. The negative thesis, common to all non-doxastic accounts, is just the denial of the claim that belief (i.e., that the trustee will prove trustworthy) is central to trust of the three-place interpersonal variety.

What distinguishes non-doxastic accounts from each other is the positive theses they maintain about the nature of trust. Perhaps the most common non-doxastic proposal-type maintains that trust is, rather than a belief, an *affective attitude*.[35]

On Karen Jones's (1996) influential account, that affective attitude is *optimism* 'that the goodwill and competence of another will extend to cover the domain of our interaction with her' (1996, 4) .[36] For Lawrence Becker (1996) the relevant affective attitude is, instead, 'a

[34] Might it be worth one final push for the line that norms of good trusting are doxastic norms – by pointing out that (i) trusting testimony is a paradigmatic form of trust; and (ii) that trusting testimony is something you do if and only if you actually uptake what the speaker says? The answer here is 'no'. The reason is that it might just be that testimonial trust incidentally involves belief because of what testimony demands but that such belief isn't essential to trust *as such*. Compare: trusting someone with a secret involves sharing the secret with them; but trust doesn't essentially involve anything like this. We should thus be wary about generalising norms of trust, as such, from the testimonial case where believing the word of another is the standard shape that trusting takes.

[35] For some representative defences of this kind of view, see de Sousa (1987), Calhoun (1984), Rorty (1980), and Lahno (2001).

[36] Jones clarifies that this kind of optimism she takes to be central to trust needn't involve any tendency to – as optimism is often taken to imply one would do – 'look on the bright side', given that, in the context of a very difficult joint task, you could trust someone, through your optimism directed at their goodwill, without any optimism about the success of the task (Jones 1996, 6).

sense of security about other people's benevolence, conscientiousness, and reciprocity' (1996, 43) . For Victoria McGeer (2008), it is a kind of 'hope', for Guido Möllering (2001), a 'leap of faith'.

Despite their differences, each of these affective attitude accounts of trust implies that any standard for good trust – a standard often captured by our talk of 'justified trust' – will take a different shape from the standards we expect good or justified belief to meet. Here's McGeer (2008):

> The question of whether our trust can be justified, then, becomes a question of whether certain feelings towards others can be justified, which is not to say they can't be, but rather that their justification conditions are different from, and perhaps not as stringent as, those on belief or on belief-based predictions of reliability (2008, 241).

Likewise, as Jones (1996) puts it:

> we can be justified in trusting even when we would not be justified in predicting a favourable action on the part of the one trusted. Our evidence for trusting need not be as great as the evidence required for a corresponding justified prediction. In this respect trusting is *more like hoping than like predicting* (1996, 15, my italics).

As we'll see, there is a serious problem for the thought that good trusting is a species of good hope. And the reasons why this is so generalise to other positively valenced affective attitudes about others actions and intentions – including optimism and a sense of security–, such that we should be sceptical that good trusting is something we can encapsulate under the heading of good affect.

So when is an instance of *hope* good hope? An initial reaction here might be as follows: "hope is good if it makes you better off with respect to getting the thing you're hoping for." This line of thought, however, faces a challenge in the form of Luc Bovens' (1999) decision-theoretic 'dominance argument' against the value of hoping.

Just suppose you that, for some projected state of the world, σ you have a choice between (i) hoping for it or (ii) not hoping for it. The projected state, σ will either come about or it will not. If it *does not* come about, then you would have been worse off having hoped than not having hoped, given that you will then be left with a greater sense of frustration after hoping than after not hoping. But suppose that σ *does* come about. Are you thereby better of having hoped? Perhaps, the contrary is the case. As Bovens puts it:

> [...] is there anything to be gained from having hoped for it? In hoping for something, I tend to fill in the contours in the brightest colors. Suppose that my hopes come true, but not precisely in the bright colors that I had pictured. Had I not hoped for anything, I would have been delighted. But having hoped as I have, I experience a sense of frustration rather than satisfaction (1999, 670).

But then the idea is that, whether the state of the world *does or does not come about*, I am always better off not having hoped for it rather than having hoped for it. Hence, by dominance, I should not hope.

Let us zoom out for a moment to see why this line of reasoning looks like it poses a problem for the thought that good trusting is a kind of good hoping. An *ex ante* theoretical desideratum on any account of what good trusting involves is that the value of good trust isn't 'swamped' by the value of trust. However, the dominance argument seems to show that the idea that good trusting is a kind of good hoping will inevitably fail to meet this desideratum. This is because, if the dominance argument goes through, it looks like the value of hoping will *trivially* swamp the value of good hoping, given that hoping will *never* be better than not hoping. Thus, as the thought goes, it is not plausible that good trusting is a kind of good hoping.

But let's not get carried away. Maybe there is a simple reply to the dominance argument, which is as follows: the dominance argument goes through only if we are assuming no causal dependency between states of

the world and choices.[37] But, as the reply continues, there often *is* such dependency. When you are tied up and locked in a room, a hopeful as opposed to a defeatist attitude might cause you to explore alternative ways for you to escape which you wouldn't otherwise have considered, the exploring of which might then lead you to stumble upon a way that actually works. And, *mutatis mutandis*, for many more mundane circumstances where exploring additional alternatives is instrumentally valuable in achieving a hoped-for objective.

Unfortunately, the above reply will be of little use for the proponent of 'good trusting is a kind of good hoping'. This is because, in the specific case of trust, there is not plausibly any such causal dependency between the relevant states of the world (whether, holding fixed that one has trusted, the trustee proves trustworthy) and whether one hopes that she has. And this is the case even if sometimes or even often, in cases of hoping, there is such a causal dependency. (Compare: your hoping might straightforwardly influence how you perform; but it won't influence how your *trustee* performs). Granted, your hoping the trustee proves trustworthy might lead *you* to direct attention to your trustee's performance *qua* trustee. But – and this is a point we've explored in Section 2 – to the extent you're monitoring their performance, you're not trusting them. In sum, then, it looks like the dominance argument remains a problem for the proponent of 'good trusting is a kind of good hoping'.

Yet, here is a further card for such a proponent to play. "Granted, once you've trusted someone, then hoping they'll prove trustworthy will never be a better strategy than refraining from hoping. *However*, hoping can itself *lead* to good trusting in the first place. Consider that without some hope, one might never trust at all, insofar as trusting involves incurring risk. In this respect, hoping isn't 'causally idle' *vis-à-vis* the trustee's proving trustworthy; it is an enabling condition[38] for this result's coming about." The continuation of this line maintains that good trusting is good hoping when (and only when) one's hoping enables successful trust

[37] Bovens anticipates this line of reply in his (1999, sec. III.a).

[38] For a notable articulation of enabling conditions, see Dancy (2004, 51, 64, 172).

moreso than it enables trust that is then betrayed. This is, at least, a *prima facie* coherent picture for how we might think of good trusting in terms of good hoping.

But this picture quickly breaks down under scrutiny. For one thing, if the trust-relevant value of hoping is to be found *outside* of trusting itself – such that we locate its value *prior* to one's trusting, as an enabling condition for that trust to have initially to come about, then identifying this value in hoping simply fails to qualify as a vindication of what *good trusting* is which adverts to good hoping.

The proponent of the idea that good trusting is a species of good hoping, then faces a dilemma. On the one hand, lies the dominance argument, which seems to imply that the value of mere hoping will trivially swamp the value of good hoping – a result at tension with the *ex ante* theoretical constraint on an account of good trusting which is that it exceeds the value of mere trusting. But, the dominance argument assumed no causal dependency between states of the world and hoping. While there is no such causal dependency in the special case of trusting, as hoping the trustee will prove trustworthy doesn't increase the likelihood the trustee will in fact do so, there *is* a causal dependency in the sense that hoping can function as an enabling condition for incurring risk constitutive of trusting. *But*, and this is the other horn, if one sidesteps the dominance argument in this way, one is then giving a story about how hope has some trust-relevant value, but is no longer giving a vindication of what *good trusting* is which adverts to good hoping. The above dilemma looks difficult to overcome for a proponent of the view that good trusting is a species of good hoping. But – even setting the dilemma aside entirely – there remains a further reason why such a proposal is sure to face an uphill battle.

This final reason has to do with an additional kind of *ex ante* constraint on good trusting – one that involves rationality and risk assessment. As we saw in Section 2, it is platitudinous that trusting essentially involves subjecting oneself to non-negligible risk of betrayal. This is platitudinous about trust in a way that, by parity of reasoning, the claim that knowledge is factive is platitudinous about knowledge. But given that trusting es-

sentially involves incurring some risk, we should thereby *expect* that good trusting, as such, will be incompatible with *poorly navigating these risks* that, by trusting, one thereby incurs.[39]

But here is where, again, the prospects for assimilating good trusting to good hoping look dim, as hoping tends to make us *worse* at the sort of risk assessment good trusting plausibly demands of us. This is for two reasons, which are related. First, and as Bovens notes, the very act of hoping for something inclines us to a predictable error in reasoning, which is to 'overestimate the subjective probability that the [hoped for] state of the world will come about' (1999, 680). A well-studied way in social psychology in which this kind of overestimation occurs is *via* the mechanisms of the availability heuristic.[40] But perhaps even worse, rationally speaking, is that hoping for an outcome has been demonstrated to encourage – as McGeer puts it – 'superstitious ideas of our own agential powers' such that we are led, via hoping, to overestimate the sense in which our hoping *itself* raises the likelihood that the hoped for event will come about – and this is *two* rational mistakes bundled into one. That hoping is, psychologically, a kind of invitation to misperceive the causal efficacy of our own agency (in connection with the hoped for event) is a common view in the psychology of hope (e.g., Snyder et al. 1991) and it reveals an important way in which hoping of any sort stands to throw a spanner in our capacities for risk assessment (*vis-à-vis* the hoped for event) that good trusting can't plausibly afford for us to compromise.[41]

[39] For discussion on this point, see Carter (2020b) and Coleman (1990).

[40] For a seminal discussion, see Tversky and Kahneman (1973). As studies by Vaughn (1999) suggest, the heuristic is strongest in cases where the outcome of an event is uncertain – which will always be the case when one is trusting given that trusting essentially involves the incurring of some risk.

[41] This is not to say, of course, that hoping can't have beneficial practical consequences, apart from anything to do with the relationship between hoping and trusting. For a discussion of some of the benefits of hoping, see McGeer (2004) and, for psychological benefits specifically, Snyder (1995). For the present purposes, the fact that hoping stands to imperil risk assessment in the case of trusting – when the probability of the trustee proving trustworthy is independent of whether we hope they prove trustworthy – does not bode well for any view on which good trusting, which will presumably preclude defective risk assessment, is meant to be a species of hope.

This concludes the case for rejecting the idea that the evaluative normativity of trusting is going to line up with the evaluative normativity of any kind of hoping. Let's now generalise. What goes for hope plausibly goes – *mutatis mutandis* – for affective states in the neighbourhood of hope, including faith and optimism directed at the object of trust. Just consider that an assimilation of good trusting to good faith or optimism will inevitably face both (i) variations on the dominance dilemma; as well as the (ii) norms of rationality of risk assessment objections. After all, regarding (i), it looks like the value of good faith and good optimism will trivially swamp the value of mere faith and mere optimism whenever the relevant states of the world about which one is optimistic or faithful are causally independent (as they necessarily are in the case of trust[42]) on one's having that faith or being optimistic. Although faith and optimistic, like hope, can be efficacious in *prompting* one to then trust in the first place when one might not have, by pointing to this efficacy one is no longer characterising good trusting, as such, with reference to either of these attitudes that might prompt trust. Regarding (ii), it suffices to note – with reference to the availability heuristic – that faith and optimism, no less than hope, will tend to incline a truster to overestimate the likelihood the trustee will prove trustworthy, and in doing so, come into tension with the kind of good risk assessment that good trusting, as such, plausibly demands.

In sum, the evaluative normativity of trusting simply does not line up with the evaluative normativity of any kind of positively valenced affect of the sort that trust has been identified in the literature.

1.3.2 Good trusting as good conation

The standards that regulate what counts as good trusting must be other than those that regulate good belief *or* good affect. What about good *conation* – viz., goodness with respect to some motivational state or states

[42] This is due to the constitutive tension between trusting and monitoring. For discussion, see Section 2.

that one has?[43]

Within the category of non-doxastic accounts of the nature of trust, affect-based theories like Jones' are but one type of proposal. A different albeit prominent non-doxastic account is the 'participant stance' account, due to Richard Holton (1994), and which takes trust not to consist in the manifesting of any affective attitude, *per se*, but rather, in a kind of 'normatively laden stance' that implies a readiness to react in certain fitting ways to the trustee's e.g., betrayal or cooperation[44]; as Holton puts it:

> [...] you have a readiness to feel betrayal should it be disappointed, and gratitude should it be upheld (1994, 67).

Is trusting perhaps a matter of doing *this* in a good way – viz., is it a matter of good readiness to feel (certain kinds of) reactive attitudes? We can envision at least two dimensions of conative quality here. One dimension concerns *how* ready one is to feel betrayal if trust is disappointed, gratitude should it be upheld. Along this dimension, presumably, the readier, the better. A separate dimension of quality concerns not the extent of the readiness, but *what* one is ready to feel. Here, the gold standard would seem to be a matter of *fittingness*: what one is ready to feel is betrayal (rather than something else) if and only if trust is betrayed, gratitude (rather than something else) if and only if trust is upheld.

Bearing in mind these two dimensions of conative quality implied by the participant stance view – in short, 'readiness quality' and 'fittingness quality' – consider now the following case:

[43] For relevant discussion on conative states and (some representative views about) how they are taken to be motivating, see, e.g., Rosati (2016, sec. 3), Björklund et al. (2012), and Mele (2003b).

[44] A recent participant stance view that is difficult to taxonomise straightforwardly is due to Berislav Marušić (2017) and which is strictly a kind of doxastic account, though one that incorporates elements of a participant-stance account, in the sense that the account maintains that trust is a belief held from the participant stance. From the perspective of assessing what it is to trust well, this kind of doxastic participant stance proposal will be committed to the position that norms of good trusting must incorporate norms of good believing. However, as I argued in Section 2, this commitment turns out to be problematic.

> THREE EASY MARKS: *X*, *Y*, and *Z* share a common flaw: deep-seated naivety. Too easily and often, each trusts unreliable websites, used car dealers, and people peddling get-rich-quick schemes. For simplicity, suppose all three are betrayed 90% of the time when they trust, and that they trust to the exact same extent – viz., none distrusts more than any other. But each trusts differently with reference to the two key quality dimensions of trusting that are implied by the participant stance view. *X* is consistently ready, when *X* trusts, to feel disappointment at perceived betrayal and gratitude at perceived trustee cooperation; however, *X*'s perceptions are not well calibrated with reality. *X* too often misjudges when the trustee has in fact betrayed versus cooperated; consequently, *X* too often, though very readily, fits disappointment with cooperation, gratitude with betrayal. Put simply: *X* scores high in 'readiness quality', poor in 'fittingness quality'. *Y* is in the opposite position. *Y*'s perceptions (of betrayal and cooperation) are, unlike *X*'s, well calibrated with reality; but *Y* is *inconsistently ready*, whenever *Y* trusts, to actually feel disappointment at betrayal when *Y* (accurately) perceives it and gratitude at cooperation when *Y* (accurately) perceives it. Put simply: *Y* scores high in fittingness quality, but low in readiness quality. *Z* shares, *ex hypothesi*, the common flaw of naivety with *X* and *Y*, *however*, *Z* scores as high as *X* in readiness quality and as high in *Y* in fittingness quality.

Here are two observations about the THREE EASY MARKS case. First, all three are – in an obvious sense – bad trusters! All three, we are assuming, trust in ways that lead to betrayal more often than not.[45] This, crucially, includes *Z*, whose overall conative score is impeccable – that is, *Z* does great, and clearly better than either *X* or *Y*, by those combined

[45]Note that we are assuming the social-epistemic environment for trust here is a normal one and so is not unusually epistemically hostile. What best explains their poor reliability is thus not going to be any abnormal features of their environment.

metrics that would seem to matter for good trusting on the participant stance view – viz., readiness to feel certain fitting attitudes in response to betrayal and cooperation on the part of the trustee.

The proponent of the participant stance view has a few moves available in reply, but none is promising. One move is to simply bite the bullet and say that Z represents good trusting in virtue of the dimensions of trusting quality that distinguish Z favourably from X and Y. This move, though, looks like a non-starter, given what we've already stipulated about Z's betrayal ratio. A more sophisticated move would be to insist that Z's goodness as a truster as represented by Z's admirable readiness to feel certain appropriate attitudes to the trustee, given different ways the trustee might behave, distinguish some important dimensions of good trusting, even if good conation doesn't capture all good ways in which one might trust.

I think we should regard this more sophisticated reply with some suspicion, however. The reason why can be put in terms of an additional *ex ante* desideratum we should expect any account of good trusting to satisfy: namely, that an explanation of what makes for good trusting can't be orthogonal to the value of a trust's being successful - i.e., the trustee's taking care of things as entrusted.

Here a brief analogy to epistemic norms will be of use. Evaluative norms of belief are obviously *not* orthogonal to successful belief – viz., true belief and knowledge. Consider, for example, the evaluative epistemic norms that aim to capture *justified belief* – e.g., these norms tell us that justified beliefs are 'reliably produced' beliefs; 'beliefs that fit the evidence', etc. Both of these are, as Sanford Goldberg (2015) notes, 'standards of success in connection with our pursuit of truth (and avoidance of error)', or perhaps in connection with our pursuit of knowledge (and avoidance of ignorance).[46] Put differently: it is because in believing we aim at truth and knowledge that the evaluative norms of belief capture (in different ways) standards of success in connection with *these* rather than some other aims. The norms are not orthogonal to, but rather importantly constrained by, what counts as successful attainment of the aim of the kind of attempt

[46]For a recent function-driven defence of this view, see Simion (2019).

one makes by believing.

But, as THREE EASY MARKS illustrates, the norms of good conation – of which the norms of good trusting will be a proper subset on the participant stance view – are *entirely* orthogonal to successful trusting; this is because the satisfaction of conative norms (i.e., readiness to feel certain attitudes in response to trusting outcomes) floats entirely freely of the aim, in trusting, that the trustee take care of things as entrusted. '*Z*' in our example case illustrates this, maximially satisfying conative norms while trusting in ways that rarely ever result in the attainment of the aim *Z* makes an attempt at attaining in trusting.

1.4 Concluding remarks

This chapter took as a starting point a question that an account of trust ought to be able to answer: what is *good* trusting? A the very least, what we say about the nature of trust ought to be compatible with a plausible view of the evaluative normativity of trust.

What we've seen, however, is that getting this right is easier said than done. If the leading contenders on offer are right, then we should expect good trusting to be principally a matter of good believing (e.g., Hieronymi 2008; McMyler 2011), or good affect (e.g., Jones 1996; Baier 1986), or good conation (e.g., Holton 1994). What this chapter has attempted to show is that good trusting doesn't plausibly line up with any of these things.

Rather than to simply try to select and then make do with the lesser of the known evils, I will – in the next chapter – suggest a different way forward, one that involves the identification of trust as a *performative kind*. As we will see, if we think of trust as a performative kind, we avoid the problems that face accounts of the evaluative normativity of trust that are restricted to theorising about good trusting as a species of good believing, good hoping, good emoting, good conation, etc. And this is the case even if trust sometimes or even usually involves combinations of these attitudes (both doxastic as well as non-doxastic) or stances.

Chapter 2

Trust as Performance

2.1 Introduction

Good trusting is not something we can capture in terms of good believing, good affect, or good conation. So where do we go from here?

Attempting to salvage any of these views, with some special pleading, does not look particularly attractive. Neither does opting for some kind of disjunctive proposal. But the good news is that we needn't resort to such strategies. This is because there is a simple view that gets us everything we could want – and more – out of a view of the evaluative normativity of trust, and with none of the baggage that comes with any of the other views.

Here is the key thesis I will defend and further develop in what follows:

> (†) Trust is a performative kind. The evaluative normativity of trust is a special case of the evaluative normativity of performances generally.

Several key ideas here need some unpacking. In this section, I will:

- briefly outline the normative structure of performances, construed as aimed attempts, giving special attention to the three central eval-

uative norms that apply to any performance type: *success*, *competence*, and *aptness*;

- sketch and defend the thesis that trusting is a performance-type, and in doing so, characterise (with reference to (a)) the three central evaluative norms that apply to trusting: *successful trust*, *competent trust*, and *apt trust*;
- show how the key thesis (†) satisfies key desiderata on any account of good trusting which other proposals canvased in Chapter 1 (i.e., good trusting as good believing, good affect, good conation) failed to meet;

Let's take these in turn.

2.2 Telic Normativity

A certain kind of normativity – *telic normativity* – is applicable to all performances that are attempts with aims.[1] A simple example is the archer's performance of shooting an arrow at a target.

There are three central ways we can evaluative this performance, *as* an attempt. First, we can evaluate the attempt against the norm of *success*. An attempt is a 'better' attempt if it succeeds in attaining the aim internal to the kind of attempt it is, than if it fails – viz., a shot that hits the target is better than one that misses.[2] Second, we can evaluate the attempt against the norm of *competence*. Regardless of whether an attempt actually succeeds in attaining its aim – that is, regardless of how the attempt stacks up against the success norm – the attempt is better if competent

[1] The key ideas of telic normativity originated from Ernest Sosa's 2005 John Locke lectures, which were later published as *A Virtue Epistemology* (2007, 2009) and refined in Sosa (2015) and, in more recent work, redescribed as *telic normativity* (previously: *performance normativity*) in Sosa (2021). For overviews of recent work on performance normativity, see Kelp (2020a) and Vargas (2016). For critiques and developments of performance normativity, see, e.g., Chrisman (2012), Kelp et al. (2017), and Carter (2020a).

[2] See Sosa (2020) for the most recent presentation of the evaluative normativity of attempts as attempts.

than if incompetent. A competent attempt – to a first approximation[3] – will issue from a disposition to succeed (at attaining the aim internal to the performance-type) reliably enough when one tries in normal conditions. Third, an attempt is a better attempt if is not just successful and competent, but *apt*, viz., successful *through* competence rather than luck.

These three evaluative norms – *success*, *competence*, and *aptness* – point to *three distinct ways that any performance might be good*. The 'goodness' here is attributive goodness; it applies to performances *qua* the kind of aimed attempt they are. The executioner's skilled movements might be successful, competent as well as apt, while at the same time reprehensible.

Finally: it is important to note that the formula 'success + competence = aptness' is incorrect. Suppose the archer's shot is fired competently but – due to a freak gust of wind – is blown off target. But *then*, due to a second freak gust of wind, is blown *back* on course, so that the arrow lands in the bullseye. Here the shot is competent and successful, but not apt, because the success is not through competence but through luck.[4]

2.3 Trust as performance

Consider a simple case of three-place interpersonal trust which is betrayed. Suppose you trust your friend with a secret, and you find out later that your friend spilled the beans. There is a clear since in which your trusting your friend with that secret *did not succeed* in attaining *what it was* at which, by trusting them with that secret, you thereby aimed. They did *not* – put generally – *take care of things as entrusted*. Their having done so would have involved, in this case, their *not* repeating what you had told them.

[3] This initial characterisation, which will be superseded by a more developed view in Chapter 5.

[4] Performances that are successful and competent but inapt have a 'Gettier' structure, where the success is disconnected from the good method used. For discussion, see Sosa (2007, Ch. 2, 2010a, 467, 474–5) and Greco (2009, 19–21, 2010, 73–76, 94–99). Cf., Pritchard (2012, 251, 264–8).

Recast in the language of performance normativity: your trusting here didn't do very well by the lights of the evaluative norm of *success*, and in a way that is broadly analogous to how an archer's shot would be better if it hit the target than not, a belief better if it is accurate (true) rather than inaccurate (false). Accordingly:

> **The Evaluative Success Norm of Trust (ESNT)**: *S*'s trusting X with ϕ is better if X takes care of ϕ as entrusted than if X does not.

But just as missed shots and false beliefs can be *competent* despite failing to secure the relevant aim, likewise, trust can be competent even when it is not successful. When it is competent, it will derive from exercise of trusting *skill*[5], which one has only if is disposed to trust successfully reliably enough when one trusts in proper shape and properly situated.

Why is trusting skill indexed to 'proper shape and situation'? Compare: it does not count against your having the skill to drive a car if you would fail to perform reliably behind the wheel when attempting to drive *if* drugged and placed on slick roads. Likewise, it doesn't count against your skill to trust well if normal bounds of – to a first approximation – risk, effort and skill (required of the trustee) are not present. In a bit more detail: it doesn't count against someone's having a skill to trust well if the truster would not trust successfully reliably enough in conditions where the (a) risk to the truster is excessively high and gains of betrayal are enormous; or where the level of (b) effort or (c) skill that would be required by the

[5]Note that Sosa's own terminology has shifted over the years when it comes to skill, in connection with shape and situation. For one thing, earlier discussions (Sosa 2010a, 465, 470) use the term 'seat' rather than 'skill'. For another, Sosa sometimes describes what one has, when one has the skill (i.e., what one retains even when in improper shape and while improperly situated) as an *innermost competence* (2015, 83) distinct from the complete competence one has when one's skill is conjoined with proper shape and situation. By contrast, the term *inner competence* (2017, 191) is meant to pick out what one possesses when they possess both skill and shape, but not the situational element of the complete competence. For general discussion on these points, see Sosa (2015, Ch. 3, 2017, Ch. 12).

trustee to take care of things as entrusted is abnormally high.[6]

When a skilled truster is in proper shape and properly situated – we will look at these condition in more detail in Chapter Five – the truster then has the (complete) *competence* to trust well, and not *merely* the skill to do so. Trusting that issues (non-deviantly) from such a competence is good, *qua* trust, in an important respect – viz., the very same respect in which other kinds of competent performances are good (*qua* their performance type) *even if* it is one of those times where the performance does not succeed. This is implied by the more general performance-theoretic idea that a given attempt is better if it issues from a disposition to reliably attain its aim in normal conditions than otherwise.[7] Thus:

> **The Evaluative Competence Norm of Trust (ECNT)**: *S*'s trusting *X* with ϕ is better if *S* trusts *X* with ϕ competently than if *S* does not.

Because (as noted in §2.2) *any* performance could be both successful and competent without being *apt* – which is of a higher quality *qua* performance than either successful or competent performance, or a conjunction of them – the same goes for trust. For example, suppose you competently trust a reliable colleague to pay back a loan on a particular date (say, 1 January). On 31 December, your colleague is in an accident which causes total amnesia. Struggling to regain memory, your colleague begins to remember who you are and then simply fabricates a specific memory (which luckily happens to be veridical) that they owe you money which must be repaid by 1 January. Because the friend is of a good and trustworthy character, they are on the basis of this fabricated but veridical memory moved to repay the loan, which they do. Your trusting them is thus successful, they have taken care of things as entrusted; the trust is also competent; but *qua* performance it nonetheless falls short in that it is not successful through competence, but successful just by dumb luck (i.e., that the

[6] For related discussion on normal boundaries within which good trusting is valued, see Carter (2020b, sec. 6).

[7] This is a first approximation, to be superseded with more detailed discussion in Chapter 4, §4.3.1.

trustee fabricated a veridical memory rather than a fictitious one).[8] Thus, in addition to ESNT and ECNT, trust is also evaluable with respect to the following norm of *aptness*:

> **The Evaluative Aptness Norm of Trust (EANT)**: *S*'s trusting *X* with ϕ is better if *S* trusts *X* with ϕ aptly than if *S* does not.

Apt trust is a kind of *achievement*, a success through competence. A common view in the axiology of achievements is that the value of an achievement does not reduce to the value of attaining the relevant success any old way (including through luck *even when* the attempt was a competent attempt – as in the 'lucky success' case above).[9] And this idea is captured nicely by EANT, according to which the attributive goodness of apt trust asymmetrically entails the attributive goodness of successful, competent trust – just as we should expect it would.[10]

In sum, then, the key claims advanced thus far are that (i) trust is a performative kind; (ii) the evaluative normativity of trust is a special case of the evaluative normativity of performances generally; and (iii) ESNT, ECNT, and EANT capture three distinct evaluative norms against which any instance of (three-place interpersonal) trust can be evaluated as better or worse *as an instance of trusting* – with EANT representing a higher standard of good trusting than ESNT and ECNT.

---

[8] Another variation on this kind of case, with the same results, will appeal not to amnesia but to what Sven Bernecker (2010, 137–38) calls 'trace creation' and 'trace implantation' – where memory traces are created in vitro and implanted. While this is perhaps less plausible than amnesia – it makes for a cleaner case given that there is no worry that the amnesia would undermine one's trustworthy character.

[9] For discussion, see, e.g., Greco (2010, Ch. 6), Sosa (2010a), Pritchard (2009a, 2009b), and Bradford (2013, 2015b, 2015a).

[10] Put another way: the idea that apt trust is better than inapt trust implies that apt trust is better (*qua* trust) then either mere successful trust or mere competent trust, or (as the veridical memory case above suggests) mere successful *and* competent but inapt trust.

2.4 Taking stock

Before adding anything further to the picture just developed, let's see how it fares against the problems (from §§1.2-3) facing the competing views of good trusting surveyed.

2.4.1 vs. the doxastic account

There were problems with the idea – implied by doxastic accounts of trust – that the evaluative norms of trust are (a subset of) evaluative norms of belief. In short, we saw that there is a constitutive tension between trusting and complying with evidence and knowledge norms of belief. These problems are are not applicable to the performance-theoretic account, which does not assimilate good trusting to good believing in the first place. Moreover, the proposal is not committed – problematically, as the doxastic account is – to predicting that good trusting will be a function of doing something well *involuntarily* to the same extent that belief is involuntary. Finally, the proposal is not challenged by cases – distinctively problematic for doxastic accounts – where good trust is achieved by suspending belief rather than by believing anything about the object of trust well.

2.4.2 vs. the affective account

One worry for the assimilation of good trusting to good affect was that such a proposal is in tension with a plausible *ex ante* constraint on an account of good trusting, which is that the value of good trusting should exceed, on any plausible account of what good trusting involves, the value of mere trusting. However, the dominance argument threatened to show – in tension with this constraint – that the value of mere hoping will trivially swamp the value of good hoping, and *mutatis mutandis* for other positive affect.[11] The 'trust as performance' view does not succumb to the dominance argument, given that each of EANT, ESNT, and ECNT is better

[11] This argument assumed no causal dependency between states of the world and hoping. See Section 3 for further discussion.

than *mere* trust that is neither apt, successful or competent – and the value of neither successful nor competent trust, nor the conjunction of the two, swamps the value of apt trust.

Further, a problem for the would-be assimilation of good trusting to good affect is that we should expect that good trusting, as such, will be incompatible with poorly navigating risks that, by trusting, one thereby incurs. However, as we saw in Section 3.1, the prospects here aren't promising, given what social psychology tells us about the kinds of rational mistakes that the very act of hoping or being optimistic vis-a-vis an outcome inclines us towards. On the performance-theoretic view, *successful* trust is of course compatible with poor risk assessment[^57] (in the sense that, analogously, e.g., hitting the bullseye and guessing correctly are compatible with using poor form and getting lucky), but – crucially – competent trust and apt trust are *not*.

2.4.3 vs. the conative account

As the THREE EASY MARKS case illustrated, the norms of good conation – of which the norms of good trusting will be a proper subset on the participant stance view – are *entirely* orthogonal to successful trusting, given that the satisfaction of conative norms (i.e., readiness to feel certain attitudes in response to trusting outcomes) floats entirely freely of the aim, in trusting, that the trustee actually take care of things as entrusted. The performance-theoretic account, by contrast, takes that aim as the normative starting point – in that it is with reference to this aim that we understand not only the evaluative norm of successful trust, but also by extension competent and apt trust.

2.4.4 An ecumenical advantage

Question: "Surely apt trust, as well as very often successful and competent trust, will require good believing – including accurate and reasonable beliefs about the trustee's good will and competence – as well as (often) combinations of affective and conative attitudes. How does the performance view countenance this observation?"

Reply: Trusting will often, like other performances, require the good execution of other subsidiary activities. But the performance *itself* is evaluated with reference to the standards of success, competence and aptness, *qua* the performance type it is, and *not* with reference to the standards that regulate what make the subsidiary activities that are often necessary for good trust good as the kind of things they are.[^58]

While the performance-theoretic account doesn't make the mistake of assimilating good trusting to good ϕ-ing for some ϕ (or any set of ϕ-ings) the execution of which is among the subsidiary activities that good trusting usually involves, it *does* accommodate the data point that good trusting will plausibly often, as Bernd Lahno (2004) puts it, 'include behavioral, cognitive and affective dimensions or aspects' (2004, 30). Other proposals are comparatively more restricted in how this data point could be accommodated – as each predicts the goodness of good trusting will primarily be a matter of the goodness of one of these things but to the exclusion of others. In this respect, the performance-theoretical account can claim a kind of ecumenical edge.

2.5 Concluding remarks

This chapter has laid out the key contours of a core idea that will now be developed in more detail, extended, and applied in various ways throughout the book. The idea can be summed up, in slogan form, as 'trust *as* performance'; the norms that govern what count as good and bad trusting are performance-theoretic 'telic' norms of success, competence, and aptness. And these are norms that apply to trust construed as as a performance constitutively aimed at the trustee taking care of things as entrusted; when the trustee does take care of things as entrusted, that's what it is for trust to succeed. But just as there's more to believing than getting true beliefs, there's more to trusting than trusting (merely) successfully. Thus, we can evaluate trusting for competence (did the trust issue from a disposition on the part of the trustor to reliably enough trust successfully?) and also for *aptness* (was trust not only successful and competent, but successful *because* competent?).

As we've seen (in §2.4) the above simple picture of trust and its constituent normativity easily sidesteps the problems that faced (respectively) account of trust that would have us assimilate good and bad trusting to good and bad believing, affect, and conation. This is an initially promising start, though plenty of questions remain.

For one thing, it seems that sometimes, the trust we place in others is merely implicit, below the surface of conscious reflection. Consider, the trust you might place in a family member without giving a moment's consideration as to whether to trust them. Other times, trust seems much more deliberative and calculating.

Can such a distinction in the *way* we place our trust be reconciled with our core evaluative norms ESNT, ECNT, and EANT? Relatedly, just as a skilled inquirer will in some circumstnaces skilfully forbear from belief (e.g., by withholding judgment), presumably something closely analogous will go for trusting. But how might the view presented so far account for this – viz., for cases where it seems as though the good truster skilfully *refrains* from trusting, in a way that is to her credit as a trsutor? Is this even something we can make sense of when we take, as we have, as a starting point the guiding idea that trusting is aimed performance?

With these questions in mind, the next chapter takes up the task of developing the further the account sketched thus far, with a particular focus on distrust and forbearance.

Chapter 3

Forbearance and Distrust

3.1 Introduction

This chapter adds to the framework developed so far in two ways. We begin by distinguishing two core species of trust – *implicit* and *deliberative*, which differ in their constitutive aims. The implicit/deliberative distinction (which we return to in later chapters) offers us a useful vantage point from which we may extend our framework from trust to *distrust*, and in doing so, to recognise both (i) how distrust, like trust, may be implicit or deliberative; and (ii) how the distinction between wide-scoped distrust – what I call Pyrrhonian *mistrust* – and narrow-scoped distrust allows us to appreciate how the latter (in both its implicit and deliberative varieties) though *not* the former kind of distrust is answerable to telic norms of success, competence, and aptness.

3.2 Varieties of trust *qua* performance: some distinctions

According to (EANT), S's trusting X with ϕ is better if S trusts X with ϕ aptly than if S does not, and S's trust is apt just in case it is successful

because competent.

Let's now take things a bit further. Consider that trust can be apt even when trust is *implicit*. We very often trust others with small things without ever consciously deliberating about *whether* to have trusted them in the first place; and we can do *this* better or worse – in ways that are successful, competent and apt. Compare: we believe many things *are particular ways* around us, implicitly, which guide action despite our having never attempted, through any conscious deliberation, to 'settle the question' for which a stance (i.e., on whether p – e.g., that the table is an arm's length away) would constitute an answer.[1] Some of these beliefs are apt, as they are when their correctness manifests competence, others aren't.

Within a performance-theoretic framework, the aim at which implicit trust constitutively attempts to secure is best understood as *teleological*, not intentional.[2] And it is the teleological aim of implicit trust with reference to which we assess implicit trust for success, competence and aptness. No intentional aim is needed for such performance-theoretic assessment. By way of comparison, and as Sosa (2021, 25, fn. 12) notes, we can assess our implicit or 'functional' beliefs for success, competence and aptness – those that guide behaviour below the surface of conscious reflection – not because a thinker intentionally aims at anything, but just because teleologically our perceptual systems aim at correctly representing our surroundings.

On many more substantial matters, however, our trust is not merely

[1] For an extended discussion on this point in connection with performance epistemology, see Sosa (2015, Ch. 3, n. 5).

[2] On this point, it is useful to consider Sosa's (2015) remarks on what is needed for performative assessment as follows: 'functional states can have teleological aims. Thus a state of alertness in a crouching cat may be aimed at detecting vulnerable prey. Whether as a state it can count as a "performance" in any ordinary sense is hence irrelevant to our focus on "performances" that have an aim and to which we may then apply our AAA aim-involving normative account. All that really matters for this latter is that the entity have a constitutive aim, whatever may be its ontological status or the label appropriately applicable to it in ordinary parlance' (2015, Ch. 5, n. 5). For other discussions of functional and teleological assessment, see Sosa (2017, 71–72, 129–30, 152, 2021, 24–31, 52–58, 64, 110, 118).

implicit but *deliberative*, and it is deliberative in a way that is broadly analogous to how some of our considered inquiries (i.e., inquiries in to 'whether *p*' questions) are deliberative. In such cases, we consciously consider whether to judge or suspend (intentionally omitting judgement). Likewise, when one faces a salient choice whether or not to trust someone X with something ϕ one deliberates on *whether* to trust X with ϕ or whether instead to forbear (intentionally omit trusting.)

Let us continue this analogy further. Following Sosa (2015, Ch. 3, 2020, 2021), it is plausible that when we deliberately judge whether *p*, we intentionally aim not just at getting it right any way, but at getting it right *aptly* (for performance-theoretic virtue epistemologists: aiming to get it right aptly is tantamount to aiming at apt belief – *knowledge*[3] –, rather than at truth any old way).[4] This is, by way of an athletic comparison, just as a basketball player[5] aims by shooting not *merely* to make the shot any old way, but to make it competently, to make a well-selected shot.[6] And, plausibly, *mutatis mutandis*, for deliberative – rather than mere implicit – trust: in deliberately trusting, we intentionally aim, in trusting, not *merely* at successful trust (like a basketball player chucking from half court, or an inquirer who aims at truth through a guess), but at *aptness* – viz., at *apt trust*.

The relevant performance-theoretic analogies are thus: implicit belief aims (teleologically) at truth, deliberative belief (judgement) aims (intentionally) at apt belief (knowledge). Implicit trust aims (teleologically) at successful trust, deliberative trust aims (intentionally) at apt trust. Implicit belief, when apt, is knowledge; judgemental belief, when apt, is apt belief (knowledge) of a *higher quality* (i.e., *fully apt* belief, or *knowing full well*[7]) – what results when one aptly attains the aim (aptness) of

[3] For a recent defence of the idea that inquiry is knowledge-aimed, within a wider knowledge-first virtue epistemology, see Kelp (Forthcoming).

[4] Cf., Schechter (2019) for criticism.

[5] Unless, of course, time is expiring.

[6] The coach will berate a player who chucks it from half court, even if it goes in. For discussion, see Sosa (2015, Ch. 3) and Carter (2016b).

[7] This term is coined in Sosa (2021).

judgemental belief. Implicit trust, when apt, is apt trust; deliberative trust, when apt, is apt trust of a higher quality (i.e., fully apt trust) – when one aptly attains the aim (apt trust) of deliberative trust (we will explore deliberative trust, and apt deliberative trust (i.e., *convictively apt trust*), in detail in the next chapter).

3.3 From trusting to distrusting

Here is how all of the above connects with the question of *good distrust*. As a first point of note: the right way to characterise the *way* we aim at aptness, when we deliberate about *whether* to make the relevant attempt (or *not*) in the first place is in terms of a *biconditional aim*: to ϕ iff one's ϕ-ing would be apt.

Just as we can actually *make some attempt*, X, in the endeavour to attain that biconditional aim, we can also *forbear* from X'ing in the endeavour to attain that *very same* biconditional aim – viz., to ϕ iff one's ϕ-ing would be apt. For example, the inquirer pursues the 'positive' Jamesian aim (attaining aptness) by (positively) affirming whether p in the endeavour to affirm if and only if doing so would be apt; but the inquirer can also contribute to the biconditional aim by contributing to its subsidiary (negative Jamesian) aim, *avoiding inaptness*, by *forbearing* on whether p, and doing so *also* in the endeavour to affirm if and only if doing so would be apt.[8]

We are getting close now to seeing how *distrust* is itself is something we can evaluate for success, competence, and aptness – given that (put generally) forbearances, like the performances of which they are omissions, can be aimed. But first, there an ambiguity that needs addressed.[9] The following locution <*forbear from X'ing in the endeavour to attain aim A*> is crucially ambiguous between a narrow-scope and a wide-scope reading. *Narrow-scope forbearance* should be read as: (Forbearing from *X*'ing) in the endeavour to attain a given aim *A*. By contrast, *wide-scope forbearance*

[8] See Sosa (2021) for a detailed discussion of this idea within telic virtue epistemology.

[9] See Sosa (2021, 49) for discussion.

should be read as: forbearing from (X'ing in the endeavour to attain a given aim A).

Widescope forbearance from *trusting* is something akin to the Pyrhhonian analogue in the case of human cooperation as opposed to inquiry. It is a wide-scope abstaining from trusting simpliciter, thus, including from trusting in the endeavour to attain any aim. Call this *Pyrrhonian mistrust*, a omission from trust – though not an omission, in the endeavour to *trust* if and only if that trust would be apt, of trusting.

By contrast, *narrow-scope forbearance* from trusting, but not wide-scope forbearance, is constitutively *aimed* forbearance from trusting.[10] Take a simple case of three-place interpersonal narrow-scope forbearance from trusting: when I consciously deliberate whether to trust the stranger with my keys, and intentionally forbear, my forbearance is aimed; I forbear in the endeavour to, *by forbearing*, avoid trusting inaptly. Call this narrow-scope kind of aimed forbearance from trust *deliberative distrust*; it is (forbearing from trusting) in the endeavour to avoid inapt trust, and *not* forbearing from (trusting in the endeavour to avoid inapt trust).

Deliberative distrust is subject to the evaluative norms of *success*, *competence*, and *aptness*. The success norm on deliberative distrust says: S's (deliberative) distrusting X with ϕ is better if S's forbearing from trusting X with ϕ avoids inaptness than if S doesn't. (Deliberative distrust *fails* if, (i)

[10] It is worth noting a connection between, on the one hand, the distinction between wide-scope/narrow-scope forbearance from trusting, and, on the other, Jane Friedman's (2013) remarks on how to characterise a certain kind of agnosticism of interest in epistemology. As Friedman puts it: '[...] the sort of neutrality or indecision that is at the heart of agnosticism is not mere non-belief and can only be captured with an attitude. This means that the attitude will have to be one that represents (or expresses or just is) a subject's neutrality or indecision with respect to the truth of some proposition. This will have to be either a sui generis attitude of indecision, or some other more familiar attitude' (2013, 167). On the present proposal it is worth noting that whilst wide-scope forbearance from trusting, mere omission, needn't involve any positive characteristics, narrow-scope forbearance – much like the kind of positive attitude of indecision that Friedman associates with agnosticism – will often have (some combinations of) attitudes, affect, and/or conation. In this respect, narrow-scope distrust is like both the agnosticism that Friedman describes, as well as trust itself (see Section 4.3 for discussion).

one deliberatively distrusts X with ϕ; and (ii) *were* one to have trusted (X with ϕ), one's trusting (X with ϕ) would have been apt.) That would have been bad distrust along the success dimension. This is so *even if* that distrust was *competent*.

The competence norm on deliberative distrust says: S's (deliberative) distrusting X with ϕ is better if S forbears from trusting X with ϕ competently than if S doesn't; this will require that S's deliberative distrusting X with ϕ manifest S's competence to (narrow-scope) forbear in ways that reliably lead to avoiding inaptness, when the truster is in proper shape and properly situated. The *overly cynical* person might forbear from trusting a reliable trustee successfully (and avoid inaptness given that the reliable would-be trustee would have failed to prove trustworthy on this occasion) but would still fail to do so competently. Finally, the aptness norm on deliberative distrust says: S's deliberative distrusting X with ϕ is better if S deliberatively distrusts X with ϕ aptly than if S does not.

Question: judgemental belief is to deliberative trust as implicit belief is to implicit trust. But what about *this* analogy: Deliberative trust is to implicit trust as deliberative forbearance is to ?

The answer, of course, is *implicit distrust*, which rounds out our picture. Implicit distrust, like implicit trust, is teleological rather than intentionally aimed forbearance from trusting. *Widescope* forbearance from implicit trusting differs from Pyrrhonian mistrust (widescope forbearance from deliberative trust) not necessarily in respect of involving some different policy or stance, *per se*, but simply with respect to what kind of trust is, in fact, omitted.[11] *Narrow-scope* implicit distrust is constitutively aimed, teleologically, not (as is narrow-scope deliberative distrust) at *avoiding inaptness*, but at avoiding *unsuccessful trust* – viz., trust where the trustee fails to take care of things as entrusted. As such, implicit distrust is evaluable as successful, competent and apt in connection with the aim of avoiding unsuccessful trust: implicit distrust is better if successful (i.e., if *not* forbearing would have resulted in unsuccessful trust) than if not, competent

[11] Possibly, an individual could wide-scope forbear from one kind of trusting without wide-scope forbearing from the other.

than if not (i.e., if one's forbearance manifested a competence to not too easily forbear when, had one refrained from forbearing, one would have trusted successfully) than not, apt than if not.

In sum, by modelling the evaluative normativity of distrust performance-theoretically as norms of forbearance, we get the following picture: (i) first, widescope forbearance from trusting (Pyrrhonian mistrust and non-Pyrrhonian mistrust) is not performance-theoretically evaluable; (ii) narrow-scope forbearance from trusting (just like trusting itself) comes in two varieties, deliberative distrust and implicit distrust, each of which is performance-theoretically evaluable, for success, competence and aptness in connection with (a) the intentional aim of avoiding inaptness (in the case of deliberative mistrust); and (b) the functional/teleological aim of avoiding unsuccessful trust (in the case of implicit distrust).

In order to make these distinctions in distrust quality more concrete, I'll conclude with an example case illustrating each – beginning with the widescoped varieties of forbearance (i.e., from deliberative and implicit trust) that lie *outside* performance-theoretic evaluation.

3.3.1 Widescope forbearance from trust (Pyrrhonian mistrust and Non-Pyrrhonian mistrust)

Pyrrhonian mistrust, i.e., forbearance from (deliberative trusting in the endeavour to trust iff one would do so aptly), is simply an *omission from deliberative trusting*. In the three-place case of interest, it is an omission from deliberatively trusting X with ϕ. I'm using the term 'Pyrrhonian' only because *what it omits* is the taking of an intentional stance – (broadly) analogous to the kind of omission from intentional judgement that characterises a Pyrrhonian withdrawal from intentional judgement (viz., of whether something is so). What one omits is *not* necessarily implicit trust (though omissions from deliberative trust are compatible with omissions from implicit trust). For example featuring an omission from mere deliberative trust but not implicit trust, suppose a misanthrope's plan is to withdraw from all cooperation; the plan is, however, flawed, as the misanthrope implicitly trusts a neighbour N with a small task, T, though

she would not have done so had she deliberated. Even though the misanthrope omits deliberative trust in the case of N with T, she (despite herself) fails to omit implicit trust of N with T. While her *implicit* in this case is performance-theoretically assessable for success, competence and aptness, her non-Pyrrhonian mistrust, viz., her forbearing from (deliberative trust in the endeavour to attain apt trust) – is not.

Contrary to the situation of the misanthrope, who exhibits Pyrrhonian mistrust[12], we can imagine a *recovering misanthrope* in the inverse position: she wide-scope forbears from implicit trust *but not* from deliberative trust. Her misanthropic tendencies are so deeply ingrained, that she – assume *ex hypothesi* – is simply *incapable* of implicit trust. She trusts, if ever, only with a strong will to do so, to overcome these tendencies. Suppose, then that with such will overcoming her tendencies, she intentionally, deliberately, trusts her neighbour N with T. Omitted here is implicit trust, not deliberative trust. While her *deliberative trust* of N with T in this case is performance-theoretically assessable for success, competence and aptness, her non-Pyrrhonian mistrust is not.

3.3.2 Narrow-scope intentionally aimed forbearance from trusting: Deliberative distrust (successful, competent and apt)

Suppose A is deciding whether to trust to trust a friend, B, to watch A's very young children (W) for the weekend. The choice is complicated, but after deliberation on such things as risk of betrayed trust, difficulty of task, trustee goodwill and competence, etc., A decides *not* to trust B with W, and does so in the endeavour to trust if and only if that trust would be (not merely successful) but apt.

This intentional forbearance from trusting *succeeds* even if B *would* have

[12] I am using 'mistrust' in the case of widescope forbearance (compared to mere distrust) to indicate the wholesale character of the relevant omission from trusting – in that widescope cases omit not only trusting but also omit (omission in the endeavour to trust iff doing so would be apt).

succeeded in taking care of the children for the weekend, and so long as *B* would very easily have failed. This is because, if *B* very easily would have failed, then had *A* trusted, she would not have *avoided inapt trust*, even though she would have attained successful trust. However, *A*'s deliberative distrust of *B* with *W fails* if *were A* to have trusted *B* with *W*, her trusting would have been *apt*. Likewise, *A*'s deliberative distrust of *B* with *W* is competent if that distrust reliably enough would have avoided inaptness, and apt iff *A*'s successful deliberative distrust of *B* with *W* manifests this competence.

3.3.3 Narrow-scope functionally aimed forbearance from trusting: Implicit distrust (successful, competent and apt)

One simple example case that can be used to illustrate narrow-scope functionally aimed forbearance from trusting involves bias-driven *testimonial injustice*.[13] Suppose a member of a marginalised group, *M*, testifies that *p* in an open trial, in which *A* is a racist juror.[14] Due to racism, *A* implicitly distrusts *M*'s testimony, including *M*'s assertion that *p*, despite never intentionally determining whether to trust *p* (any more than, say, *A* is reflecting on whether to believe *that* the testifier is speaking, something *A* believes implicitly). In addition, were *A* to deliberate about whether to trust *p* specifically, *A* would have been triggered to reflect on racism in a way that would have offset the racist and implicit tendency to forbear.

In such a case, *A*'s implicit distrust is successful if, had *A* trusted *M*'s testimony' that *p*, *A would* have ended up believing falsely whether *p*. Likewise, the implicit distrust is *not* successful if, had *A* trusted *M*'s testimony that

[13] See Fricker (2007) for the canonical presentation. It is worth noting that an alternative kind of example that might be used to illustrate this species of implicit distrust involves implicit distrust of *experts*. For discussion of this phenomenon, and some of the biases that lead to it, see Baghramian and Croce (forthcoming, sec. 4).

[14] Note that the details of this case are importantly different from a more familiar case of a racist juror in social epistemology, due to Lackey (2007a, 598), the latter of which is meant to illustrate the phenomenon of 'selfless assertion'.

p, A would have ended up believing truly whether *p*. Likewise, *A*'s implicit distrust of *B* with *W* is competent if that distrust reliably enough would have avoided unsuccessful trust, and apt iff *A*'s implicit distrust of *M* with *p* manifests this competence. If we assume, *ex hypothesi*, that *M did testify truthfully that p* in trial, *A*'s implicit distrust will be inapt because unsuccessful. Because racially driven implicit bias is unreliable[15], *A*'s implicit distrust of *M* whether *p*, will, in addition, fail to be competent.

3.4 Concluding remarks

This chapter has taken some initial steps to expand the telic theory of trust, the core ideas of which were outlined in Chapter 2. We began by registering a distinction that will be important in much of the rest of the book – between implicit and deliberative trust. This distinction was then used to illuminate the wider point that just as trusting can be implicit or deliberative, so can *distrusting*. The focus of the chapter then lay squarely on distrust itself, its nature and normativity. It was shown that – and here was the second key expansion of the view – at least one kind of distrust, what we called *narrow-scope distrust* – is plausibly an aimed performance in its own right no less than trusting is, and that it is thus answerable (in both its implicit and deliberative modes) to norms of success, competence, and aptness of *forbearance* from trusting.

Stepping back from a moment, our framework is now much richer than we began with. But it also invites new questions, particularly in light of the implicit/deliberative distinction. Given that implicit and deliberative trusting are distinct *performances*, distinguished by their constiutive aims, one might wonder what kinds of considerations would ever make one variety more appropriate than the other? Answering *this* question requires that we look beyond trust itself as an aimed performance, but also at the wider kind of cooperative practice within which trusting (and distrusting) of both varieties constitute performative moves. This is the task that the next chapter will take up.

[15] See, for instance, Saul (2017), Munroe (2016), and Díaz and Almagro (2019).

Chapter 4

Trust, Pistology, and the Ethics of Cooperation

4.1 Introduction

On the picture developed so far, *implicit trust* is to be distinguished from *deliberative trust*.

Implicit trust is not the result of conscious deliberation any more than are (for example) the beliefs you have that guide your behaviour but which aren't a response to any 'whether' question. For example: "My phone is ringing," "A dog is barking," "My leg hurts." Nonetheless, such beliefs can be successful (if true), competent (if reliably enough true), and apt (if true because competent).[1]

Likewise, on the view advanced so far, we might trust someone with something, ϕ, implicitly, and not as a result of ever having deliberated about whether to trust them with ϕ.[2] But *even so*, this implicit trust – which

[1] See Chapter 3.1. For discussion in the context of performance-theoretic virtue epistemology, see Sosa (2015, Ch. 3, n. 5, 2017, 71–72, 129–30, 152, 2021, 24–31, 52–58, 64, 110, 118).

[2] Granted, we might *also* take implicit attitudes of trust towards individuals them-

aims constitutively just at the trustee's taking care of things *as entrusted* – is subject, like implicit belief is, to evaluative norms of *success* (i.e., *did* the trustee take care of things as entrusted or did they not?), *competence*, and *aptness*. ESNT, ECNT, and EANT detailed in Chapter 2, straightforwardly characterise the evaluative normativity of implicit trust.[3]

But do ESNT, ECNT, and EANT also satisfactorily capture the evaluative normativity of *deliberative trust*? Recall that in deliberately trusting, we intentionally aim, in trusting, not *merely* at successful trust (like a basketball player chucking from half court, or an inquirer who aims at truth through a guess), but at *aptness* – viz., at *apt trust*.[4]

This means that in deliberatively trusting someone to ϕ, one is making a kind of attempt that will be competent just in case the truster manifests *not merely* a disposition to trust in ways that don't too often lead to betrayal (this is what competent *implicit* trust requires), but in ways that *don't too often issue in inapt trust*.[5]

The achievement of *apt deliberative trust* is more impressive, as a trusting performance, than the achievement of mere apt trust. It demands more, namely, the *apt attainment* of apt trust. (One fails to attain *this* aim even if one trusts aptly but might too easily have not done so.)

---

selves; you might implicitly trust your partner, and in such a way that this implicit trust then carries over to particular tasks. As I noted in previous chapters, I am taking three-place-trust as a theoretical focus. This does *not* mean however that what is said about implicit three-place-trust does not have potential ramifications for what we say about implicit two-place-trust. On the contrary, it might be that implicit two-place trust is to be explained in terms of implicit three-place trust. On this point, I'm remaining neutral. See Hardwig (1991) and Domenicucci and Holton (2017) for discussion.

[3] That is: (i) *S*'s trusting *X* with ϕ is better if *X* takes care of ϕ as entrusted than if *X* does not (from ESNT), (ii) *S*'s trusting *X* with ϕ is better if *S* trusts *X* with ϕ competently than if *S* does not (from ECNT); and (iii) *S*'s trusting *X* with ϕ is better if *S* trusts *X* with ϕ aptly than if *S* does not.

[4] Note that it is the constitutive aim of any given performance-type that determines what counts as 'success' for success for that performance-type and, by extension, competence and aptness (see Chapter 2).

[5] For a detailed discussion of the distinction between these competences in the case of belief, see Sosa (2019).

When deliberative trust (which aims at apt trust) is *itself* apt, then trust is *convictively apt* (the choice of terms will be clearer in Chapter Five). Convictively apt trust is, *qua* performance, analogous in important respects with what is called *fully apt judgement* in performance-theoretic virtue epistemology.[6] In particular, whereas a fully apt judgement is the highest grade of achievement-type in the theory of knowledge, likewise, convictively apt trust is the highest grade of achievement-type in the theory of trust.

However, there is quite a bit to be said about convictively apt trust that takes us beyond its structural analogies with fully apt judgement.

This chapter and the next will be organised around two central questions about deliberative (and convictively apt) trust. In the course of answering these questions, the aim will be to understand this kind of trust more deeply, how it relates to implicit trust more broadly in the philosophy of trust, what kind of specific skills it demands of us, and (in Chapter Six) what kinds of things this high-grade trust permits us to non-negligently take for granted.

Our guiding questions in this chapter and the next are as follows:

> **Appropriateness Question:** Implicit and deliberative trust differ, but under what conditions is one kind of trust more appropriate than the other, and what kinds of considerations determine this?
>
> **Substance and Structure Question:** How should we think about what is involved – not just structurally, but also substantively – in the *apt attainment* of apt trust?

The remainder of this chapter will answer the Appropriateness Question, and Chapter 5 will answer the substance and structure question.

[6] See, e.g., Sosa (2015, 2021); cf., Carter (2020a).

4.2 Implicit and deliberative trust

Consider the following two characters, Fidel and Misty, who navigate their respective potential trusting relationships very differently, in one key respect. Fidel almost never trusts deliberatively, almost always only implicitly.[7] Misty almost never trusts implicitly, almost always only deliberatively.

Is either doing better than the other? We of course don't know without more details. But before supplying them – and in order to get a better grip on what details *matter* – it will be instructive to consider by way of analogy their *zetetic* counterparts[8], Fidel-I and Misty-I, who are inquirers who navigate *inquiry* very differently. Fidel-I almost never inquirers into any question of whether *p*; he forms his beliefs unreflectively; they guide his actions despite never constituting the answer to any deliberate inquiry into whether *p*. Misty-I is, by default, unusually sceptical – her inquiries are guided by constant wonder, and she almost always affirms whether something is so only after careful scrutiny.[9]

Let's consider what we should say about Fidel-I and Misty-I, and then how this might organise what we should say about Fidel and Misty. The first thing we might wonder about Fidel-I and Misty-I is what kind of *epistemic environment* each is in, whether it is a cooperative and friendly one (few deceptions and liars about), or whether it is a risky environment to navigate, full of easy error possibilities. If they are both in a friendly environment, then, as one line of thought would go, Fidel-I is – *all else equal* – using a better strategy than Misty-I, one that results in *far more true beliefs and knowledge*. If they are both in an epistemically *unfriendly* environment, however, then – on this same line of thought, Misty-I is – all else

[7] It is worth noting that there is some overlap between the kind of defaul implicit truster imagined here and what Jones (2004) calls a 'basal' truster. For discussion, see Nickel (2015).

[8] I am using 'zetetic' in the sense of Friedman (2020), and as pertaining to *inquiry*.

[9] This is, by stipulation, that she does so. Granted, such a character might be highly unrealistic, given that various assumptions we must make that guide action. For the sake of this thought experiment, let's assume Misty is on the extreme end of what is psychologically possible.

equal – using a better strategy than Fidel-I, one that results in *far less false beliefs and ignorance*. This kind of diagnosis lines up with John Greco's (2013, 2020b) recent work in the epistemology of testimony: according to Greco, the optimal strategy among close-knit communities is *not* to scrutinise testimony at all; whereas, the better strategy when testimony is being introduced into a community for the first time is not to accept it without careful scrutiny. Greco's rationale is that different strategies align better with different epistemic environments.[10]

But if we stipulate that both Fidel-I and Misty-I are in the very *same* epistemic environment – and further, that this epistemic environment is equally friendly and unfriendly – then should we say then that their strategies break even?

When evaluating a thinker's beliefs and judgements as the kinds of things that are truth- and knowledge-directed, the answer is probably that we simply don't have enough information. For example, how *good* is Misty-I at the scrutiny that she subjects her beliefs to prior to affirming when she does? At any rate, when assessing their strategies from *this* point of view, the main information we need to evaluate their strategies is information that helps us work out the extent to which the beliefs they form via these strategies *succeed* in the kind of attempts that they are. Ernest Sosa (2021) calls this kind of telic assessment of beliefs qua attempts *gnoseological* or 'knowledge-related' assessment. This is the *kind of assessment* that tells us that (all things equal) Fidel-I will do better than Misty in an good epis-

[10] In a bit more detail, Greco's tack is to reject the widely held assumption that either all testimonial knowledge requires inductive evidence on the part of the hearer, or none does. His rationale for rejecting this presumption draws from what he takes to be two distinct kinds of activities governed by the concept of knowledge: knowledge *origination* (e.g., uptake) and knowledge *distribution*. As Greco sees it, it's reasonable to suppose that the norms (e.g., quality control) governing originating activities differ from norms (e.g., easy access) governing distributing activities. Greco's key move at this point is to claim that 'testimonial knowledge *itself* comes in two kinds' (2013, 20), insofar as it is sometimes serving the distributing function, sometimes the originating function. Accordingly, some knowledge-generating intellectual virtues will be apposite to one kind of testimonial knowledge, some to the other, given that 'what is required for reliable distribution is different from what is required for reliable origination' (2013, 23).

temic environment, Misty-I better than Fidel-I in a bad epistemic environment, and 'it's unclear' in a middling environment.

However, gnoseological assessment – e.g., where evaluative norms of success, competence, and aptness are central – has its limits. It doesn't tell us other things we might care about when comparing Fidel-I and Misty-I more broadly, as inquirers. Gnoseological assessment is silent about *what questions one should or should not take up and sustain*.[11] These kinds of considerations are proper to the *ethics of inquiry*, viz., to zetetic ethics.[12]

The norms of zetetic ethics forbid (unlike gnoseological norms, e.g., of success, competence and aptness) for example, spending one's time inquiring into trivial or pointless truths (e.g., truths about grains of sand on a beach, blades of grass, etc.[13]), *even when* one inquiries into such truths in ways that that pass scrutiny in terms of success, competence, and aptness. Norms of zetetic ethics (e.g., Alston's 2006, 32, "maximise true beliefs on matters of importance") unlike gnoseological norms – can help us to explain why there is something defective about an inquirer with a wealth of trivial knowledge and no false beliefs. And, in the particular case of Fidel-I and Misty-I, norms of zetetic ethics (e.g., "Pure moral deference is prohibited[14]") can help us answer questions like the following: Is Fidel-I inquiring well if he accepts – even if *knowledgeably so* – all of his philosophical and ethical views second-hand, without seeking first-hand insight into why they are true? Relatedly, is it permissible for Fidel-I to *terminate* his inquiry into whether democracy is better than autocracy by sheer deference to a reliable authority? (Conversely, is Misty-I making some kind of mistake when she fails to simply defer to what the CDC says about the transmissibility of Covid-19, seeking additional first-hand

[11] For a presentation of the difference – in the context of performance-theoretic virtue epistemology – between gnoseology and 'intellectual ethics', see Sosa (2021, Ch. 2).

[12] Though, cf., Kelp (2020b) for resistance to drawing such a clear distinction between these domains.

[13] For discussions of pointless and trivial truths and the problems they pose theories of zetetic ethics (as well as, relatedly, theories of epistemic value), see, Kvanvig (2008), Treanor (2014), and Carter (2011).

[14] For some defences of this idea, see, e.g., McGrath (2011, 2011); Hills (2009), Crisp (2014), cf., Enoch (2014).

confirmation?)

Capturing what are the correct zetetic norms is a big task[15], not one I'm going to take up here. The point of the foregoing discussion of Fidel-I and Misty-I is to highlight that there are two importantly different ways – viz., with respect to telic gnoseological norms and the norms of zetetic ethics – by which we might *answer* the question *"under what conditions is implicit belief more appropriate than judgmental belief?"* Certain kinds of things matter for gnoseological assessment of these strategies, other things matter for zetetic assessment.

4.3 A structural analogy

What goes for Fidel-I and Misty-I, I want to now suggest, goes for what we should say, *mutatis mutandis*, about Fidel and Misty in our original case. The guiding analogy here is that **gnoseology is to zetetic ethics as pistology is to the ethics of cooperation**.

Just as zetetic ethics concerns which inquiries to take up and sustain, and gnoseology concerns good and bad thinking *once an inquiry has been taken up* (and here norms of success, competence and aptness are pertinent), likewise, we can think of the ethics of cooperation as concerning which kinds of people and tasks we should develop cooperative relationships with and place our interest in *initially*, such that holding fixed these relationships and interests, we are then positioned to place our

[15]For a 'knowledge-first' approach to zetetic norms, see Kelp (Forthcoming). However, it is worth clarifying that, on Kelp's view, epistemology itself – all included – is best understood as the theory of inquiry. With this in mind, norms governing inquiry will for Kelp be a larger class of norms than those that I am calling zetetic norms here (as distinct from gnoseological norms). For another more expansive conception of the zetetic, see Friedman (2020).

trust (or not) to further these interests.[16] And pistology[17] (and norms such as EANT, ESCT, and EAPT, as well as performance-theoretic norms governing distrust in Chapter 3) concerns good and bad trusting (and distrusting) *given* the kinds of relationships and interests we are working with. Just as gnoseology and zetetic ethics are two kinds of complementary projects in epistemology (where the norms we are looking to uncover are distinctive to each kind of project), pistology – the philosophy of good trusting as such – and the ethics of cooperation are best understood as two separate and analogously complementary projects in the wider philosophy of trust.

With this broad picture of the philosophy of trust in view, we can now, in a principled way, answer the Appropriateness Question: "Implicit and deliberative trust differ, but under what conditions is one kind of trust more *appropriate* than the other, and what kinds of considerations determine this?"

When we restrict ourselves to purely *pistological* assessment, it looks like the 'conditions most appropriate' for implicit and deliberative trust are just those more likely to facilitate success, competence, and aptness in each

[16] What is the *scope* of the ethics of cooperation? We should think of it as, in principle, very wide. The analogy *gnoseology is to zetetic ethics as pistology is to the ethics of cooperation* invites us to think of the ethics of cooperation as standing to trust as zetetic ethics stands to gnoseology, viz., as encompassing norms on trusting that (like zetetic norms on belief) are not restricted to whether the relevant attempt is successful, competent or apt. For the purposes of the analogy, the ethics of cooperation will at least include norms about how we build and sustain relationships that can serve as preconditions for trusting. However, this is compatible with recognising that the ethics of cooperation has other dimensions to it where the importance of trust lies alongside other ethical values in connection with cooperation, including justice. To make one example, a substantial part of Plato's *Republic* (especially book VIII) can be understood as a project in the ethics of cooperation, in so far as the ideal state is conceived of in terms of whether the state is organised in a way that would facilitate harmonious cooperation between individual members.

[17] The term 'pistology' derives from the Greek 'pistis' (Πίστις), which means trust or faith. Philosophical pistology (the theory of good trusting as such) should not be confused with theological 'pistology', as this term has been used by Niebuhr (1991), which is concerned moreso with faith than trust.

case. Broadly analogous with Greco's 'two pronged' approach in the epistemology of testimony, we might then say that in close-knit environments – where there is a shared history of trust[18] – implicit trust has advantages (all else equal) over deliberative trust, and that the opposite is the case in environments where risks of betrayal are higher.[19]

However, we get a completely different answer when we approach the same question from the 'other side' of the philosophy of trust – viz., from the ethics of cooperation. Here the kind of normativity involved is not restricted to good trusting as such, and includes more broadly how to orient ourselves so as to establish conditions conducive to cooperation.

With this much broader kind of normativity in play, it becomes relevant (for instance) that deliberative as opposed to implicit trust can be positively *counterproductive* in close-knit relationships. For example, as Fay Niker and Laura Sullivan (2018) have recently argued, deviations from implicit trust can lead to ruptures in trust in cases of 'thick' interpersonal relationships, e.g., with there is an shared history of trust.[20] Imagine, for example, the effects on a trust relationship that might arise from a romantic partner *explicitly* and deliberately weighing up whether to trust you on a particular occasion, rather than to simply to so implicitly. Niker and Sullivan take it that suitably close relationships are often predicated upon such deliberation being moot.

Likewise, just as plausible norms of zetetic ethics forbid intellectual apathy as a general tendency, plausible norms governing the ethics of coop-

[18] Within environments where trust has been established, Greco thinks not only that scrutiny is not necessary, but also that we should think of trust (and reciprocity on the part of the trustee) as a product of 'joint agency'. For discussion, see, e.g., Greco (2019, 2020a).

[19] And given that cooperative activities often take places across the spectrum of such environments, from the perspective where we want our trust to be successful, competent, and apt, it is of course going to be best to traffic in both varieties of trust rather than to have any one as a default.

[20] As Niker and Sullivan (2018) note, this category "most often will refer to committed romantic relation- ships, close friendships, familial relationships such as that between a parent and his/her (adult) child, and so on" (2018, 2). For related discussion, see Tsai (2018).

eration will be violated by an individual who never (or too rarely) establishes and sustains the relevant kinds of relationships that are themselves the very *preconditions* for successful, competent and apt trust. Crucially, part of sustaining such relationships will involve being *trustworthy* and cooperative oneself. (Trustworthiness will be the central focus of Chapter 9).

The theory being developed here is in the main a theory of pistology (more akin by way of analogy to the theory of knowledge, in epistemology, than to zetetic ethics), one that is developed further in the remaining chapters.

4.4 Concluding remarks

The key 'takeaway' from this chapter is the contrast between two distinct though complementary domains of normativity of interest in the philosophy of trust. There is the kind of telic normativity that applies to trust *qua* aimed attempt. Telic evaluation of trust (and distrust) for success, competence and aptness is *pistological* assessment, in the very same sense that telic evaluation of beliefs (and disbelief) for success, competence and aptness is – in Ernest Sosa's terminology – gnoseological assessment.

Moreover, just as Sosa distinguishes between gnoseology and intellectual ethics, where the norms of the latter pertain to what kinds of inquiries to take up and sustain, by way of analogy, we should distinguish between *pistology* and the ethics of cooperation, where the norms governing the latter pertain to which cooperative relationships we should pursue and sustain.

Distinguishing between telic pistological assessment and the kind of assessment that is applicable more widely in the ethics of cooperations was needed in order to answer the Appropriateness Question that we began with. As we've seen, the question gets entirely different answers depending on whether the kind of assessment at issue is purely pistological or not.

The telic theory of trust that I'll continue to develop in more detail in fur-

ther chapters is concerned primarily with pistological assessment and the telic norms go hand in hand with such assessments. That said, this is not the last we'll hear of the ethics of cooperation, or of cooperative practices. As it will be later shown, the fact that trusting is a kind of performative move within our wider practices of cooperation turns out to be essential to understanding how trust relates to risk, negligence, vulnerability and monitoring – the topics of Chapters Six and Seven.

Chapter 5

Deliberative Trust and Convictively Apt Trust

5.1 Introduction

On the telic theory of trust developed so far, deliberative trust that is *itself* apt is convictively apt trust. The structure of this kind of trust is more sophisticated than the structure of (mere) apt trust, given that apt trust is *that at which we (intentionally) aim* in deliberatively trusting, and so deliberative trust is itself apt only if *that* aim, the aim of apt trust, is itself aptly attained.

Apt deliberative trust is of special interest because it is what we very often aspire to in trusting[1], much like, in inquiring, we want our deliberative judgements (viz., of whether *p*) to be not just *successful* in attaining their aim (i.e., of the aim apt belief, viz., knowledge) but *apt* in attaining that aim.

In order to bring the achievement of apt deliberative trust – viz., convictively apt trust – into sharper view, let's zero in on the following question:

[1] We are, after all, only rarely in the fortunate position of Fidel from Chapter Four.

> **Substance and Structure Question:** How should we think about what is involved – not just structurally, but also substantively – in the *apt attainment* of apt trust?

The answer given to this (two-component) question – over the rest of this chapter – will be organised in two parts, which focus (i) first (in §5.2) on *substance*, and (ii) second (in §5.3) on *structure*.

Important to understanding the *substance* of apt deliberative trust will be to clearly distinguish between *first-order trusting competence* and *second-order trusting competence* and especially how the latter is paired with a different skill/shape/situation profile than the former; likewise, key to understanding the *structure* of apt deliberative trust – and where *conviction* enters into the story – will be to appreciate, by way of analogy with fully apt judgement, the relationship in cases of apt deliberative trust, between (i) the exercise of second-order trusting competence; and (ii) the attempt one makes in deliberatively trusting.

5.2 The substance of apt deliberative trust

5.2.1 First-order trusting competence

Apt deliberative trust involves two kinds of trusting competence, *first-order trusting competence* and *second-order trusting competence*.

We've already introduced the first variety in Chapter Two. It is (in short, to be developed further) a disposition to trust in ways that don't too often lead to betrayal. As such, one could possess and exercise this disposition without realising they have it and thus without appreciating its limits (and what situations lie beyond those limits).

As was noted in passing in Chapter Two, *all* competences – viz., dispositions of an agent to perform well – are indexed to both *shape* and *situation* conditions appropriate for their exercise. Understanding these aspects of competence is important to appreciating the *substance* of any particular competence – viz., *what it is a competence to do and in what conditions*. Here, it will be helpful to recall Sosa's (e.g., 2017) handy way to flesh

out this point – within the wider theory of performance normativity – in terms of the 'SSS' acronym: *skill, shape, and situation*.[2]

To possess the *skill* to do something, ϕ, is to possess a disposition to succeed reliably enough at ϕ-ing *when you try and are in proper shape and situation for ϕ-ing*. For example, it simply doesn't matter – when it comes to assessing whether you possess the skill to land a plane – whether you can always land a plane safely via emergency water landing, if that's *all* you can do. What we are interested in, implicitly, (when attributing or withholding the skill to land a plane to you) is whether you can land the plane reliably enough in situations that feature normal runways; that is, after all, the kind of situation where reliable performance at landing a plane is principally valued.[3] By the same token, it *doesn't* count against your skill to land a plane if, even when approaching a normal runway, you almost always get in a crash when you attempt to land the plane *after* someone has covertly drugged you with Dimethyltryptamine. The *skill* to land a plane is accordingly best unpacked as the disposition to land a plane reliably enough when sober/awake/alert (i.e., *proper shape*), while behind the cockpit of a working airplane, approaching a suitably flat runway, with plenty of pressurised ambient oxygen (i.e., *properly situated*).

Of course, what counts as 'proper' shape and situation for any kind of performance with an aim varies from domain to domain, as the conditions where good performance is valued differ across domains; the same goes for the threshold for 'reliable enough' performance.[4] On the former point,

[2] See, along with Chapter 2, also Sosa (2015, 2017, 2021, 2020), Kelp (2020a), Vargas (2016), Turri (2011), and Carter (2020a).

[3] Because our competence-discerning judgements need to keep track of who would perform well in situations where (as Sosa puts it) 'human accomplishment is prized (or otherwise of special interest),' it is our *own human interests and needs* that – as we should expect – play a role in fixing the limits of proper shape and proper situation that circumscribe a given competence-type.

[4] Consider, for example, that we don't test for one's driving competence by asking: would the driver perform reliably enough (make it to the destination safely, avoid accidents, etc.) if deprived of oxygen and placed on abnormally slick roads; driving poorly in those conditions doesn't count against one's possessing a competence to drive reliably enough when in proper shape and properly situated—viz., in normal driving conditions.

consider that although a pilot *does* need to be able to land a plane in the dark to count as having the skill to land a plane, it isn't, by contrast, a mark against an *archer's* skill to hit the target if the archer (as opposed to the pilot) would perform unreliably in the dark.[5]

A final point of clarificatory groundwork before we return to the substance of first-order trust: one can have the *skill* to do something without having the *complete competence* to do that thing; possessing the complete competence to ϕ requires not only possessing the skill to ϕ (i.e., which one might retain while drugged or poorly situated) but also that one possess this skill while being proper shape and properly situated.[6] When one exercises one's skill *in those conditions* – viz., when one possess the complete – one's *performance* is competent.

The above offers us a more sophisticated vantage point from which to revisit ECNT (and by extension EANT) from Chapter 2. According to ECNT, *S*'s trusting *X* with ϕ is better if *S* trusts *X* with ϕ competently than if *S* does not. Because any performance is competent just in case it issues from a skill exercised in appropriate shape and situation, the performance of *first-order trust* is competent just in case it issues from a truster's (first-order) trust skill *exercised in appropriate shape and situation.* The *substance* of a *first-order trusting competence* is just a specification of those conditions, along with an indication of what counts as sufficient reliability in those conditions.

What exactly are the *shape and situation* conditions pertinent to first-

The same goes for more mundane competences, like visual-perceptual competences: one possesses the (innermost) visual-perceptual competence if one's visual-perceptual beliefs are reliably enough correct when one is in proper shape (i.e., awake, alert) and properly situated (not in the dark, in thick fog, etc.).

[5]Darkness lies outside the bounds of the situations in which reliable performance matters for good archery. Even the best archer might be terrible in the dark.

[6]The idea that one can retain one's innermost competence to (for example) drive a car even while drunk or on slick roads comports well with the familiar idea – defended variously by Tony Honoré (1964), Anthony Kenny (1976) and Mele (2003a, 447–70) – that one can retain a general ability to do something, ϕ, even when one lacks a specific ability to ϕ.

order trusting competence? Consider first the *shape* of first-order trusting competence. Presumably, this will involve at least certain healthy levels cognitive functioning (e.g., the sort by which we might spot obvious *betrayal indicators*[7]) and which preclude various kinds of mental incapacitation (i.e., being drugged, sleep deprived and unusually pliant, etc.).

Moreover, it is plausible that one is not in proper shape for first-order trust if one is cognitively compromised due to *ex ante* manipulation or coercion. Here a distinction is needed between (i) *manipulation ex ante* into trusting (e.g., manipulation prior to one's placing one's trust[8]); and (ii) manipulation *post hoc* by the trustee, after one has placed one's trust. (One may be in proper shape – even ideal shape – to trust competently even if one's trustee happens ultimately to betray one's trust – viz., a kind of manipulation or deception *post hoc*.)[9]

Regarding the relevance of manipulation or coercion *ex ante* to shape: just as we don't test for a driving competence by checking one's reliability in conditions where (for instance) one is non-culpably misled about the correct speed limit – for example, if pranksters swapped out a 20 mph sign for a 30 mph sign, such that even the most skilled driver would accept

[7] Granted, empirical work on our capacity to spot betrayal reliably is far from settled. Early work (e.g., Kraut 1980; Vrij 2000; Bond Jr and DePaulo 2006) has suggested that reliability is not much better than chance that we can detect deception from visual cues. However, more recent studies are slightly less pessimistic. For discussion in the context of the epistemology of testimony, see Simion and Kelp (2018).

[8] For instance, suppose you are duped into trusting the medical advice of someone who presents as a doctor, but under conditions in which this deception would be undetectable even by the most cautious.

[9] This is just a corollary of the more general idea that a performance's being competent does not entail that it is successful. Compare: Even when every prerequisite is in place for a basketball player's making a free-throw, the shooter may still miss on occasion, despite attempting a competent shot. A shooter's competence, after all, is (as noted in §5) a competence to hit the target reliably enough via one's method exercised in proper shape and when properly situated. Whereas, in baseball, such a method need only be 30% reliable to qualify as a competence, an archery competence may require a more reliable method, though not an infallible method. For discussion see Carter et al. (2015) and Carter (2019).

the 30mph sign as reflecting the law – *nor*, in the case of assessing for first-order trust competence, do we check whether one performs reliably enough in conditions in which (for instance) one is manipulated *ex ante* into trusting, as when one is the subject of an elaborate prank or deception (i.e., a surprise party). These are after all atypical circumstances, not the circumstances in which we generally value the accomplishment of trusting well.

That covers *shape*. But what are the *situation* conditions pertinent to first-order trusting competence? Here I want to discuss three distinct betrayal-relevant thresholds that can vary independent of each other – namely, the extent, present in a given trust context, of the (a) *gains to the trustee* that would come from betrayal; (b) the *effort*; and (c) the *aptitude* required by the trustee to *avoid* betrayal.

(a) Gains to the trustee. In typical situations where reliable trusting is valued, we can assume that there would be *some* gains or rewards the trustee might attain through betrayal.[10] One's first-order trust skill is thus implicitly indexed to a shape/situation pair that includes the presence of *some* gains that would be reaped by betrayal. However, if we adjust the level of these gains dramatically, it then becomes less clear that we should expect even one who is skilled at first-order trust to reliably avoid betrayal when gains of betrayal are so unusually high. An example here is a situation where the gains of betrayal are ratcheted up to the point that *only* through betrayal (i.e., suppose, by *not* returning a car as entrusted to do) can the trustee save her family by fleeing an imminent threat of harm.

(b). Effort to avoid betrayal. Just as we can assume there will be *some* gains or rewards the trustee might reap through betrayal in any given situation where trust is placed, *likewise*, we should assume that avoiding betrayal will never be entirely effortless.[11] Almost always, taking care of things

[10] Note that this is compatible with it being all-things-considered valuable to cultivate in the long-term a reputation of trustworthiness.

[11] Of course, in rare cases, 'doing nothing' might be exactly what one is entrusted to do (and so in such rare cases one can trivially take care of things as entrusted by doing nothing).

as entrusted involves, on the part of the trustee, performing some *non-trivially effortful* task (often at a designated time, or by a designated time) at the exclusion of performing some other tasks.[12] However, if we ratchet up the level of effort required for the trustee to take care of things as entrusted to an abnormally high level, avoiding betrayal will then be unlikely even for a skilled truster, and thus reliability in those conditions should not be expected of a skilled truster. An example here – where effort by the trustee to avoid betrayal is abnormally high – is a situation where *A* entrusts *B* to keep *A*'s child safe while *A* is away on a work trip. As things transpire, it turns out that B can do this only by monitoring *A*'s child for a five year period (rather than 24 hours, as anticipated), during which A has been unexpectedly detained.[13]

(c). Aptitude to avoid betrayal. Finally, just as we can assume there will be *some* gains or rewards the trustee might reap through betrayal *and* some non-trivial effort level involved in avoiding betrayal, likewise, we should assume that there is always going to be some non-trivial *aptitude* level involved in avoiding betrayal. For example, entrusting someone to deliver a message doesn't require much *effort*; though it does at least require the aptitude to communicate. One without this aptitude couldn't take care of things as entrusted even though the level of effort such a task would

[12] 'Effort' needn't be limited to physical effort; more generally, the idea is that effort will involve some kind of 'exertion of the will', and this could include, e.g., cognitive effort. For discussion of effort and its connection with the will, see Bradford (2013).

[13] Of course, one might quibble here that what one is plausibly entrusted to do is implicitly *only* babysit for 24 hours or so. The question of what the *content* is that captures that which one is entrusted to do, in a given context in which trust is placed, is complex. In many ways, the philosophical problem of delineating this content dovetails with a related philosophical problem of distinguishing the *content* of that which one consents to when one consents (see, e.g., Dougherty 2013, 2019). In both cases, we may ask whether determines such facts are some combination of mental states, verbal articulations of these states, social norms, etc. But we needn't settle this general problem here (one that applies no less to theories of trust than to theories of consent) in order to see that in at least some cases, there can be something X that is both (i) something that a trustor entrusts a trustee to do; and (ii) where X involves a level of effort that is both beyond typical bounds *and* (iii) that it's being above typical bounds does not thereby undermine the fact that the trustor has trusted the trustee with X.

typically require from one is low. That said, if we ratchet up the level of aptitude required to take care of things as entrusted so that the aptitude is far outside typical bounds, we should no longer expect that one skilled at first-order trust would reliably avoid betrayal (in a situation where such a high aptitude by the trustee is required). For example, suppose *A* entrusts *B* to help *A*'s child solve all of her math homework problems before *A*'s child turns in her homework. *A*s it turns out, *A*'s homework includes a prank question by the teacher, which asks for a proof of Goldbach's conjecture.

In sum, then, the *substance* of *first-order trusting competence* – of which we should distinguish the *skill* from the complete competence – is as follows. The skill associated with *first-order trusting competence* (a skill is referred to in a performance-normative framework, alternatively, as an *inner-most* competence[14]) is a disposition to trust successfully (viz., trust such that the trustee then takes care of things as entrusted) reliably enough whenever one trusts *while in proper shape and while properly situated*. We've seen in this section how 'proper shape' and 'properly situated' should be unpacked – viz., we've spelled out the *substance* of the skill. Further, one possesses not merely the skill but the *complete competence* for first-order trust when they possess the skill (i.e., the counterfactual is true of them that they *would* trust successfully when in proper shape and properly situated) and, additionally, while *in fact* in proper shape and properly situated for first-order trust. And trust is itself (first-order) competent just in case it issues from a first-order trusting skill in these conditions.

We now have a more substantial picture of what is required for one to satisfy ECNT (from Chapter 2). But, more importantly at present, this fuller picture of the substance of first-order trusting competence is needed

[14]See, e.g., Sosa (2017, 191–2). Note that 'skill', within a performance-normative framework, is a technical term (in so far as it is meant to pick out an inner-most competence); there are other uses of 'skill' in the literature that have been unpacked differently. One way to unpack 'skill' differently is on an intellectualist account of intelligent action and intelligence states. For a notable example of an intellectualist account of skill, see Stanley and Williamson (2017). See also Pavese (2016) and (eds). Fridland and Pavese (forthcoming) for related discussion of skill in epistemology and elsewhere.

in order to understand the substance of *second-order trusting competence*.

Whereas competent *implicit trust* requires only that trust be first-order competent, competent (and by extension, apt) *deliberative* trust is *more demanding,* in that competence at that kind of performance – a performance that aims intentionally at apt trust – requires a further additional skill set, with its own shape/situation profile.

5.2.2 Second-order trusting competence

In order to think about the difference between *first-order trusting competence* and *second-order trusting competence*, consider the following case, where one's trust is first-order competent (and, indeed, even first-order *apt*), but *not* second-order competent.

> Mr. X: Mr. X, having read The Art of the Deal along with several books by Tony Robbins, fancies himself a charismatic dealmaker, overestimating his influence. Mr. X entrusts Mrs. Y with information I, in a situation within normal bounds of risk, effort and aptitude, and Mrs. Y does not betray Mr. X. Mr. X's (first-order) trust on this occasion may be apt – his successful trust manifests his competence to trust reliably enough in the shape/situation pertinent to first-order trust. However, suppose that while Mr. X in trusting Mrs. Y has trusted aptly, he very easily would have trusted inaptly. Although in entrusting Mrs. Y with information I, the risk to Mr. X is in fact not excessively high and gains of betrayal are within normal bounds, Mr. X. (with a distorted view of his charisma and influence, thanks to the Tony Robbins books) would easily have entrusted Mrs. Y with information I *outside such bounds* (e.g., had I been information that would have given Mrs. Y huge gains if divulged with little threat of her detection), in a situation where he would not have been a reliable enough truster.[15]

[15] This case is based closely on a case that appears originally in Carter (2020b, 2308).

In the above case, Mr. X's trusting is first-order competent. *And* it is first-order apt. Though, crucially, very easily, Mr. X would easily have trusted inaptly, outside of his range of sufficient reliability, given the distorted view he has of his own (first-order) trusting competence, a distorted view which precludes him from accurately gauging the risks of trusting *inaptly* that are present. In short: Mr. X is a good enough truster to trust aptly in this situation, but he thinks he's better than he is and this is why he (though not one with a more accurate view of their own first-order trusting competence) would easily have trusted aptly.

That Mr. X trusted first-order aptly on this occasion – as opposed to inaptly – accordingly doesn't owe to any awareness of his of the threshold of his own first-order competence (an awareness he lacks *ex hypothesi*), but rather just to good fortune. In this respect, Mr. X is (performatively speaking) is very much like a basketball player who is reliable enough to shoot and make a shot aptly from a particular distance, D, but who – *oblivious to the fact that they are unreliable just a few inches beyond D* – would have just as well taken a shot had they been a few inches behind D – and then even if it went in, it would be inapt.[16]

A shot taken in the above scenario is first-order competent. And it is first-order apt. But it is not *second-order competent*; the shot is not made in the light of any competent assessment *that* the shot would (likely enough) be apt. And the same goes, *mutatis mutandis,* for Mr. X's trust.

As with the case of first-order trusting competence, we can distinguish – in the case of second-order trusting competence, between (i) a second-order trusting *skill*; and a (ii) *second-order trusting (complete) competence*. The second-order trusting *skill* – which Mr. X lacks in the above case, despite possessing first-order trusting competence – is a disposition to trust not merely successfully but *aptly* reliably enough, when in appropriate shape and appropriately situated; and the second-order trusting complete competence is possessed when one possesses the skill while in fact in proper shape and properly situated. (And, then, trust *itself* is second-

[16]Compare: this is different from the player, aware of the limits of their own first-order competence, who not only shoots aptly just inside of th

order competent if it issues from the exercise of the second-order skill in these conditions).

Just as trust could be apt without being second-order competent (i.e., in the Mr X case), trust could be second order-competent *without being first-order apt.* With this in mind, contrast Mr. X. with Ms. Y.

> Ms. Y: Mrs. Y is the manager of a bank, which includes a valuable safe, which Mrs. Y makes sure to lock herself before closing up each evening. One day, Mrs. Y is called away and needs to leave this task with one of her employees. Because it is an important task, she doesn't make the decision lightly. Mrs. Y looks at her 10 employees' track records and reflects on their character – as well as on her own reliability in trusting employees in similar situation (i.e., with similar gains that would be had from betrayal, and similar effort and aptitude that would be required for the trustee to avoid betrayal) – finally selecting her most trustworthy and reliable employee, Mr. Lockit, as the employee entrusted to lock the safe. By fluke, Mr. Lockit happened to flake out this one time, forgetting to lock the safe, and thereby ruining what was otherwise an impeccable track record of trustworthiness.

The trust here that Mrs. Y places in Mr. Lockit to lock the safe is not first-order apt, given that first-order aptness requires that the trust placed be (first-order) successful – viz., it requires that Mr. Lockit *actually take care of things* as entrusted, which he does not. However, *even though* Ms. Y's trust is not first-order apt, it is *second-order competent.* Her trust is made in the light of a competent second-order assessment *that* the trust would likely enough be apt in the conditions in which the trust was placed. This time, it just wasn't apt – bad luck!

The reader might wonder whether we are being too generous in attributing a second-order trusting competence to Ms. Y in the above case. After all, her risk assessment here led her to place her trust in someone who betrayed it. One might think, 'some risk assessment that was!' The right

reply here, though, is to hold firm. It *was*, indeed, excellent risk assessment: Ms. Y *not easily* would have trusted inaptly in this situation, as she might have had she (for example) selected the employee to trust with locking up the safe randomly, or if she gave no consideration to such things as the gains of betrayal and the effort and aptitude required to avoid betrayal. But she assessed this unimpeachably, *ex hypothesi*. What bears emphasis here is that second-order competences needn't be *infallible;* they needn't be infallible any more than first-order competences do. (A basketball player can shoot the ball competently after all even when the shot misses).[17]

In sum, trust can be first-order competent *or* second order competent, *or* both. *Implicit trust* is apt so long as it is successful because *first-order competent*; the aptness of implicit trust doesn't depend on the possession or exercise of second-order trusting competence.

Deliberative trust – our central focus in this chapter – is different. For *deliberative trust* to be apt, trust must be not just first-order competent, and first-order apt, but it must also be *second-order competent*, as well as *second-order apt*.

But is deliberative trust apt *if* it is first-order apt *and* second-order apt? The answer here is 'no'. First- and second-order aptness is *necessary but not sufficient* for apt deliberative trust. To appreciate why, though, we'll need to move from thinking about the *substance* of the skills involved in apt deliberative trust, and instead look more carefully at its *structure*.

[17] See, e.g., Carter (2019) for an extended defence of this claim. See also Sosa (2007, Ch. 2). One dissenting view here is due to Millar (2009), according to whom exercising an ability or competence entails succeeding in doing the thing that the competence is a competence to do. Millar defends this view in the service of a disjunctivist account of perceptual knowledge and the perceptual-recognitional abilities that give rise to them; however, the thesis is one he maintains holds for abilities or competences generally. For criticism of Millar's position, see Carter (2019).

5.3 The structure of apt deliberative trust

Consider now the case of Sherlock:

> *Sherlock*: Sherlock trusts Mrs. Hudson to complete an important task, the trust is successful (she takes care of what he entrusted her to do as he entrusted her to do it), and the trust is apt: his trusting successfully manifested a complete first-order trusting competence. Suppose further that his trusting was aptly risk assessed at the second-order; Sherlock aptly appreciates that not easily would he trust inaptly in these conditions. But because life has gotten a bit too boring, Sherlock decides whether to *actually* trust Mrs. Hudson with the task or not by flipping a coin, and so his apt risk-assessment in this case is in fact disconnected from his apt trusting.[18]

In the above case, Sherlock's trust 'ticks' a lot of boxes. It is first-order apt, *ex hypothesi*. It is, importantly, also second-order apt. But here is the problem, there is a *disconnect* between Sherlock's first-order trust and his apt second-order awareness *that* his first order trust not easily would be inapt; we should think that the latter awareness would *bear* on Sherlock's trusting, as opposed to being an achievement in its own right that stands as a kind of 'idle wheel' in deliberative process that results in his trusting.

Compare: here Sherlock is not unlike a political leader who has read and digested an expensive and rigorous intelligence report, only to then disregard all he has learned from this report before acting. What we would say of such a leader is that their action (on the matter about which the intelligence pertained) was in no way better, *qua* action, for the leader having read the report; the report was *wasted* on the action. (And this is so even if the leader should get credit for reading and understanding the report very well!) And by parity of reasoning, Sherlock's trust is no better than were he to have *not* aptly appreciated that not easily would he trust inaptly in these conditions. This apt appreciation was *wasted* on his trusting. (And

[18] Compare with Sosa's case of Diana the huntress in Sosa (2015, 69).

this is so even if he should get credit for assessing risks aptly – viz., even if his second-order assessment is better than it would be were it not itself apt.)

Stepping back for a moment, what the case of Sherlock shows is that the following 'conjunctive" formula must be mistaken – a formula

> **Conjunctive view of the structure of apt deliberative trust:**
> (!) A's deliberative trusting B to ϕ is apt – if and only if:
>
> (i) *First-order condition:* A's trusting B to ϕ is first-order apt > (first-order condition); and
>
> (ii) *Second-order condition:* A's risk assessment (that A's trust > wouldn't easily have been inapt) is second-order apt (viz., it > manifests A's complete second-order trusting competence).

> The problem with the conjunctive view is that *S*'s trust is no better, as an instance of trust, for satisfying both (i) and (ii) than for satisfying just (i) alone whenever *A*'s satisfying (ii) *plays no role* – as it plays no role in the case of Sherlock – in *A*'s satisfying (i). The conjunctive view leaves second-order competence as a *pistologically* idle wheel, even if it is (qua knowledge) a kind of intellectual accomplishment in its own right.

But how, then, should *A*'s satisfying (i) and (ii) be related, structurally, when *A*'s deliberative trusting *B* to ϕ is apt? Recall that Sherlock *settles the question of whether to trust* by flipping a coin. But how does the coin flip *do* that; how, exactly, does Sherlock's flipping the coin settle the question of whether to trust? It does *not* do it by providing him any new information *about* whether to trust. Rather, it just plays a kind of imperative function in that it introduces (albeit, arbitrarily) a preference between alternatives (trust or not trust).[19] The problem, of course, is that the preference introduced (i.e., heads = trust, rather than forbear) is a preference entirely

[19] For a defence of the view of imperatives, generally, as playing this role in language and thought, see Starr (2020).

unrelated to the second-order risk assessment.

5.3.1 The guidance view of the structure of apt deliberative trust

So how, then, might a preference between the alternatives (trust or forbear) be introduced in a way that is *not* unrelated to the second-order risk assessment? One *prima facie* plausible – but ultimately problematic – type of proposal that we find in Ernest Sosa's epistemology would encourage us to link the first-order condition and the second-order condition with a *guidance relation*.

Here's Sosa (2017):

> The fully desirable status for performances in general is [...] is aptness on the first order *guided* by apt awareness on the second order that the first-order performance would be apt (likely enough) (2017, 96).

In a bit more detail, Sosa says:

> [...] the aptness on the first order be attained under the guidance of the second-order awareness. The performance on the first level must be *guided to aptness* through the apt second-order awareness (explicit or implicit) that the subject is in that instance competent to avoid excessive risk of failure. This would comport with the subject's apt awareness that if he performed on the first level, he would (likely enough) do so aptly [...] (2017, 96).

It should be granted that any kind of 'guidance' linking the first and second order, however we construe it (we'll return to this), would surely be missing in *Sherlock*, where Sherlock's satisfying the second-order condition, (ii), bears in *no way* whatsoever on his satisfying the first-order condition, (i). So Sosa's suggestion looks initially promising.

If we supplement the conjunctive view of apt deliberative trust with a Sosa-style guidance condition linking (i) and (ii), we get the following:

Guidance view of the structure of apt deliberative trust:
(!) A's deliberative trusting B to ϕ is apt – if and only if:

(i) *First-order condition:* A's trusting B to ϕ is first-order > apt; and

(ii) *Second-order condition:* A's risk assessment (that A's trust > wouldn't easily have been inapt) is second-order apt (viz., it > manifests A's complete second-order trusting competence).

(iii) **A's satisfying (ii)** ***guides*** **A to satisfying (i).**

In assessing this proposal, the tempting first question is "but what do you mean by *guidance*?" Sosa does not give us many concrete clues here. Of course, this needn't be problematic in itself, given that we are free to charitably interpret 'guidance' in different ways and consider which way would help us see how satisfying (ii) – viz., and thus aptly assessing risks to inaptness – really *would* make satisfying (i) better as an instance of trusting.

But I want to suggest instead is that the guidance view of the structure of apt deliberative trust is really a dead end from the start.[20] Here is the problem, in a nutshell. Stipulate that a truster *A* has just satisfied the second-order condition on apt deliberative trust – viz., that *A*'s risk assessment (that *A*'s trust wouldn't easily have been inapt) is second-order apt (viz., it manifests *A*'s complete second-order trusting competence). The output of this apt risk assessment is a known proposition *about* whether one's trust would be (first-order) apt.

So far, so good. Such knowledge settles a 'whether *p*' question: specifically, the question whether *A*'s trust *would* be likely enough apt. But, importantly, that is not the *kind* of question a deliberative truster is deliberating about. Here we should distinguish between *whether p* questions and *whether to p* questions. In deliberating whether to trust, we are deliberating about a *whether to p* question, not a whether *p* question. This matters because what counts as *settling* a whether to *p* question and a whether *p*

[20] Importantly, my reasoning here does not generalise beyond pistological assessment. In fact, for all I will suggest, it might be that Sosa is entirely right that 'guidance' is exactly the kind of link needed to connect the relevant first-order condition and second-order condition *in the case where the performance at issue is that of apt (deliberative) judgment*.

question differ. The difference is that the latter is always settled by belief (or knowledge), but plausibly the former is settled only when one forms an *intention*.[21] In this case, the relevant intention would be an intention to trust or forbear.

Bearing the above in mind, an argument against the guidance view of the structure of apt deliberative trust begins to take shape. Condition (iii) on the guidance account is satisfied if and only if the aptness on the first order be attained under the guidance of the second-order awareness that *A*'s trust wouldn't easily have been inapt. But second-order awareness that *A*'s trust wouldn't easily have been inapt would plausibly succeed in guiding one to apt first-order trust only if it that second-order awareness would settle a *whether to p* question: the question of whether to trust or forbear. Otherwise, the matter of whether one forms the intention to trust, rather than to forbear, remains underdetermined.

However – and putting this all together – although second-order awareness that *A*'s trust wouldn't easily have been inapt suffices to settle a whether-p question (i.e., the question of *whether the trust would easily have been inapt*), it does not suffice to answer a whether-to-p question (the question of whether to trust or forbear). Second-order risk assessment is, therefore, insufficient to guide one to apt trust. It is, after all, compatible with one's both forming an intention to trust and with one's refraining from forming that intention. But if the foregoing is right, then condition (iii) won't in principle be satisfied by a deliberative trustor: this is because, put simply, satisfying condition (ii) isn't *the sort of thing* that we should expect would suffice for guiding one to satisfying condition (i).[22]

[21] For discussion on this point, see, e.g., Shah (2008).

[22] One might object to this reasoning as follows: even if we grant that it is an intention rather than merely a belief or knowledge that would suffice to settling the whether to *p* question at issue in deliberative trust (i.e., whether to trust or forbear), it remains that satisfying condition (ii) could of course *cause* one to form such an intention. And thus satisfying condition (ii) could cause one to settle the question of whether p. But the reasoning underlying this objection overgeneralises. After all, as we saw in *Sherlock*, the flipping of a coin could also cause one to form such an intention.

In sum, the problem with the guidance view is that satisfying (ii) just provides information. But this risk assessment actually adds value to the trust only if the risk assessment is somehow connected in the right way to the agent's forming an intention (or not) to trust. Otherwise, one's satisfying (ii) will not suffice to have settled the whether-to question that distinguishes deliberative trust from mere first-order trust that is not deliberative. This is so even if satisfying (ii) gives us knowledge about whether our first-order trust would not too easily have been inapt.

5.3.2 A basing view of the structure of apt deliberative trust:

So the guidance view of the structure of apt deliberative trust of doesn't pan out, despite its initial promise. Where should we go from here?

A helpful way forward will be to notice a structural analogy between the *Sherlock* case and another kind of case – well-studied in mainstream epistemology – where we find a parallel kind of general structure: viz., where subject possesses some belief or knowledge K, that would help improve the quality of S's ϕ-ing *were S to use K* in ϕ-ing, but where ϕs without doing so.

The analogy I have in mind here involves cases of *bad basing* in epistemology. The basing relation is what it is that makes the difference between two ways you might have a 'good reason for a belief' you have – and which line up with the distinction in epistemology between *propositional justification* and *doxastic justification*.[23] In the first case, suppose you believe it is raining via reading tea leaves, and that you *also* believe that the weather forecast says it is raining. In this case, you *have a good reason* (i.e., that the weather report says it is raining) for believing that it is raining; the proposition that it is raining is *propositionally justified* for you. And yet, you are *not* doxastically justified in believing that it is raining; even though you

[23] See Korcz (2019) and Carter and Bondy (2019) for overviews of recent work on this distinction in connection with basing. For a notable exception to the orthodoxy of explaining doxastic justification in terms of propositional justification, see Turri (2010); cf., Silva (2015).

believe it is raining and are propositionally justified in believing that it is raining. This is where basing comes in: the reason you're not doxastically (or fully) justified in believing that it is raining is that you did not actually base your belief that it is raining on the good reason you had for believing it is raining, and which propositionally justifies you in believing it. Instead you based your belief on a bad reason for believing it is raining, i.e., that the tea leaves said so.

Bad basing can occur both with or without one's possessing propositional justification (in the latter case: one simply bases one's belief that *p* on a bad reason for believing *p*, *and* S is not propositionally justified in believing that *p* in the first place).[24] The kind of case that exhibits parallel structure with *Sherlock* occurs *with* propositional justification.

A propositionally justified bad baser *has* propositional justification but doesn't use it to improve the quality of the belief that the target proposition is true; as a result, the belief is not doxastically justified. Likewise – and here is the structural analogy – Sherlock *has* an apt awareness that his trusting not too easily would be inapt but he doesn't use it to improve the quality of his trust. As a result, his trust is not fully apt.

The reader should now be able to see where this is headed. What the propositionally justified bad baser needs to do in order to use the propositional justification she has to upgrade the quality of her belief is to *base* her belief on the good reason she has that propositionally justifies it.

By parity of reasoning, the idea I want to now explore is that a parallel structural diagnosis is promising in the case of Sherlock: to a first approximation, what Sherlock needs to do in order to use his apt second-order risk assessment to improve the quality of his trust at the first order is to *base his trusting on his apt risk assessment.*

[24] In such a case, imagine that a thinker believes a conspiracy theory on the basis of some reason R, which is a bad reason to believe the conspiracy theory. Here, the subject not only bases her belief on a bad reason, but even more, the belief she holds (i.e., the belief that some conspiracy theory is true) is also not *propositionally* justified for her, given that we may suppose she lacks any good reasons for thinking it's true.

The structure of the proposal, then is that we replace a 'guidance' condition with a 'basing' condition. Rather than to say that what 'connects' the first and second-order aptness, in cases of apt deliberative trust, is that the latter 'guides' one to the former, a more promising idea is to require that the former is 'based' on the latter. The proposed structural analysis is accordingly as follows:

> **Basing view of the structure of apt deliberative trust:** A's deliberative trusting B to ϕ is apt – if and only if:

(i) *First-order condition:* A's trusting B to ϕ is first-order > apt; and

(ii) *Second-order condition:* A's risk assessment (that A's trust > wouldn't easily have been inapt) is second-order apt (viz., it > manifests A's complete second-order trusting competence).

(iii) **A's satisfying (i) is based on A's satisfying (ii)**

What makes this view initially promising is that it maintains parity of structure with what is already a well-established way of thinking about how to 'link' a (ii)-type condition with a (i)-type condition – through a basing relation – in a way that 'gets the goods' when it comes to showing how the former stands to *improve* the quality of the latter. In epistemology, there is, unsurprisingly, disagreement about how to characterise the *nature* of the basing relation[25]; but it bears emphasis that there is remains consensus *that* basing a belief on a good reason improves the belief's quality.

In order to develop and defend a plausible *basing view of the structure of apt deliberative trust*, we need to address a potential issue that we encounter right out of the gate. The issue arises given a *disanalogy* between (i) epistemic basing, where, when all goes well, *beliefs* are based on good reasons; and (ii) practical basing, where when all goes well *actions* are based on good reasons.

[25] For some overviews, see, e.g., Korcz (2019) and Carter and Bondy (2019, Ch. 1). For a recent and sustained treatment of epistemic basing, which critiques recent work, see Neta (2019).

The disanalogy is that the kinds of reasons in the two cases are different. In both kinds of cases, the kind of reason that a belief (or action) *X* must be *based on* in order to improve the quality of *X* is a *normative reason* (viz., alternatively called a justifying reason) a reason that 'favours' *X* (see, e.g., Alvarez 2010). It must as such be a *good reason* for *X*.

But *what makes reason a good reason* differs when the reason is a reason for belief as opposed to a reason for action. In the epistemic case, there is a relatively simple – even if not entirely uncontroversial – story: a reason is a good reason for a *belief* if it supports the belief's being *true*. And this is because belief is the kind of attitude that, by taking it up, we are aiming to get things right.[26]

Good (i.e., normative) reasons for action are different. A reason is a normative reason for acting because it *favours* someone's *acting* in some particular way (as opposed to supporting the truth of a proposition).[27] Here is why things are trickier in the case of reasons for action: Whereas the *basis of the normativity* of (normative) epistemic reasons is relatively straightforward (i.e., the line is that in believing we aim at truth), the basis of the normativity of practical reasons is much more contentious.

This matters for us presently because it means that if we are going to say that *A*'s apt awareness that *A*'s trust wouldn't easily have been inapt is not a mere explanatory or a mere motivating reason[28] but a full-blown *normative reason* for trusting (one that by basing one's trusting on that normative reason one *thereby improves one's trust*) we need some explanation of what the *basis of that normativity is*.

What would constitute a good explanation here? As Maria Alvarez (2017)

[26] See Shah (2003) Shah and Velleman (2005), Velleman (2000a), Wedgwood (2002), and Whiting (2013b) for defences of this idea; and, for discussion, Chan (2013). Cf., Gluer and Wikforss (2009) and Steglich-Petersen (2006).

[27] The favouring here is *pro tanto*: there might be other considerations that count against the action. (Compare: in epistemology we say that a reason leads defeasible support to a belief).

[28] For discussion of normative reasons in connection with explanatory and motivating reasons for action, see, e.g., Alvarez (2009) and Raz (2009).

notes, one important desideratum that needs to be met is the following: Any story we give of the basis of the normativity of normative reasons for action ought to be able to account for 'the relationship between the normativity of reasons *and the capacity that reasons have to motivate agents to act'*. This is best interpreted, according to Alvarez, as a desideratum an account of normative reasons for action meets only if the account can explain (in short) how my having a normative reason for me to perform an action can (i) motivate me to act, and (ii) to act *for* that reason (2017, sec. 2).

Putting this all together then, it looks like the following is the case. If we are going to vindicate the claim that a deliberative truster's second-order apt awareness that her first-order trust would likely enough be apt is a *normative reason* for trusting – which it needs to be if trust is to be improved by being based on it – then we'd need some story for why a truster *for whom this is a normative reason to trust would be capable of being motivated to trust for that normative reason*.

One 'shortcut' for securing this result would be to throw all-in with a strong Humean theory of normative reasons, according to which something is a normative reason for S to ϕ only if S has a desire that would be served by her ϕ-ing. Part and parcel with this idea is that normative reasons, given their connection with desires, are intrinsically motivating – viz., what is called 'reasons internalism'.

The Humean theory of reasons, and 'reasons internalism' which is closely associated with it, are deeply controversial.[29] Fortunately, there is a way to get everything we want without needing to appeal to any thesis that applies to all normative reasons for action as such.

Here is the broad idea in outline. Regardless of whether all normative reasons bear any essential connection with motivation – this is where a lot of the quibbling lies – we can still offer a relatively straightforward expla-

[29] See, for discussion, Williams (1979), Wong (2006), Shafer-Landau (2003, Ch. 7), and for an overview of the debate between reasons internalist and externalists, see Finlay and Schroeder (2017).

nation for how a deliberative truster's apt awareness that her first-order trust would likely enough be apt is a normative reason for her to trust; we do this by appealing to a feature distinctive of trust's being deliberative rather than merely implicit in the first place – viz., to a deliberative truster's *intentional aim to trust if and only if trusting would be apt*. Recall, from Chapter Three, that the right way to characterise the *way* we aim at aptness generally, when we deliberate about *whether* to make the relevant attempt (or *not*) in the first place is in terms of a *biconditional aim*: to ϕ iff one's ϕ-ing would be apt. In the special case of deliberative trust, the biconditional aim is to trust iff one's trust would be (first-order) apt. Because the deliberative truster is already intentionally aiming at aptness in this way, we can see how her knowledge that her trust if placed wouldn't easily have been inapt (when combined with her intentional aim to trust if and only if that trust would be apt) would have the capacity to motivate her to trust for this reason. Thus, we can see how the deliberative trustor's apt second-order awareness that her trust would likely enough be apt is capable of playing the related roles that we should expect it to play *qua* normative reason for trusting: that is, we see how – when one acquires this reason in the context of deliberative trust – it has the capacity to motivate her to trust rather than forbear, and to trust *for* this reason.

We're now in a position to put all the pieces together. Recall, again, the problem that faces the guidance view of the structure of apt deliberative trust. In a nutshell, the problem with the guidance view is that possessing the information one possesses when one satisfies condition (ii) actually adds value to the trust only if the risk assessment is somehow connected in the right way to the agent's intentionally trusting. Otherwise, one's satisfying (ii) will not have settled the whether-to-p question that distinguishes deliberative trust from mere first-order trust that is not deliberative.

The basing view offers a different and more complete picture, one that tells us exactly how satisfying condition (ii) is used to both answer *this* question and to improve your trust in the course of doing so.

On the basing view of the structure of apt deliberative trust, *you* settle the whether-to-*p* question, through second-order risk assessment; when that

risk assessment is apt, you then attain counterfactual knowledge that you not easily would have trusted inaptly. This knowledge that her trust if placed wouldn't easily have been inapt (when combined with her intentional aim, in deliberatively trusting, to trust if and only if that trust would be apt) has the capacity to motivate her to trust *for* this reason.

When you then do trust for this reason, you base your trust on the good normative reason you have for trusting, and through this basing, you use the good risk assessment to improve the value of your trust. In this way, you improve the value of your trust by basing it on apt second-order risk assessment in a way that is structurally analogous to how one might improve the quality of a belief by basing it on a normative epistemic reason.

5.4 Concluding remarks

Let's sum up. This focus of this chapter has been deliberative trust, and what is involved for such trust to be apt – both substantively and structurally. On the 'substance' front, we distinguished between first-order and second-order trusting competences – and importantly – the SSS conditions that are distinctive to each. On the 'structure' front, we considered how not just first- and second-order competence, but first- and second-order *aptness* must be related to each other when deliberative trust is itself apt.

In a bit more detail, we outlined the kind of answer that would be most naturally recommended by Ernest Sosa's approach to telic normativity which powers his virtue epistemology. Drawing inspiration from that approach, we should expect the two levels would be linked by the nexus of 'guidance' – viz., the deliberative truster would be guided to trusting first-order aptly by her apt second-order risk assessment that her first-order trust wouldn't easily have been inapt. As we saw, though, even if appealing to guidance to 'connect' the two levels move works in virtue epistemology, it runs in to problems as a move within the theory of the structure of apt deliberative trust. In place of 'guidance', I've opted in this chapter for 'basing'. On the view advanced, A's deliberative trusting B to ϕ is apt

if and only if (i) *A*'s trusting *B* to ϕ is first-order apt; (ii) A's risk assessment (that *A*'s trust wouldn't easily have been inapt) is second-order apt (viz., it manifests *A*'s complete second-order trusting competence); and (iii) A's satisfying (i) is based on *A*'s satisfying (ii). This view, I've argued, not only avoids problems that faced a guidance view, but it also offers us the resources to neatly explain how the quality of apt trust is improved through second-order risk assessment.

With the answer to the Substance and Structure Question we began with at the start of this chapter, we've now got a full view of what the highest grade of trusting *demands* of us. But what does it *permit*? Put another way: What kind of risks to the inaptness of trust can the convictively apt truster *non-negligently* ignore? This is the question that will be taken up in the next chapter.

Chapter 6

Trust, Risk, and Negligence

6.1 Introduction

Consider the following paradigmatic case of apt deliberative trust.

> LOAN PAYMENT: *A* deliberatively trusts *B* to pay back a (modern, online) financial debt, which *B* repays as entrusted to do. Let's assume further that all conditions for apt deliberative trust (i.e., convictively apt trust) are met; that is, *A*'s trusting *B* to pay back the loan is (i) *first-order apt*; that (ii) *A*'s risk assessment (that *A*'s trust wouldn't easily have been inapt) is *second-order apt* (viz., it manifests A's complete second-order trusting competence); and (iii) *A*'s satisfying (i) is *based on A*'s satisfying (ii).

Notice that – in this paradigmatically good case – *A*'s trust *minimises more risk* than were *A*'s trust *merely* (first-order) apt. This is because *A*'s trust minimises not only *risks to the trust's (first-order) success* but – given that, in LOAN PAYMENT, *A*'s trust is (*ex hypothesi*) based on apt second-order risk assessment – A's trust also, thereby, minimises *risks to the trust's (first-order) aptness*.

But *to what extent* should we expect that a convictively apt truster heed

risks to (first-order) aptness of trust? This turns out to be a hard question to get right. The difficulty of this question comes into focus when we consider cases where risks to trust's first-order aptness are simply 'inherited' from risks to the obtaining of 'background conditions' for trusting.

Suppose, for example, that while *A* is deliberating whether to trust *B* to pay back the loan, there is, unbeknownst to *A* and *B*, a secret war on the cusp of breaking out. Two rogue leaders – armed to the teeth with powerful bombs – are themselves debating whether to engage in all-out war (igniting their entire arsenal, which would destroy the world) or to agree to a truce. The rogue leaders decide to make this decision by flipping a coin: heads = war, tails = truce. The result, fortunately, is tails: so no war. Meanwhile, *A* (at that moment) signs the loan to *B*, and *B* as expected pays it back – both *A* and *B* remain oblivious to the disaster that nearly took place, but which did not.

Obviously, the war could *very easily* have broken out. It came down to a coin flip. And, *had the war broken out*, then *A*'s trust would have been (trivially) inapt, given that *A* and *B* would, on this scenario, have both perished. *But,* is the mere modal proximity of the disastrous war enough to spoil the quality of A's trust in the above case, when the war in fact *doesn't* transpire? Put another way, does *A*'s trust fall short of being convictively apt *simply* because this disaster – a disaster that would have surely rendered all trust inapt (because moot) – could so easily have taken place?

Plausibly, not. And the inclination to answer 'no' here sharpens when we swap out deliberative trust for other kinds of aimed performances. For instance, suppose that while *A* is deliberating whether to trust *B* with the loan, *B* is (entirely unrelatedly) picking a warded lock, in doing so aiming intentionally to unlock the lock aptly. Applying expert precision, the lock opens. Is *B*'s lock-picking performance of less quality, *qua* lock-picking performance, than otherwise simply because, very easily, *that* performance could have ended up (trivially) inapt (given that, very easily, the rogue leaders could have agreed to war rather than to truce?) It seems not. The lock-picking performance quality didn't seem to suffer at all.

The thesis now in play (which we will refine later) looks like the follow-

ing: *there are at least some kinds of risks to the inaptness of trust that a convictively apt truster could non-negligently ignore.* These seem to include *at least* (though perhaps not exclusively) risks to the inaptness of trust that are implied by risks to the *basic preconditions* for trusting at all. We'll refine this more as we go on.

But already we are in a position to pose a philosophical problem. Call this the *Specific Non-Negligence Question*, 'specific' because it concerns exclusively the performance of *trusting*.

> **Specific non-negligence question:** What kind of risks to the inaptness of trust can the convictively apt truster *non-negligently* ignore?

Any decent answer to the Specific Non-Negligence Question will need to account for – in some principled way – the relevant difference between the war-risk variation on LOAN PAYMENT and (e.g.,) the case of MR. X (Chapter 5). And even more, such an answer should also be able to deal with intermediate kinds of cases; suppose we substitute 'risk of world-ending war' with 'risk of sabotage to electricity grids necessary to support online banking payments', which would make paying back debt online impossible.

It doesn't seem promising that we would be able to provide a principled as opposed to ad hoc answer to the specific non-negligence question unless we first answer a more basic question about risk and negligence, what I'll call the *general non-negligence question*:

> **General non-negligence question:** What kind of risks to the inaptness of any performance, ϕ, with an aim, A_ϕ, can the fully apt ϕ-performer non-negligently ignore?

The remainder of this chapter will proceed as follows. First, I will criticise the kind of answer Ernest Sosa has given to the General Non-Negligence Question, and I will then propose and defend a very different kind of alternative, one that appeals to the concept of *de minimis risk*.[1] This answer

[1] The basis for §§6.3-4 is the theory of *de minimis* risk developed in Carter (2020a).

to the General Non-Negligence Question will then be applied in the service of defending an answer the Specific Non-Negligence Question, one to which the idea of *de minimis risks* in trusting play a central role.

6.2 Sosa's answer to the general non-negligence question

According to Sosa (2017), there is an important distinction *within the class of things that could cause any aimed performance to fail*, between:

(i) the kinds of things a fully apt performer must heed in order to safeguard against credit-reducing luck; and

(ii) the kinds of things he or she is free to *non-negligently* assume are already in place.

Let's look at the first category. As Sosa puts it, an athlete, in order to meet the second-order (i.e., reflective) competence condition on fully apt performance, needs to:

> [...] consider various shape and situation factors: how tired he is, for example, how far from the target, and so on, for the many shape and situation factors that can affect performance (2017, 191).

For example, even if a basketball player – say, Steph Curry – shoots a jump shot aptly, his shot isn't fully apt if he easily could have shot inaptly *because* he easily could have been too tired, or easily could have been shooting unaware from outside his range of sufficient reliability. Tiredness, distance from the target, etc., are factors pertinent to *basketball*.

Let's now look squarely at Category (ii):

> But there are many factors that he need not heed. It is no concern of an athlete as such whether an earthquake might hit, or a flash tornado [...] and so on. As an athlete, he is *not negligent* for ignoring such factors (2017, 191, my italics).

And such things are of 'no concern' to the athlete, as such, even though earthquakes, tornadoes (as well as earth-destroying bombs and massive power outages) are the sort of things that *could* spoil a performance if they in fact materialised.

Category (ii) – viz., the kinds of things an apt performer can *non-negligently* assume are already in place – corresponds to what Sosa is calling *background conditions*.

> **Background conditions:** Background conditions for a given performance, ϕ, on Sosa's view, are entailed by the presence of pertinent seat, shape and situation conditions that correspond with a complete *ϕ-competence*; they are conditions that must hold if the relevant 'S' [seat/shape/situation], corresponding with a complete ϕ-competence, is in place at the time of a subject's ϕ-ing.

To make Sosa's idea of background conditions more concrete, just consider as a reference point one's figure skating competence. The *shape* pertinent to a figure skating competence will include being alert, awake, etc., and *this* entails whatever is necessary to support being alert, awake, etc. – viz., among other things, a properly functioning thalamus. And so a properly functioning thalamus is thus a shape-relevant background condition for competent ice skating.

The *situation* pertinent to ice skating competence includes a sufficiently flat ice rink; necessary for the existence of this rink is that a large sinkhole under the ice rink does not suddenly materialise as a result of water eroding the underlying rock layer. (Even more dramatically, the existence of the *earth* is necessary to support the obtaining of *any* seat/shape/situation conditions of interest to human performances of any kind).

As Sosa sees it, the *quality* of (say) a skating routine to Carmen that includes three triple Axels isn't going to be spoiled in any way if the skater who lands these jumps flawlessly was oblivious to the fact that they nearly had a sudden aneurysm while skating which would have wiped out the 'thalamus possession' background condition to being in proper shape, or

that they were oblivious to the near possibility that, deep underground, the chemical dissolution of carbonate rock via the Karst process *nearly* but did not cause a sudden sink-hole under the rink. (Or, for that matter, the near possibility of world-destroying bomb.) Thus, though background conditions for any performance must obtain for competent performance, they needn't *obtain safely* for that performance to be fully apt. And this is so even if the unsafety of background conditions implicates a risk to the aptness of a performance.[2]

We are now in a position to frame Sosa's answer to the *general non-negligence question.*

> **Sosa's answer:** A fully apt performer can't non-negligently ignore risks to the inaptness of a performance, ϕ, *unless those risks are due to the unsafety of* ϕ-relevant *background conditions*.

6.3 An underdetermination problem for Sosa's answer to the general non-negligence question

The answer Sosa gives to the general non-negligence question is meant to show precisely *how* we can distinguish between (i) the kinds of things a fully apt performer must heed in order to safeguard against credit-reducing luck; and (ii) the kinds of things he or she is free to *non-negligently* assume are already in place.

But here is the problem. The *criterion* Sosa offers for making this distinction – a criterion framed in terms of (performance-indexed) *background*

[2] Let's apply this answer to the lock-picking example from §6.1. The complete competence to pick a warded lock corresponds with a seat/shape/situation profile that represents the conditions under which good lock picking performance is of value. And the obtaining of any of these S/S/S conditions entails that certain other things obtain. You can't be suitably alert if you don't have a brain. You can't be in a situation featuring suitable light and ambient oxygen (assume these are appropriate lock-picking conditions)

conditions – doesn't always succeed in sorting risks to inaptness neatly in to either one of these two categories. But if that's right, then as a candidate answer to the general non-negligence question, Sosa's criterion is not a satisfactory one.

To flesh out this worry, it will be helpful to focus on an example that features what is, for Sosa, going to count as a very typical kind of background condition: 'normal atmospheric pressure'. Consider that the obtaining of normal atmospheric pressure is going to qualify as a background condition on Sosa's view for almost any athletic performance type, given that one cannot be in proper shape without the presence of ambient atmospheric pressure; it is a precondition for breathable oxygen.

Thus, Sosa's view implies that a performer – say, a basketball player – ought to be able to shoot fully aptly while ignoring entirely risks to the presence of normal atmospheric pressure, even when such risks to the presence of normal atmospheric pressure are modally close.

So far, so good. But here is where things complicate quickly. Just consider that *dips* in atmospheric pressure are well known to lead to one's shape being compromised. Dips in atmospheric pressure pose a risk to joint stiffness (bad shape) just like – for example – the more familiar experience of burning, lactic acid build-up carries with it the risk of fatigued muscles (bad shape).

However, when thinking about atmospheric pressure in *this* way, though, it looks as though a fully apt athletic performer could non-negligently ignore nearby threats to normal levels of atmospheric pressure *only if* they can also non-negligently ignore more mundane threats to being in proper athletic shape, including the burning sensation of lactic acid buildup[3], pains in one's muscles, etc. But these are exactly the kinds of things Sosa

[3]Sensitivity to the kinds of sensations, such as lactic acid build up and the burning experience that usually corresponds with it, can be critical to understanding one's physical limits and when one is approaching them; such sensitivity plays an important role in some athletic practice regimens, including competitive swimming, where these sensations are closely monitored so that athletes are prepared to adjust strategy appropriately. See, e.g., McNarry, Allen-Collinson, and Evans (2020).

thinks a fully-apt performer *can't* be oblivious to.

Thus, when we ask, "can a fully apt athlete non-negligently ignore risks to inaptness posed by dips in atmospheric pressure?" The answer for Sosa seems unclear: his full proposal suggests 'yes' in one sense, 'no' in another.

To take another example where Sosa's view seems to generate conflicting verdicts, consider the way professional chess players monitor their glucose levels through nutrition during games. The kind of shape appropriate to elite chess includes alertness and mental acuity, the obtaining of which *entails* normal glucose levels of the very sort it is (within the standards of professional chess) taken to be negligent *not* to monitor – especially in classical format games which may last 6 hours.[4] Here again, when we ask "can a fully apt chess player non-negligently ignore risks to inaptness posed by drops in glucose levels?" the answer seems to be both 'yes' and 'no'.

In the absence of some kind of principled rule for how to adjudicate these kinds of cases which seem to fall into both categories, assessments one way or the other will end up arbitrary.

In what follows, I want to propose an entirely different approach to answering the General Non-Negligence Question – one that sidesteps the above problem by incorporating into a theory of full aptness some insights from the (i) the theory of social norms; (ii) and the literature on *de minimis* risks. This theory will then be applied to the case of apt deliberative trust specifically.

6.4 *De minimis* normativism

Here is the core statement of view I will now defend, what I call *de minimis normativism*:

De Minimis Normativism (DMN): A fully-apt performer

[4] See, for instance, recent studies reported by Alifirov, Mikhaylova, and Makhov (2017).

> can't non-negligently ignore practice-relative risks to the inaptness of a given performance that occurs within that practice-type, except when these risks count as *de minimis* with reference to practice-sustaining rules.

This is my answer to the General Non-Negligence Question, and it avoids the problems that were shown to face Sosa's answer.

In order to see how the view works and can get results, several pieces of terminology need clarified. The first concerns *practice-sustaining rules*. Let's define generally – in a way that abstracts from athletic and epistemic domains – a 'practice' as a way of doing things and a 'rule' as a prescriptive principle[5] or standard of conduct.[6] Prescriptive rules (hereafter, 'rules') can be primary or derivative (a distinction that we will return to).[7] To a first approximation, primary rules say 'do *X*' or 'don't do *X*'. For example: don't break promises.[8] Derivative rules are generated by primary rules and take the form: 'do what a person disposed to satisfy the primary rule would do'. (Example: try to bring it about that you don't break promises.) Rules are important to practices: they 'hold practices together'. But *how* do they do this?

A straightforward recent answer that has been defended by John Turri (2017) is value-driven:

> **Practice-sustaining rules**: A rule normatively sustains a practice if and only if the value achieved by following the rule explains why agents continue following that rule.

"Don't break promises" is a sustaining rule for many kinds of practices: the value achieved by following this rule explains why clergy as well as bankers continue to follow it. Yelling "bingo" if and only if you have a bingo is a

[5] For discussion on the difference between prescriptive and evaluative norms, see Simion et al. (2016) as well as Chapter 1.

[6] Here I am following John Turri (2017, sec. 1).

[7] See Williamson (2016) and Simion et al. (2016) for discussion.

[8] This is the example typically used by Williamson (2016) to characterise primary norms.

practice sustaining rule just for bingo: the value of doing this explains why players of bingo keep doing this.

A practice might have many rules, though only some of these play the role of sustaining it, by leading to 'reproduction via value produced' – alternatively, by having *reproduction value*. A simple (albeit imperfect) heuristic for assessing whether a rule has reproduction value is to check whether it has derivative rules that themselves have reproduction value. If not, then this counts against the rule itself being a rule that sustains the practice, as opposed to one that merely features in the practice.

Many practices *include* performances. They do so when performances are prescribed, in certain conditions, by rules that sustain the practice. For example, the practice of archery includes the performance of shooting an arrow at a target. The practice of playing chess includes the performance of castling to defend the king. The practice of inquiry includes belief.[9]

Practice-relative risks to the inaptness of a performance, within a given practice, can now be defined in terms of practice-sustaining rules as follows:

> **Practice-relative risks**: Risks to the inaptness of a performance, ϕ, within a given practice ψ, are ψ-relative risks if and only if, were the performance inapt, it would constitute a violation of (at least one) primary ψ-sustaining rule or rules.

Consider again basketball and the risk a shooter runs when she is *barely inside* her reliability threshold and, oblivious to this, shoots the shot aptly.[10] In the nearest worlds where *that* shot is inapt, a primary ϕ-sustaining rule is violated. After all, these are worlds where the shooter steps just a few inches back before shooting. And it is a primary practice-sustaining rule

[9] Recall that this idea also featured in our discussion in Chapter Four, which distinguished performances from the wider practices in which they feature. In the epistemic case, we distinguished the performance of believing from the wider practice of inquiry, and the (as will be relevant again in this chapter) the performance of trusting from the wider practice of cooperation.

[10] For reference: this case is structurally analogous to the case of Mr. X. from Chapter 5.

in basketball that, *ceteris paribus*, you should not shoot too far from the basket, beyond your sufficient threshold for reliability.[11] (Note: this rule has its own derivative rules that have reproduction value – viz., 'check how far out you are before you shoot!'.) But what makes a risk to the aptness of a performance *practice-relative*, with respect to a practice ϕ, is whether, *were it inapt*, a primary ϕ-sustaining rule would be violated.

Of course, when basketball sharpshooter Steph Curry is about to take a shot, when there just so happens to be hungry beaver outside chewing on a utility pole that controls the lighting of the arena, there is (like there would be were Steph nearly outside his reliability threshold) *also* a modally close risk to the aptness of the shot. But, crucially, the risk to the aptness of the shot posed by the beaver chewing on the electric pole is not a *practice-relative risk*. In the nearest world in which sudden darkness spoils the aptness of Curry's shot, it's *not* the case that Curry violates any plausible primary basketball practice-sustaining rule. 'Don't shoot in the dark' is not a good candidate for such a rule. It lacks any obvious derivative rule with reproduction value. (After all, it's implausible that the value achieved by following some rule taking into account *immediate darkness* possibilities explains why agents continue to follow any such rule. On the contrary: the value achieved by *ignoring* such possibilities would explain why players carry on ignoring them.)

The final key component of the view concerns the *de minimis* risk proviso that features in DMN. The phrase *de minimis* derives from the Latin sentence *de minimis non curat lex*, which translates (roughly) to 'The law should not concern itself with trifles' (e.g., the crime of stealing a penny).[12] In decision theory, risks are termed *de minimis risks* whenever

[11] Within the basic rules of basketball (e.g., a shot counts as 2-points, you can not walk while carrying the ball) it is a practice sustaining rule that you not shoot willy nilly any time you have the ball. The value of shooting (all else equal) only high percentage shots explains why this continue to be followed. For related discussion from Sosa on the impropriety of 'blind shooting', in the context of discussing why shooters *aim* to shoot not only successfully but aptly, see Sosa (2015, 71–72).

[12] See Peterson (2002, 47).

they are judged to be so 'small' that they should be ignored.[13] The concept is an especially important in health and environmental decision making.[14]

So why, exactly, is a proviso of this sort needed in (†)? After all, if the account *already* gets the result that full aptness isn't undermined by the nearness of the beaver/darkness possibility – given that this risk to the inaptness of the shot isn't a *practice-relative* risk – then isn't the inclusion of a *de minimis* proviso redundant?

The answer here is 'no'. Whereas there is no (in Turri's terms) 'reproduction value' achieved in basketball by attending to immediate darkness possibilities, there *is* reproduction value achieved by not shooting when too tired. Usually this is done in basketball by following derivative rules such as 'keep an eye on tell-tale signs that lactic acid is building up[15]' etc. But there are rarer sources of tiredness in basketball, viz., ingesting halothane gas emitting from a small hole in the court. Suppose that Curry unfortunately does exactly this, which causes him to become tired *without* lactic acid buildup, *almost* tired enough that his shooting form is compromised, but (beyond his ken) it's not. *Within* his sufficient reliability threshold, Curry then shoots and makes, oblivious that he was nearly too tired to shoot with reliable enough form *because* he is oblivious to the fact that he's ingested what is nearly a reliability-compromising dose of halothane gas.

Without some kind of *de minimis* proviso in DMN, it looks like the account in DMN would problematically lump the halothane gas version of the case with the lactic-acid *rather* than with the beaver version of the case, as one where full-aptness is undermined. But this is a bad result: it seems after all that the holothane and beaver risks – the monitoring for

[13] See Peterson (2002) for discussion.

[14] See, for example, Sandin (2005), Rulis (1986), Rhodes et al. (2011), Whipple (2012), and Mumpower (1986).

[15] Note that this is an example of a rule that one who is disposed to comply with the primary rule would aim to comply with. It is derivative because it prescribes a way of attempting to comply with the primary norm rather than prescribing simply that the primary norm be complied with.

each of which seems equally irrelevant as the other is to quality basketball – should stand and fall together, *even though* monitoring for signs of tiredness is itself prescribed by derivative practice-sustaining rules and monitoring for signs of a sudden power-outages is not.

With this in mind, the formulation of the *de minimis* proviso I want to now defend, as a component of the wider principle DMN, is the following:

> **De minimis proviso**: For any practice ψ, a ψ-relative risk, R, to the inaptness of a given performance, ϕ, is *de minimis* with reference to ψ, if and only if the safety of ϕ against R can't be easily increased through adherence to one or more derivative ψ-sustaining rules.

The safety of a performance against a risk concerns how easily the risk event would materialise.[16] Doing something to *increase* the safety of a performance against a risk is to do something that makes it less easy (holding fixed that you've done that thing) that the risk event will materialise – viz., that, holding fixed that you've done that thing, the risk event materialises in further-out worlds than before.[17] The *de minimis* proviso above says that a risk is *de minimis* when, specifically, adhering to derivative rules of the relevant practice is *not* among the things that can easily increase the safety of a performance against a practice-relative risk.

Two clarifications are in order here. First, regarding the 'derivative' qualifier. Here is why it matters. Remember: a risk is practice-relative only if, were the performance inapt, it would constitute a violation of (at least one) primary ψ-sustaining rule or rules. In the holothane gas case, the primary practice-relative rule that would be violated in the nearest worlds where the shot is inapt is "don't shoot when too tired", a rule with clear reproduction value in basketball. Now, *trivially* one can increase the performance's safety against that risk by adhering to *that* primary rule – viz.,

[16] For some representative discussions of safety in epistemology, see Pritchard (2005, 2007, 2012, 2016), Luper-Foy (1984), Sosa (1999), Rabinowitz (2011), Comesaña (2005), Ballantyne (2012), Engel (1992), Hetherington (2013), Madison (2011).

[17] For related discussion, see Pritchard (2016).

by *not* shooting when too tired. What one can't do, however, is easily increase the safety of the performance against that risk by adhering to *derivative* ϕ-sustaining rules. This is because adhering to derivative rules that have reproduction value in basketball (e.g., check for familiar signs that one is too tired) doesn't *easily* increase the safety of the performance against the risk to inaptness that is posed by halothane gas; monitoring in *those* ways, you'd never see it coming – at least, by following such rules. You could, by contrast, safeguard swimmingly against the holothane gas risk by toting around a gas monitor and checking it regularly while on the court. But attempting to comply to the primary rule 'don't shoot when too tired' by adhering to *this* derivative rule blatantly lacks reproduction value in basketball. (You'd surely be kicked off the team.)

Second, regarding the '*can't be easily* increased' locution. Why not just: '*can't be* increased?' Consider that one who gets very lucky could increase the safety of the shot against the halothane gas risk to its inaptness by adhering to a derivative rule with reproduction value, like 'check for familiar signs that one is too tired'. As it happens, one of the *less common* side effects of exposure to halothane gas is difficulty breathing, a side effect that overlaps with one of the tell-tale signs of lactic acid buildup as when one typically becomes tired. One *could*, as it were, 'get lucky' and experience this rarer symptom and correctly identify it as a marker of tiredness. In doing so, the safety of the performance against the halothane gas risk to inaptness *would* be increased through adherence to one or more derivative ϕ-sustaining rules. It just wouldn't be *easily* increased.

With the *de minimis* proviso now fully unpacked, it should be clear how it neatly separates the lactic acid and halothane gas cases (which for Sosa's view generated contradictory verdicts), by ruling the latter in as *de minimis*, and the former out. And this means that the wider account – *de minimis* normativism – intuitively rightly classifies the halothane gas and beaver cases as both cases where a performance is (unlike in the lactic acid case) fully apt, and despite the fact that halothane gas poses a practice-relative risk to the aptness of the shot, and the beaver does not.

On DMN, an apt chess move that could easily have been inapt because

the player (oblivious to her crashing glucose levels) could easily have been in improper shape is *not* fully apt. This is because on my view a fully apt performer can't non-negligently ignore practice-relative risks to the inaptness of a performance, and *this* is an instance of practice-relative risk. After all, were the performance inapt, there would be a violation of a primary practice-sustaining rule or rules – viz., "don't play with compromised mental acuity" a primary rule that generates derivative rules with reproduction value, such as "try to keep acuity sharp by monitoring glucose levels." Moreover, on my view, this practice-relative risk is not *de minimis* because monitoring for glucose levels in this case *would* increase the safety of the performance against inaptness.

It is important to note that the above rationale does *not* imply that, to make a chess move fully aptly, the chess player must be thinking throughout the game about glucose levels. Doing *that* would not have reproduction value in chess, even if it would increase the safety of the performance against inaptness from mental sluggishness. (Compare: monitoring not only for regular signs of tiredness but for signs one has ingested halothane gas would not have reproduction value in basketball even if it would increase the safety of a performance against inaptness from tiredness when there happens to be halothane gas about.)

Likewise, the proposal can diagnose the atmospheric pressure case – the other case ruled as an 'overlap' case on Sosa's view – in a principled way. Though, in this case, the risk (at least in basketball) – will be *de minimis*, unlike in the chess case. The thinking here is as follows: *even though* dips in atmospheric pressure can lead to one's shape being compromised, e.g., by bringing about joint stiffness, this is a *de minimis* risk for a basketball player; she can shoot fully aptly while ignoring it. This is because there is no derivative basketball sustaining rule (viz., no derivative rule *with reproduction value in basketball*) the adherence to which would increase the safety of a basketball shot against *that* risk. One could safeguard against it, no doubt – perhaps by being mindful of both the forecast and ways in which the arena could open up to the elements outside. But there is reproduction value in basketball to simply ignoring such risks, rather than to concern oneself with them.

6.5 *De minimis* normativism and the specific non-negligence question

We're now in a position to put everything together. Zooming back out, there were two key questions this chapter was attempting to answer, the General Non-Newligence Question and the Specific Non-Negligence Question.

Let's now *apply* the answer given to the former, general question – *de minimis normativism* (DMN) – to the specific case of interest to us.

In order to apply the account, we'll need to first define *practice-sustaining rules* and *practice-relative risks* in the special case of *trust*. By defining these two terms we can then spell out what *de minimis risks* are to the aptness of a performance within the wider practice in which trusting is embedded. Because practice-relative risks are *themselves* defined in terms of practice-sustaining rules, let's begin by focusing on nailing down *practice sustaining rules.*

An idea developed in Chapter Four – one that now bears relevance – is that trust is a performance embedded within the wider practice of *cooperative activity* – viz., of cooperation, and that we can helpfully think of trusting as a kind of 'move' in the wider practice of cooperation – much as believing is a kind of zetetic 'move', a move in the practice of inquiry.[18]

Taking, then, as a basic starting point 'trust/cooperation' as the operative 'performance/practice' relationship of interest, the kind of practice-sustaining *rule* (i.e., prescriptive principle) we need to provide a specification of is, specifically, that of a *cooperation-sustaining* rule.

With this in mind, it follows from DMN that:

> **Cooperation-sustaining rule**: A rule is *cooperation-sustaining* if and only if the value achieved by following the rule explains why agents engaged in the practice of

[18] See Kelp (2020b) for a detailed defence of the idea that beliefs are moves within the wider practice of inquiry.

cooperation continue following that rule.

Just as we can distinguish between primary and derivative practice-sustaining rules generally, we can also distinguish between primary and derivative cooperation-sustaining rules. An paradigmatic example of the former involves promise-keeping[19]: the *value* of following the rule 'keep promises' would explain why individuals engaged in any kind of cooperative activity continue to follow the rule.[20] One might of course *try* one's best to keep a promise but unluckily fail, for reasons outside of one's control. Such a person violates the primary norm 'Keep your promises' despite best intentions, but they will have nonetheless followed a *derivative* cooperation-sustaining rule: try to keep promises.[21]

With the notions of primary and derivative cooperation-sustaining rules in hand, we now can define *practice-relative risks* to the inaptness of performances that feature in a cooperative practice. The key idea here is as follows: a risk to the aptness of any performative 'move' one makes within the practice of cooperation is a practice-relative risk, that is, a risk relative to the practice of cooperation – if and only if, that performative move (within the practice of cooperation) is such that *were it to be performed inaptly*, it would constitute a violation of (at least one) primary cooperation-sustaining rule or rules. And when the relevant performance is that of *trust*, then, we get from DMN the following:

[19]Note that some rules, like promise keeping, will end up being practice-sustaining rules for multiple practices. For example, promise-keeping is plausibly a practice-sustaining rule for the practice of banking, also for the practice of romantic relationships. Of course, even though multiple practices can have in common certain practice-sustaining rules, the rules that hold together distinctive practices will include many rules that are distinctive to sustaining those particular practices. For further discussion on this point, see e.g., Turri (2017) and Carter (2020a).

[20]One explicit defence of this idea is found in Hume's ([1739] 2003) *Treatise*, in which Hume argues that norms of promise-keeping, even (contra Locke) in the absence of a government to sanction violations of it, would be followed on account of the value that is attained by following the rule. Relatedly, see Rawls (1955) and (2002) for rule-utilitarian defences of the value of promise-keeping.

[21]For a more recent discussion of primary and derivative norms, see Williamson (2021).

> **Cooperation-relative risks**: Risks to the inaptness of a token performance of trusting, T, are are cooperation-relative risks as opposed to a non-cooperative-relative risk if and only if, were T inapt, this would constitute a violation of (at least one) primary cooperation-sustaining rule or rules.

Recall now, from DMT, that the notion of a *de minimis risk* to the inaptness of a performance within a practice can be straightforwardly articulated in terms of the key notions of *practice-sustaining rule* and *practice-relative risk*. Since we've now defined both of these, we are in a position to see how the DMN framework can tell us when a risk to the aptness of trust is *de minimis* (and so can be non-negligently ignored).

> **De minimis proviso$_{\text{Trust}}$** A risk R to the inaptness of a token performance of trusting, T is *de minimis* (and so can be non-negligently ignored) if and only if the safety of one's trust against X can't be easily increased through the truster's adherence to one or more derivative cooperation-sustaining rules.

And we can now put the three key notions together for the final view, which answers the Specific Non-Negligence Question.

> **Specific non-negligence question:** What kind of risks to the inaptness of trust can the convictively apt truster *non-negligently* ignore?
>
> *Answer*: (DMN$_{\text{Trust}}$) A convictively apt truster can't non-negligently ignore *cooperative-relative risks* to the inaptness of trust, except when these risks count as *de minimis* with reference to cooperation-sustaining rules.

(DMN$_{\text{Trust}}$) is just an instance of the more general theory, (DMN), which tells us when risks to the aptness of *any* performance can be non-negligently ignored. (DMN$_{\text{Trust}}$) is what we get when we 'plug in' trust-relevant terms to (the components of) (DMN), so that it 'spits out' an answer to the specific non-negligence question.

Let's now put some pressure on (DMN_{Trust}) – what (DMN) has spat out – by putting it to the test in order to see if it can do what we need it to do. In the remainder of this we'll consider applications of (DMN_{Trust}) to two different categories of cases:

- first, we'll apply (DMN_{Trust}) to 'easy cases' (i.e., LOAN PAYMENT and MR. X), as a kind of 'proof of concept', to show that the view can do *at least* as well as Sosa's.
- next, we'll apply (DMN_{Trust}) to harder cases, viz., to the kinds of cases where an applicaiton of Sosa's answer to the specific non-negligence question would generate problematic results.

6.5.1 Proof of concept: easy cases

6.5.1.1 LOAN PAYMENT

A first 'test' of (DMN_{Trust}) will be to see whether it can get the right result in LOAN PAYMENT. Recall that in LOAN PAYMENT, armed warlords are, on the other side of the world, flipping a coin in secret to determine whether or not to unload their nuclear arsenal, and *were they to do so*, then *B* (i.e., the trustee in that case) would obviously not succeed in paying back the loan to *A*, as both would perish.

The idea here was that, when the coin flip by the warlords leads them to peace rather than war, *A*'s trust of *B* with the loan – when everything goes well – ought to be ruled as unimpeachable, *qua* trusting performance, *despite* the near-by risk to the trust's aptness posed by the warlords' coin-flip. Put another way: *this* kind of risk to the aptness of *A*'s trust is such that it ought to be able to be *non-negligently* ignored by a convictively apt truster, even if other risks to the aptness of *A*'s trusting *B* with the loan could not.

But *why*? But what would explain why a convictively apt truster could non-negligently ignore such a risk to the aptness of her trust? The explanation that is given by (DMN_{Trust}) would go like this: *A* can non-negligently disregard the risk to the aptness of *A*'s trust posed by the warlords' coin-flip *because the safety of A's trust against the risk posed by the warlords' coinflip can't be easily increased through A's adherence to one or more

derivative cooperation-sustaining rules.

Granted, the safety of *A*'s trust against this risk *could* be increased, easily so if *A* were powerful enough to hire spies to simply keep tabs on all warlords at all times. However, monitoring for *this* kind of risk doesn't have reproduction value in any kind of cooperative context whatsoever. (Compare: carrying around a device to detect for halothane gas risks does not have reproduction value in basketball, even if it *would* increase safety against halothane-gas-induced risks to the aptness of a shot, whenever such rare risk is present). Thus, even if one *could* safeguard against this risk effectively, and even if one could do so easily, one couldn't do so *through the adherence* to any derivative cooperation-sustaining rules.

Thus, *de minimis normativism* looks like it gets exactly the right result in LOAN PAYMENT, where we have a risk that, *prima facie*, could be ignored non-negligently by a convictively apt truster. This is the result Sosa's view got through an appeal to background conditions which (as we saw) made mischief in other kinds of cases (which we'll revisit shortly).

6.5.1.2 MR. X

Continuing a 'proof of concept' of how (DMN_{Trust}) can get the goods – let's now look at a risk to the aptness of trust that a convictively apt truster could clearly *not* simply ignore. Recall here the case of MR. X. from Chapter Five (noted again in full, for reference):

> MR. X: Mr. X, having read The Art of the Deal along with several books by Tony Robbins, fancies himself a charismatic dealmaker, overestimating his influence. Mr. X entrusts Mrs. Y with information I, in a situation within normal bounds of risk, effort and aptitude, and Mrs. Y does not betray Mr. X. Mr. X's (first-order) trust on this occasion may be apt – his successful trust manifests his competence to trust reliably enough in the shape/situation pertinent to first-order trust. However, suppose that while Mr. X in trusting Mrs. Y has trusted aptly, he very easily would have trusted inaptly. Although in entrusting Mrs. Y with

> information I, the risk to Mr. X is in fact not excessively high and gains of betrayal are within normal bounds, Mr. X. (with a distorted view of his charisma and influence, thanks to the Tony Robbins books) would easily have entrusted Mrs. Y with information I *outside such bounds* (e.g., had I been information that would have given Mrs. Y huge gains if divulged with little threat of her detection), in a situation where he would not have been a reliable enough truster.

In MR. X., Mr. X's trusting is (a) first-order apt; and (b) very easily, Mr. X would have trusted first-order inaptly. Both (a-b) line up not only with Mr. X, but also with *A*'s situation in LOAN PAYMENT. *However*, in the case of Mr. X. and *A* in LOAN PAYMENT, what explains why (b) holds is different. The fragility of the first-order aptness of *A*'s trust in LOAN PAYMENT was due to the nearness of the world-destroying possibility, which would have trivially wrecked the first-order aptness of *A*'s trust. But the fragility of the first-order aptness of Mr. X's trust is due, by contrast, to the the distorted view he has of his own (first-order) trusting competence, a distorted view which precludes him from accurately gauging the risks of trusting inaptly that are present.

As we have just seen, in LOAN PAYMENT, the safety of *A*'s trust against the risk posed by the warlords' coinflip *can't be easily increased* through *A*'s adherence to one or more derivative cooperation-sustaining rules. However – crucially – the safety of Mr. X's trust against the risk posed to the first-order aptness of *his* trust *easily could be*.

After all, while a rule like "try to rule out earth-destroying possibilities before trusting" lacks cooperative reproduction value (on the contrary, following this rule would have cooperative *disvalue*; those following it would contribute to bringing cooperative activity to a standstill), attending to a rule like "try to appraise your own influence on others accurately" *does* have reproductive value within the practice of cooperative activity, and by following this rule, the safety of Mr. X's trust against the risk posed to the first-order aptness of *his* trust easily could. After all, by following this rule, Mr. X would more easily have identified the limits of his own first-

order trusting competence, and that this situation featuring Mrs. Y lied *beyond* those limits.

6.5.2 Diagnosis of intermediate cases

The key take away from the application of *de minimis* normativism to LOAN PAYMENT and MR. X. is this: the view is able to get the goods in these cases at least as well as Sosa can via an appeal to background conditions. We turn now to some more complicated cases of trust – cases that have a structure similar to those more general kinds of performance cases (§6.3) that Sosa's view could not easily categorise.

Let's begin by registering that the existence of a well-functioning stock market (suppose *A*'s and *B*'s money is tied up in stocks) and currency system is going to qualify as a situational background condition on Sosa's view for almost any trusting performance that involves a large-scale (i.e., £100m), modern online loan, given that *A* cannot be properly situated to competently trust *B* with such a loan without both (which are closely related to each other) already being in place and functioning normally; their being in place is a precondition for *A* to entrust *B* with this kind of a large-scale loan, even if not for any loan. Thus, Sosa's view implies that *A* ought to be able to trust *B* unimpeachably with such a loan while non-negligently taking for granted the stability of the stock market and currency system, *even when* risks to both are modally close. This, again, is because their being in place is entailed by *A* being properly situated to make such a loan to *B* in the first place – which is enough to qualify these things as 'background conditions' on Sosa's proposal.

But here is where the analogy with the atmospheric pressure case discussed previously comes in to play.[22] Just consider that unexpected fluctuations in the stock market would very easily compromise one's situation for competently trusting, given that such fluctuations could easily ramp up the prospective gains of betrayal and/or effort for paying back the loan to

[22] Recall that *dips* in atmospheric pressure – the presence of which qualifies as a background condition on Sosa's view – are well known to lead to one's shape being compromised.

beyond normal bounds. *However*, when thinking about the stock market and its delicate relationship with the value of currency in *this* way, it looks as though a trustor (of large, modern £100m online loan) could non-negligently ignore nearby risks to the stock market *only if* they can also non-negligently ignore more mundane threats to being properly situated to trust – e.g., whether there is some strong incentive for the trustee to default, whether paying it back would require more skill on behalf of the trustee than usual, etc. But these are exactly the kinds of things Sosa's view rightly implies that a fully-apt performer *can't* be oblivious to. Accordingly, when when we ask, "can a fully apt truster of a modern £100m loan non-negligenty ignore the stabiity of the stock market?" The answer for Sosa seems unclear: his full proposal suggests 'yes' in one sense, 'no' in another – much in the same way his proposal gave this kind of mixed answer to the atmospheric pressure/basketball and glucose/chess cases from §6.3.

($\text{DMN}_{\text{Trust}}$) offers a more principled way to diagnose this case. In short, *rather than* to try to get an answer by asking whether the stock market's being in place and functioning normally is a *background condition vis-à-vis* the performance of trusting someone with a large modern loan, we should be asking a *different* question entirely: the question we should ask, on ($\text{DMN}_{\text{Trust}}$), is whether (in the case of this kind of performance) safety against the risk posed by the stock market can be easily increased through adherence to one or more derivative cooperation-sustaining rules.

Asking this question offers us a more nuanced way to diagnose the case, one that gets the right results. First, note that a bank loan lender's adherence to a rule like "monitor for possible threats to there existing a properly working stock market" is going to lack any reproduction value – and carry clear reproduction disvalue – in just about any kind of cooperative activity.[23] Adhering to this kind of rule would would demand of one actions

[23] One might attempt to press back here by pointing out that some philosophers have defended certainty norms on action and practical reasoning (see, e.g., Beddor 2020b; Stanley 2008). If such a line is plausible, then – as the thought would go – why not think also that the kind of certainty that would be implicated by requiring one adhering to extreme risk-aversion rules would be so implausible when the action is part of a

the undertaking of which would prevent successful cooperation. It would demand, for example, at *least* forestalling, e.g., entrusting one with a large loan until one has first taken a series of wide-reaching epistemic precautions – precautions against threats to the *existence* of the stock market – which include keeping tabs on, e.g., the potential for world-devastating events.

Just as assuming the world remains in place has reproduction value in the practice of *inquiry* – viz., the practice of particular inquiries would be brought to a standstill if we could did not make any kind of zetetic move (i.e., belief) until we had first followed a rule like "monitor for risks against the annihilation of the world", the same holds *mutatis mutandis* for monitoring for such risks within the practice of cooperating with others.

Crucially, though, this does *not* mean – within the ($\text{DMN}_{\text{Trust}}$) framework – that one thereby has a *carte blanche* to assume 'all is well' with the stock market before entrusting one with a large modern loan. This is because safety against the risk posed normal variations (including even crashes) in the stock market *can* be easily increased through adherence to one or more derivative cooperation-sustaining rules, including "speak with a finance expert before before making a £100m loan". Note that adhering to this rule – which has clear reproduction value – *does not* involve, checking for, e.g., asteroids that might collide with the earth ruining the sock market. And this is so even if the nearness of such an asteroid might threaten the aptness of one's trust.

($\text{DMN}_{\text{Trust}}$), accordingly, has the flexibility to deal with cases that Sosa's own proposal seems to lack a principled answer to.

---

cooperative activity? In response, it is important to note that in neither Beddor's nor Stanley's variations on a defence of a certainty norm for action do they conceive of certainty in such a way that action would be prohibited were one to not 'rule out' all kinds of far-off error possibilities. This is so even if there are some construals of certainty (e.g., Unger 1975) on which certainty does require that kind of epistemic work.

6.6 Concluding remarks

This chapter has further developed the account of convictively apt trust the key positive features of which were advanced in Chapter Five. The focus in this chapter has been on the other side of the coin – by clarifying not what convictively apt trust demands but what it *permits*. The guiding question has been the Specific Non-Neligence Question: What kind of risks inaptness of trust can the convictively apt truster *non-negligently* ignore?

This question was approached by first outlining an answer to a more basic question for any about risk and negligence, what I called the General Non-Negligence Question, which asked: What kind of risks to the inaptness of any performance, ϕ, with an aim, A_ϕ, can the fully apt ϕ-performer non-negligently ignore?

Because a fully developed answer to the General Non-Negligence Question already appears in Ernest Sosa's recent virtue epistemology, Sosa's own answer to the question was our natural starting point. Sosa's preferred strategy for answering the General Non-Negligence Question maintains that the fully apt performer can't non-negligently ignore risks to the inaptness of a performance, ϕ, unless those risks are due to the unsafety of ϕ-relevant background conditions. This strategy was critiqued and shown to face some problems. A better answer to the General Non-Negligence Question was then defended, according to which a fully-apt performer can't non-negligently ignore *practice-relative risks* to the inaptness of a given performance that occurs within that practice-type, except when these risks count as *de minimis* with reference to practice-sustaining rules. After unpacking the key components of this view and showing how they fit together to issue verdicts in cases, I showed how the proposal has advantages over Sosa's, and is not faced with the same problems that his answer faced.

My favoured answer to the General Non-Negligence Question was then used as a basis for answering the Specific Non-Negligence Question of primary interest to us. The answer to the Specific Non-Negligence Ques-

tion that was defended holds that a convictively apt truster can't non-negligently ignore *cooperative-relative risks* to the inaptness of trust, except when these risks count as *de minimis* with reference to cooperation-sustaining rules. This view was then put to work; as we've seen, the view not only gets us the right verdicts in straightforward cases, but it also does so in more difficult cases.

The answer given to the Specific Non-Negligence Question rounds out our complete account of deliberative trust, and what *apt* deliberative trust (i.e., convictively apt trust) does (and does *not*) demand of us. The telic theory of trust by this point is now much more fleshed out than when we first laid the core ideas on the table in Chapter Two. However, at the same time, there remain some rather fundamental issues in the philosophy of trust that we have not yet addressed. For one thing, in what sense, exactly, does trusting essentially involve subjecting oneself to risk of betrayal – as I've thus far simply taken for granted that it does – and why? Relatedly, many philosophers of trust have signed on to the view that trusting is (in some non-trivial sense) incompatible with *monitoring* the trustee for risks of betryaal; but if this is right, *why* is it the case? And can our telic theory be reconciled with it? It is these questions about the nature of trust, vulnerability, and monitoring that the next chapter will take up.

Chapter 7

Trust, Vulnerability, and Monitoring

7.1 Introduction

Here are two perennial questions in the philosophy of trust, both of which concern the relationship between trust and risk:

> *Vulnerability question*: In what sense does trusting essentially involve subjecting oneself to risk of betrayal?
>
> *Monitoring question*: In what sense is monitoring for risks of betrayal incompatible with trusting?

These questions have traditionally been pursued independently from one another.[1] It will be shown that they are much more closely connected than has been appreciated. The central objective will be to demonstrate how a performance-normative framework can be used to answer both the Vulnerability Question and the Monitoring Question in a principled way,

[1] For discussions of the relationship between trust and monitoring, see, e.g., Hieronymi (2008) and McMyler (2011) and Wanderer and Townsend (2013). For some representative discussions of trust's relationship to vulnerability, see e.g., Nickel and Vaesen (2012, 861–2). Cf., Pettit (1995, 208).

one that reveals a deep connection between not just the questions themselves, but also between the concepts of vulnerability, monitoring, and *de minimis* risk.

7.2 Trust and Vulnerability to Betrayal

The very idea that trusting constitutively involves subjecting oneself to the risk that one's trust is betrayed is platitudinous in the philosophy of trust.[2] But what counts as 'subjecting oneself' to risk of betrayal? Getting this right is important to understanding the nature of trust and what is distinctive about it.

One tempting starting point – widespread in the social and behavioural sciences[3] – is to begin with the role that trust plays in facilitating cooperation between parties with competing interests. And here a common view maintains that trust functions as a strategy to mitigate, without entirely eliminating, uncertainty.[4]

This way of thinking suggests a natural, even if imperfect[5], contrast between trusting someone X to ϕ (as entrusted) with *knowing* that X will do so – one that invites us to link trust-relevant vulnerability to betrayal with (some non-negligible degree of) *ignorance* about whether trustee will come through.[6]

---

[2] For various expressions of this idea, see, along with Hardin (1992), e.g., Baier (1986, 244), McLeod (2020, sec. 1), Nickel and Vaesen (2012, 861–2), Becker (1996, 45, 49), Dormandy (2020, 241–2), Kirton (forthcoming), O'Neil (2017, 70–72), and Hinchman (2017). Cf., Pettit (1995, 208).

[3] See, e.g., Krishnan et al. (2006), Waston and Moran (2005), Beck (1992).

[4] As Frederiksen (2014) puts it, 'Contemporary trust research regards trust as a way of dealing with uncertainty and risk. Predominantly, it suggests that trust reduces uncertainty by means of risk assessment and rational calculation'.

[5] The Cartesian position that knowledge entails subjective certainty no longer enjoys much popularity in mainstream epistemology. Though cf., Beddor (2020a) for discussion.

[6] The idea that knowledge obviates the need for trust is broadly analogous to the thought, due to Plato, that knowledge obviates the need for inquiry. In the *Meno*, Plato maintains that one 'cannot inquire about what he knows, because he knows it, and in that

Unfortunately, this kind of a starting point only gets us so far. It invites us to ask – what *kind* of ignorance suffices here? On the one hand, one might be ignorant that a trustee X will come through simply because there is some *actual* risk, R, (above some threshold) to X's coming through, and *regardless* of whether S perceives this to be the case. This is called *objective risk*[7]; it is objective because its status as a risk doesn't non-trivially depend on its being perceived as such. For example, an impending storm presents a risk that you will not be able to finish painting the house as entrusted, even if you are in denial – or misinformed about – the weather forecast. On the other hand, one might be ignorant that X will come through simply because one *perceives* there to be some risk (even if, objectively, there is not). *Perceived risk* is such that its status as a risk *does* (non-trivially) depend on its being perceived as such.[8] For example, the *perceived risk* that 5G towers increases the spread of Covid is such that its status as a risk is entirely dependent upon its (mistaken) perception as such.[9]

The distinction between objective and perceived risks maps naturally on to two different ways of answering the Vulnerability Question. According to a simple perceived-risk account of trust-relevant vulnerability to betrayal, trust essentially involves subjecting oneself to *perceived* risk of betrayal, though not to *objective* risk of betrayal.

I will argue that the simple perceived risk account is untenable. Trusting essentially involves subjecting yourself to at least some objective risk of betrayal. But this raises a question: what is the right way to characterise the kind of objective risk to which, by trusting, one essentially subjects herself? I will then consider and reject two answers: as (i) the product of the estimated objective probability of betrayal multiplied by the disvalue of

case is in no need of inquiry (Plato [385BC] 2011, sec. 80.e). The idea under consideration proceeds by a similar reasoning: 'one cannot trust another to do what he knows he will do, because he knows he will do it, an in that case there is no need for trust'. A contemporary variation on this idea is found in the sociology of George Simmel, who explicitly contrasts trusting with knowing (see, e.g., Wolff 1950).

[7] For discussion, see Hansson (2018).

[8] See, e.g., Sjoberg (2004) and Slovic (1987).

[9] For discussion of this perceived risk, and the extent of its uptake on social media, see Ahmed et al. (2020).

betrayal (i.e., as *risk expectation value*); and as (ii) the objective (frequentist) probability of betrayal alone, above some specified threshold. What the defects in these accounts reveal is the need for a *normative* objective account – framed in terms of *de minimis* risk – which is what I'll go on to propose and defend.

7.2.1 A simple perceived risk account

One initial – but ultimately misguided – line of argument against a simple perceived risk account of trust-relevant vulnerability to betrayal holds that there is a tension between (i) the presumed explicit, conscious awareness involved in risk perception; and (ii) the unconscious or tacit character of (at least some kinds of) trusting. Trust can certainly be unconscious or tacit.[10] And it *seems* plausible on first blush that risk perception is not. For example, it is a hallmark of the 'Risk Society' research programme (Beck 1992; Goddens 2013) that our perceptions of risk are often given expression through affect such as fear and anxiety.[11]

But the tension here is only apparent. The countenancing of implicit trust is problematic for the perceived risk account only if risk perception of the sort that is essential to trust can't *itself* be unconscious or tacit. But the empirical evidence – especially over the past several decades[12] – on unconscious bias and risk perception has established, uncontroversially, that *even if* some risk perception is accompanied with conscious awareness (i.e., some combination of occurrent beliefs plus affect) a significant extent of our risk perception takes place below the surface of conscious awareness. (Compare: our cognitive biases are often *unconscious* biases,

[10] For some empirical discussion on the ubiquity of tacit trust, see, e.g., Lagerspetz (1998), Burns (2006) and Guo et al. (2014).

[11] In this line of thinking, Bauman (2013) describes our modern high-tech predicament, characterised by new technologies and dangers, as pervaded by a 'derivative fear' namely 'the sentiment of being susceptible to danger: a feeling of insecurity and vulnerability'.

[12] See, e.g., Sjoberg (2000) and Slovic (1988).

and at least some of these biases consist in perceptions of risk[13]).[14]

What this means is just that *if* the perceived risk account of trust relevant vulnerability to betrayal is problematic, it isn't going to be so because of any 'mismatch' between the implicit character of (some) trust and the alleged conscious character of risk perception; just as trust itself can be deliberative or implicit, so can our perceptions of risks to its being betrayed. There is, however, a much more serious problem that faces the perceived risk account of trust-relevant vulnerability to betrayal. Consider the following case:

> SUNRISE: Having read some fringe QAnon conspiracy theories on a Reddit subthread, you come to think your friend is among a select group of people who decides how and when the sun rises, by manipulating the earth's orbit and rotation. Afraid the group might trigger an event that would shroud your hemisphere in permanent darkness (something you believe you friend has final control over), you say "Can I trust you not to prevent the sun from rising?" Your friend (though finding this request strange) says they can surely oblige, simply because they knew that betrayal here would be impossible.

Question: Did you really *trust* your friend to not prevent the sun from rising, or did you merely *think* you did? There are two good reasons to think you merely *thought* you did. The first appeals to a very weak attribution principle according to which S trusts X with ϕ only if either (i) X is in a position to have ϕ-ing attributed to her; or X is in a position to have not-ϕ-ing attributed to her. This principle is implied the platitude that trustees incur any commitments at all *vis-à-vis* what they are entrusted to do, commitments they may uphold or not depending on what the trustee does. Granted, one could reject this attribution principle, but only on pain of then losing a grip on what distinguishes trustees from those (e.g.,

[13] One classic example here is 'shooter bias' (e.g., Unkelbach et al. 2008).

[14] For some representative discussions of unconscious or implicit bias, which include some perceptions of risk, see, e.g., Saul (2013) Holroyd et al. (2017).

mere sympathisers with the trustor, bystanders, etc.) who incur no commitments to the trustor, *vis-à-vis* ϕ-ing, one way or another. But, crucially, from this attribution principle it follows straightforwardly that you didn't really trust your friend in SUNRISE, even if you thought you did.

A second reason for doubting that genuine trust is present when you think you're trusting but subjecting yourself to *merely* perceived risk (i.e., as is the case in SUNRISE) is closely related to the first. Just consider the tight relationship between trust and reactive attitudes such as gratitude and blame. A common view in the philosophy of trust is that gratitude would be an appropriate or 'fitting' reactive attitude to a trustee's coming through, as blame would be to betrayal.[15] But, as this line of thought goes, gratitude would be *clearly* misplaced, if directed by your friend to *you*, when the sun then goes on to rise the next day as expected.

These points suggest that there is an intractable kind of problem with the simple perceived risk account. If trusting essentially involves subjecting oneself to *merely* perceived risk of betrayal, then there will be pressure to rule-in cases like SUNRISE as cases of genuine trust. But we have good independent grounds for thinking that SUNRISE is not a case of genuine trust; at least, this is the case given the very weak assumptions that trust involves (i) the incurring of normative commitments by the trustee *vis-à-vis* what she is entrusted with; and (ii) the presumed connection with the reactive attitudes it is taken to have.

The proponent of a simple perceived risk account of trust-relevant vulnerability to betrayal might press back by digging in the heels and defending an *immunity from error* thesis along the following lines: if one believes one is trusting X to ϕ, then one is trusting X to ϕ. If this immunity from error thesis is correct, then I am wrong to claim that (given the lack of any objective risk) I merely *think* I am trusting you not to fiddle with the earth's orbit. Whether I'm trusting you with something is, as this line of thought goes, not something I could be mistaken *about*: it is guaranteed that I am trusting you with X if I take myself to be trusting you with X. Trusting is in this respect akin to the kinds of mental states (i.e., perhaps

[15] See, e.g., O'Neil (2012), Domenicucci and Holton (2017), and D'Cruz (2015).

like 'being in pain' or 'being confused') that have the property of being such that if one believes one is in that state, then one is in that state.[16]

But such an immunity from error thesis is false in the case of trust for two main reasons, and the perceived risk account of trust-relevant vulnerability to betrayal therefore can't press back against the objections raised by relying on it. The first reason has to to do with the the fact that we can and often are mistaken about reliance facts. Reliance is necessary even though not sufficient for trust.[17] However, I am not infallible about whether I am relying on you for X; for one thing, I might forget I am relying on you to repay a debt. Or, I might forget that you've already repaid a debt and so believe mistakenly that I am *still* relying on you to repay it. But since I can mistake reliance facts, as the thought goes, my thoughts about whether I trust can't be self-guaranteeing.

Secondly, and following here Santiago Echeverri (2020), one standard way to test whether a belief that you are in some state guarantees its own truth is to ask whether it would be either incoherent or irrational for one to *question* whether one is in that state.[18] But for any case where we have trusted someone X to ϕ, we can coherently question whether we have done so. This is thus another reason why it is a mistake to attempt to revive the perceived risk account by latching on to the idea that it's impossible to think you are trusting someone with something when you're not. All this points to is the beginnings of an answer to the Vulnerability Question. That question asks: In what sense does trusting essentially involve subjecting oneself to risk of betrayal? Our working answer is now: not *merely* to perceived risks of betrayal. Let's continue to refine this answer.

[16] For discussion, see, e.g., Shoemaker (1995) and Burge (1998).

[17] For discussion, see Carter and Simion (2020).

[18] For example, it might be incoherent or irrational to ask whether you are a thinking thing, or (to use an example from Kaplan (1979) involving indexicals to ask whether it is true that "I am here now"?

7.2.2 Towards an objective risk account

Let's explore now the idea that necessary to trusting is subjecting yourself to at least some non-negligible risk to betrayal whose status *as* a risk to betrayal doesn't (non-trivially) depend on its being perceived as such. A natural first-pass at refining this idea maintains the following: trust essentially involves subjecting oneself to risk of betrayal *beyond some objective risk 'threshold'*.

As is common in risk analysis[19], an (objective) risk threshold is set as (above or below) some specified risk *expectation value*, which is calculated as the product of (i) objective (or frequentist) probability of the risk event obtaining; and (ii) its severity (i.e., degree of harm of the risk event's obtaining). For example, the risk expectation value of a low-probability risk with significant severity were it to obtain might be very similar to the risk expectation value of a much higher-probability but less severe risk.

A qualification here needs some care. One might ask "Since we must inevitably use our own evidence to work out what the risk expectation value is for a given risk, and different people have different evidence that they will be relying on to make such an assessment, then doesn't the notion of 'objective' risk – understood as above a risk expectation value threshold – just collapse into a perceived risk account?" The answer, importantly, is 'no'. When we try to determine a given risk expectation value, we inevitably make a subjective *assessment* of the objective probability of the risk event obtaining as well as of the objective disvalue. But – and this is a crucial point of difference between the notions of objective risk and perceived risk – on the latter account, what the risk facts are do not depend on our *estimates*. In characterising risk expectation value, we are attempting to characterise something that is *mind-independent*. Perceived risks by contrast depend (non-trivially) on their being perceived as such.

Bearing these qualifications in mind, appealing to objective risk expectation value (the product of the objective probability of the risk event obtaining multiplied by its severity) would be an obvious way by which one

[19] See, e.g., Hansson (2018, 2004) for discussion.

might try to assess *risk of betrayal*, simpliciter. However, appealing to objective risk expectation value it is ultimately not a promising way to think about *trust-relevant* vulnerability to betrayal, viz., as vulnerability expressed in terms of risk expectation value threshold. The problem is not the objective frequentist interpretaion of probability at issue[20], but rather, what happens when we adjust (significantly) the expected disvalue. To see the problem, consider the following simple pair of cases:

> BABYSITTER: *A* trusts *B* to responsibly babysit their only child, *C*, for the weekend; assume the objective probability of betrayal is .001 and would generate 100,000 units of disvalue.

> PENCIL: *A* trusts *B* to use *A*'s pencil and return it; assume the objective probability of betrayal is .1 and betrayal would generate .001 unit of disvalue.

Both BABYSITTER and PENCIL are paradigmatic cases of trust, though the the risk expectation value products are dramatically different: in BABYSITTER, .001 x 100,000 = a risk expectation value of 100. In PENCIL, .1 x .0001 = a risk expectation value of .00001, which will – and here is the worry – end up being lower than any kind of plausible threshold we might appeal to in order to distinguish cases of genuine trust from cases where there is effectively no objective risk of betrayal. What's more, cases like PENCIL become even more difficult for the kind of proposal under consideration when we lower even further the disvalue of betrayal (e.g., to .00000001 disvalue).[21] What cases like

[20] It is worth noting that risk expectation value, while naturally allied to a probabilistic gloss, isn't necessarily tied to one. For a modal approach to risk expectation value, see Pritchard (2015).

[21] Another kind of case that serves to capture this kind of problem will simply shift the value of what is entrusted to near zero, where the shift takes place after trust is placed. For example, I may loan you my gold pen (my only valuable possession) so you can impress a client. I trust that you'll return it. In the meantime, a goldmine might be discovered that saturates the market and sends its value to ~£0. This fact doesn't undermine my having trusted, and continuing to trust, you to return the gold pen. Thanks to [OMITTED] for discussion of this kind of case.

PENCIL seem to suggest, then, is that if we want to characterise the kind of risk that trust essentially involves subjecting oneself to in terms of objective rather than merely perceived risk, we might do better to simply control for the severity of betrayal and then characterise the relevant risk threshold solely in terms of the objective probability of betrayal. Then, presumably, PENCIL will be above the relevant risk threshold, given that the probability is .1 (10%).

Continuing with this idea, suppose we were to set the threshold as .05 (5%). This move will get PENCIL right; and since the probability that the sun won't rise is vanishingly low, there is no pressure to rule in SUNRISE. However, the cost with this kind of a move is that we can then no longer deal with cases like BABYSITTER. After all, when stakes are high (i.e., when the disvalue of betrayal is suitably high), it seems we might, and very often do, trust one even when the objective risk is very low – i.e., 1 in 1,000 (.10%) as in the case of BABYSITTER. Cases like BABYSITTER are not aberrations: many cases of trust (e.g., with loved ones' lives and welfare) have a structure whereby something of high value is entrusted to someone very reliable, precisely *because* they are very reliable, and are accordingly very unlikely to betray the trust.

One might try to deal with the above by simply setting the objective probability even lower. On such a view, trust essentially involves subjecting oneself to betrayal in the sense that the objective probability of betrayal must be, e.g., at least 0.000001 (0.0001%, i.e., 1 in a million). Since there is probably at *least* a 1 in a million chance the the hero in BABYSITTER brings about a disaster, this tweak seems to put the threshold on the right side of BABYSITTER. But the cost of setting the threshold *this* low is that you then invite an entirely different problem, which brings us back to cases in the vicinity of SUNRISE.[22]

[22] After all, once the threshold is set this low, then it will be difficult to explain why we should rule out (as we should) cases as being cases of genuine trust where the weak attribution principle isn't satisfied, and where reactive attitudes toward the would-be trustee would be misplaced.

7.2.3 A performance-normative account

Here is where we've got to. Trust essentially involves subjecting oneself to risk of betrayal in a sense that: (i) cannot be captured exclusively with reference to perceived risk of betrayal, because trust requires at least some objective risk betrayal; however, (ii) the threshold of objective risk above which one by trusting must essentially subject herself isn't something we can plausibly capture satisfactorily in terms of either (a) objective risk expectation value; or (b) the objective probability of betrayal alone; (iii) neither (a) nor (b) could handle all three of our examples cases together; and so (iv) *whatever* level of objective risk beyond which by trusting we thereby subject ourselves accordingly needs to be characterised in some other way.

The way forward, I want to suggest, is to pursue the idea that the relevant threshold of objective risk to which by trusting one essentially subjects herself is fixed by neither (i) risk expectation value nor by (ii) simple objective probability of betrayal, but rather, it is fixed (iii) *normatively* – viz., with reference to the (objective) normative concept of *de minimis* risk, viz., risks that can be *non-negligently* ignored by a truster.

The working idea that I will unpack, refine, and then put to work in order to handle our problem cases is the following:

> *De Minimis* **Account (DMA)**: Trust essentially involves subjecting oneself to a risk of betrayal that is not merely *de minimis* within the relevant cooperative practice.

The starting point for unpacking DMA is that trusting as well as distrusting are both performative 'moves' within within the wider practice of cooperation – in a way that is roughly analogous to how belief and withholding are performative moves in practice of inquiry.[23]

'*De minimis*' is a normative term; a risk to the success of any performance (i.e., aimed attempt) is *de minimis* iff it can be *non-negligently* ignored in

[23] See Kelp (2020b) for a recent defence of this kind of picture of inquiry.

the course of making the relevant attempt.[24] *De minimis* risks are always *de minimis*, and thus have this normative standing, *relative to a practice*, where a given practice (i.e., a way of doing things) is held together by rules, either explicit or implicit.

What distinguishes the rules that *sustain* a given practice, as opposed to those rules that are merely incidental to it? A plausible general characterisation here, following John Turri (2017), is axiological: rules 'hold together' a practice whenever the *value* of following *those* rules explains why people engaged in that particular practice continue to follow them – viz., rules are practice-sustaining when they have 'reproduction value' within the practice.[25]

Accordingly, the initial idea that *de miminis* risks are practice-relative is tantamount to the idea that *de minimis* risks – those that can be non-negligently ignored – will *have that normative status they have in connection with negligence always relative to a system of rules*, the rules that hold the practice together. And this, then, raises a question: in virtue of *what* would a given risk, e.g., to the success of *S*'s ϕ-ing, within a practice ψ, attain (when it does so attain) the normative status of being such that it could be *non-negligently* ignored by *S* with reference to the system of rules that constitutes ψ?

Here is a promising initial answer: we surely *can't* non-negligently ignore risks to the success of a performance within a practice *if there are rules with reproduction value within the practice, the following of which would easily mitigate against the risk*. (The archer, for example, can't non-negligently ignore whether the wind is blowing, even if the underwater diver can.)

But then – and this is the other side of the coin – we presumably *can* non-negligently ignore risks to a performance's success if the safety against that risk *can't* be easily increased through the truster's adherence to any rule whatsoever (e.g., "Monitor for this", "Check for that, etc.") that has reproduction value within the relevant practice. For example, the basket-

[24] See Sandin (2005) and Peterson (2002) for discussion.

[25] See also Carter (2020a).

ball player *can* plausibly non-negligently ignore risks of earthquakes prior to taking a shot, even though an earthquake would spoil that shot, given that monitoring for earthquakes lacks any reproduction value whatsoever in basketball (it is a rule the following of which would be a disvaluable distraction in the practice of basketball).[26] There are no rules with basketball reproduction value that a player *could* adhere to in order to easily safeguard against *that* risk. Thus, taking the non-obtaining of an earthquake scenario for granted is non-negligent during the making of performative moves within that particular practice, no matter how nearby the earthquake risk is modally: (after all, the *neglecting* of that possibility flouts no rules that are valuable to follow within the practice.)

Putting the key pieces together, we are now in a position to see how DMA works as a substantive answer to the Vulnerability Question we began with. DMA purports to answer that question by telling us in what sense trusting *essentially* involves subjecting oneself to risk of betrayal. And the answer to this question offered by DMA makes reference to the concept of risks that are *de minimis*, viz., risks that can be non-negligently ignored by a truster. We have now got a working view of what such risks are and how to identify them: a risk of one's trusting being betrayed (alternatively: a risk to the success of one's trust) is *de minimis* and thus can be non-negligently ignored by a truster iff the safety of one's trust against that risk *can't be* increased through the truster's adherence to one or more rules that have reproduction value within the cooperative practice within which one is placing one's trust.

Let's now 'plug' this substantive characterisation of *de minimis* risk of betrayal back in to DMA in order to put the core idea of the view in full view. Since DMA maintains that trust essentially involves subjecting oneself to risk of betrayal that is *not merely de minimis* within the relevant cooperative practice, DMA tells us that trust essentially involves subjecting oneself to at least some risk or risks of betrayal that *aren't* merely *de minimis*

[26] For a different explanation of *why* far-off risks such as the earthquake risk could be non-negligently ignored, see Sosa (2017, 191) and for more recent developments, Sosa (2020).

– viz., that *aren't* such that one can't increase the safety against them by adhering to rules that have cooperative reproduction value within the relevant practice. That is the full way to spell out DMA – viz., that trusting involves rendering yourself vulnerable *beyond* mere *de minimis* risk of betrayal.

With the key components of the account now on the table, let's see what it can do, by checking whether it can – as advertised – fare better than the other accounts considered. Let's take, first, SUNRISE.

SUNRISE was not a case of *bona fide* trust. Our normative view DMA straightforwardly accommodates this. DMA says that trust essentially involves subjecting oneself to risk of betrayal that is *not merely de minimis* within the relevant cooperative practice. And the risk subjected to here is *de minimis* (i.e., it *can* be non-negligently ignored) because you can't increase the safety against *that* risk of betrayal by following any cooperation sustaining rule whatsoever. (Indeed, given the details of SUNRISE, this turns out to be *trivially* so; the likelihood of *that* risk event materialising remains the same (near zero) no matter *what* you do. Thus, by DNA, SUNRISE is not a case of genuine trust. So far, so good.

What about the BABYSITTER case? BABYSITTER *is* plausibly a case of trust, and a paradigmatic one, *despite* the very low objective probability of betrayal. If DMA is going to secure this result, then it had better be the case that you *could* at least in principle increase the (already robust) safety against risk to betrayal by adhering to rules with cooperative reproduction value. And indeed you can, and you can do so in relatively mundane ways: consider that such rules include vetting the babysitter *ex ante* (i.e., checking up on references for reliability), making babysitting itself easier (i.e., laying out emergency phone numbers, a list of medications, etc.): rules that encourage these have cooperative reproduction value. (Compare, by contrast: aiming to increase safety against the low risk present by *surveilling* the babysitter is non-cooperative; it is a form of monitoring we will discuss in the next section). Accordingly, then, the risk event that would consist in the babysitter failing to keep the child safe is not *de minimis* risk, even if it is very low due to the babysitter's impressive reliability

and the straightforwardness of the task. Thus, DMA again gets the right result.

Let's turn now to PENCIL. This was also a case of trust, despite the very low albeit non-negligible *disvalue* of betrayal, and which generated a problem for an answer to the Vulnerability Question framed in terms of objective risk expectation value. For DMA to get the right result in this case, it had better bet that you *could* mitigate against the 'pencil theft' risk by the adherence to rules with cooperative reproduction value. And so you could. Writing you name on your pencil, for example, would violate no cooperation-sustaining rules; doing so facilitates rather than hinders cooperation between trustor and trustee. (More generally: the rule in play here would be to make items you loan out identifiable, which is a rule that has reproduction value for cooperation through loaning and borrowing). Accordingly, DMA is going to countenance PENCIL, rightly, as a case of trust, even though the disvalue of betrayal is exceedingly low (problematically so for the risk expectation value account to plausibly 'rule in' this case as a case of bona-fide trust).

The scoreboard of cases, then, is as follows:

Answer to Vulnerability Question	SUNRISE	BABYSIT.	PENCIL
> some threshold (T) of perceived risk	x	✓	✓
> some (T) of expected disvalue	✓	✓	x
> some (T) of (frequentist) prob. of betrayal	✓	x	✓
> (normative) *de minimis* risk	✓	✓	✓

7.2.4 Objections and replies

So far, it is looking liker DMA outperforms the competition as an answer to the Vulnerability Question, at least in so far as the view gets on the wrong side of none of the three cases that posed a problem for at least one of the other views considered. This is a promising mark in favour of DMA. Let's now see how the proposal holds up against some anticipated objections.

Objection 1

Even if we grant that the DMA gets the SUNRISE case right, there are nonetheless 'Frankfurt-style' cases (also with effectively zero risk of betrayal) that pose a problem to *any* view that takes some (non-zero) objective risk of betrayal to be necessary for trust.

Notice that it is a feature of SUNRISE that, given the effectively zero objective chance of betrayal, which would have involved moving the earth's orbit, it was completely *out of the control* of the trustee whether they betray or not, such that betrayal (or not) isn't something that could be attributed to them. Even so, it seems like we can imagine cases where the following *both* hold: (i) there is zero objective chance of betrayal; but (ii) where it is *not* out one's control whether they betray or not such that we could attribute *at least trust fulfilment* to the trustee, thus satisfying the weak attribution principle. For example, consider FRANKFURT-BABYSITTER:

> FRANKFURT-BABYSITTER: Suppose this case is just like BABYSITTER, except that it is a Frankfurt case[27], in that *if* the babysitter were to do something that would in any way imperil the baby, a benevolent demon would rush in and course-correct, preventing any danger to befall the baby. Because the babysitter (through her own goodwill and reliability) does everything right, the benevolent demon never has to intervene.

If we are going to retain the idea that trust essentially involves 'some objective risk' of betrayal – a conclusion from the critique of the perceived risk account – it looks like we're going to get the wrong result, i.e., that this isn't a case of trust. But, as the worry goes, it *is* trust despite there being no objective risk whatsoever that the babysitter will *not* come through. So long as she behaves in such a way that the Frankfurtian demon needn't intervene, all is good.

[27] See, e.g., Frankfurt (1969).

Reply

It is important to distinguish (i) risks to *successful reliance* and (ii) risks to *successful trust*. If you rely on someone to ϕ, your reliance is successful iff they ϕ, no matter how.

Trust asymmetrically entails reliance. When you trust someone to ϕ, you trust them to ϕ *as entrusted*, where 'as entrusted' might include such things as: with goodwill toward the trustor (e.g., Baier 1986; Jones 1996), by encapsulating the interests of the trustor (e.g., Hardin 2002), by believing they have a commitment to the truster to ϕ (e.g., Hawley 2014), etc.

For my purposes, I am happy to remain neutral on which of these ways of unpacking 'as enstrusted' best distinguishes trust from mere reliance. What is relevant at present is just that the success conditions for reliance and trust differ, in that trusting someone to ϕ is successful iff they ϕ *as entrusted* (however this is to be spelled out), and not *merely* iff they ϕ.

This difference in success conditions is important in defusing the above objection. This is because *there can be risks to successful trust that are not also risks to successful reliance.* And indeed, that is exactly what is going on in FRANKFURT-BABYSITTER. It is true that there is zero objective risk to successful reliance; the benevolent demon waiting in the wings is seeing to that. But it is not *thereby* also true that there is zero objective risk to successful *trust*. The trustee's taking care of things as entrusted – however we fill this out – is plausibly going to require some exercise of autonomous agency, some *way* of taking care of things, attributable to the trustee – a point that lines up with the observation that reactive attitudes like gratitude are appropriate to fulfilled trust as well as to betrayal. Actions and mental states caused by the demon's intervention are not autonomous[28]; when compelled to act by the demon, the agent is not free to govern herself one way or the other, with respect to what she has been entrusted to do. She *cannot* fulfil trust (even if she can play a causal role in bringing about what she was relied on to do) or betray it.

[28] There are different ways to explain why. For two prominent options, see, e.g., Mele (2001) and Fischer and Ravizza (2000).

Thus, there *is* some non-zero objective risk of betrayal (i.e., a risk to successful *trust*) in FRANKFURT-BABYSITTER, even though there is no objective risk to successful reliance. And importantly, the objective risk to betrayal is (as is pertinent to DMA) not *merely de minimis*: This is because, just as there are pro-cooperative rules you could adhere to to increase safety against risk of betrayal in the original (non-Frankfurt) BABYSITTER case, so likewise, the same applies here – viz., such rules include, e.g., proper vetting, and (post-vetting) facilitating cooperative attitudes of the trustee through, being cooperative as a truster – e.g., by making duties clear. DMA therefore is able to handle not only BABYSITTER but also FRANKFURT-BABYSITTER.

Objection 2

Let's consider now a further objection to DMA, one that serves well as an entry point into the discussion in the next section on the relationship between trusting and monitoring.

Consider that on the proposal advanced, trust essentially involves subjecting yourself *beyond* mere *de miminimis* risk of betrayal – that is, it essentially involves subjecting yourself to at least some risks of betrayal that can't be non-negligently ignored – viz., such that the safety against them *couldn't* easily be increased through adherence to rules with cooperative reproduction value. But a corollary of this idea is that that *all* cases of trust are ones where you could at least potentially increase safety against betrayal by following rules with cooperative reproductive value.

But – and here is the worry – isn't *this* commitment of the view somehow in tension with the platitudinous idea that trusting is incompatible with *monitoring?* After all, there will often be no more effective way to increase safety against betrayal than to blatantly monitor the trustee's every move. Rather than to, e.g., mitigate against betrayal by carefully vetting the babysitter's references and then leaving helpful reminder notes, why not simply watch the entire time with surveillance cameras, or – better yet – hire a full surveillance team to oversee the babysitter's every move?

In short, the objection to the proposal can be put like this: the answer

given to the Vulnerability Question – viz., that trust essentially involves subjecting yourself beyond mere *de miminimis* risk of betrayal – rests on the underlying idea that when one trusts, there *are* certain things that one can do to increase safety against betrayal. But, increasing safety against betrayal is (at least in cases of monitoring, which is one very obvious way to increase safety against betrayal) incompatible with genuinely trusting. Thus, it seems that the answer given to the Vulnerability Question cannot be satisfactory: it relies on a claim that is itself in tension with the datum that monitoring kills trust.

Reply

This is a straightforward objection, and it has an equally straightforward answer. *Monitoring*, even though it increases – perhaps better than anything else! – safety against betrayal, is fundamentally *non-cooperative*. For one thing, that norms of cooperation generally prohibit monitoring or surveilling a trustee is supported by our practices of sanctioning; we tend to sanction those who purport to trust and then monitor.[29] Additionally, monitoring contributes to the erosion of conditions for cooperation; this is due to the social function of monitoring as *signalling* a lack of confidence in a pre-established commitment.[30]

Importantly, the view advanced here does not maintain that by trusting you subject yourself to risks of betrayal such that you could (while continuing to trust) *in any way* increase the safety against their obtaining. *That* would indeed be an unacceptable result. It implies rather that trusting essentially involves subjecting yourself to risks of betrayal that are not merely *de minimis*, which just means that by trusting you subject yourself to at least some risks of betrayal such that you could (in principle, and regardless of whether you do) increase the safety against their obtaining without violating any cooperation-sustaining rules.

And indeed we increase safety against betrayal without violating any such

[29] See, e.g., Kramer (1999).

[30] For some studies reporting these effects in cases where computers are used to monitor employees, see Ariss (2002).

rules like this *all the time* (and *without monitoring*): by deliberating about whom to trust, assessing their reliability, assessing facts pertinent to the likelihood of betrayal, including the extent, present in a given trust context, of the (a) *gains to the trustee* that would come from betrayal; (b) the *effort*; and (c) the *aptitude* required by the trustee to *avoid* betrayal. Through a competent assessment of these factors one can cooperatively increase safety against risk of betrayal. (Likewise, one can cooperatively increase safety against risk betrayal by facilitating the ease by which the trustee can take care of what is entrusted, e.g., by leaving a map, leaving detailed instructions, etc. None of these things involves trust-incompatible monitoring.)

In sum, then, the idea that trusting essentially involves subjecting yourself beyond mere *de minimis* risk of betrayal does not stand in tension with the platitude that trusting is incompatible with monitoring the trustee.

7.3 Trust and Monitoring

The final objection in the previous section prompts a further question - in what sense, then, is monitoring for risks of betrayal incompatible with trusting? This is the Monitoring Question. The Monitoring Question takes at face value that monitoring *is* incompatible with trusting. It invites us to explain *how so*.

Just as a good answer to the Vulnerability Question required some sense of what the threshold is beyond which by trusting one subjects oneself to risk, a good answer to the Monitoring Question requires some sense of what the threshold is beyond which by monitoring one is no longer trusting.

Here is the answer to the Monitoring Question I will now defend:

> (MON): One's monitoring is incompatible with trusting to the extent that, through monitoring, one intentionally aims (through the taking of some means) at *invulnerability* to risks of betrayal that, by trusting, one essentially subjects

oneself to.

Since by trusting one essentially subjects oneself beyond mere *de minimis* risk of betrayal, MON implies that monitoring is incompatible with trusting to the extent that it involves taking means by which one aims to render oneself *invulnerable* to all but *de minimis* risks of betrayal.

Two initial clarifications here are needed. First, the proposal does not say that one *actually* has to render herself invulnerable to all but *de minimis* risks of betrayal. This is important, because the monitoring needn't actually succeed in that aim to be incompatible with trust. Consider, for example, the following case:

> CRYSTAL BALL: *A* hires *B* to babysit *A*'s child. Highly superstitious, *A* believes, falsely, that *A* has a working crystal ball. After dropping *A*'s children off with B, A hurries home to the crystal ball in an attempt to surveil *B*'s every move. The crystal ball shows ambiguous, smoky images, which *A* mistakenly thinks provide information about *B*'s movements. *B* watches and attempts to interpret these movements, much as a less superstitious person might peer into the grainy images on a nannycam with poor resolution.

Intuitively, *A* is no longer trusting *B* when using the crystal ball – anymore than one surveilling via a poor-resolution nannycam would be doing so – and *even though A* is not succeeding in making herself invulnerable to any risk of betrayal whatsoever. An account of the incompatiblity of monitoring with trusting that required actually succeeding in eliminating, even to some degree, such vulnerability would fail to get the right result in CRYSTAL BALL.

A second clarification: why 'the taking of means by which one aims'? Why not *simply* 'aims'? The reason is that monitoring – as opposed to something less, i.e., merely intending but failing to monitor – requires an attempt to *attain* an aim (i.e., vulnerability elimination) *in some way*, viz., through some means by which one *through taking those means* (in this case,

via the means of surveilling the trustee) monitors; in this respect, monitoring that is incompatible with trusting is not 'idle aiming' (i.e., mere intending to monitor) any more than trusting is idle aiming (mere intending to trust).

These clarifications made, we can see now that MON is able to secure the following pleasing result: it can explain why surveilling the babysitter with a nannycam (or, for that matter, attempting to do so via a crystal ball) is incompatible with trusting but vetting the babysitter for reliability (prior to hiring) and leaving notes and reminders after is not, even though the latter kinds of things also minimise risk of betrayal. The explanation given is that, in the former examples, one intentionally aims (through the taking of some means) at invulnerability to risks of betrayal that, by trusting, one essentially subjects oneself to (i.e., to risks that aren't merely *de minimis* in the relevant contexts), whereas this is not so in the latter cases.

The fact that MON is able to generate different verdicts in the former and latter kinds of cases constitutes a key advantage over a more standard line of thought about trusting and monitoring in the literature (e.g., Elster 2015) according to which trusting essentially involves simply refraining 'from taking precautions against an interaction partner' (2015, 344). Such a proposal frames the relationship between trust and monitoring in a problematically coarse-grained way; it would get the former cases right, but not the latter unless it could provide us (as MON does) a principled reason for a difference in treatment.

An additional advantage of MON is that the very idea that 'aiming' at eliminating vulnerability is something that is incompatible with trusting fits snugly with a much more basic idea about trusting *qua* performance, or aimed attempt. Within the theory of performance normativity, performance types may be distinguished from each other by the constitutive aims internal to those performance types. In slogan form: change the aim, and you've changed the performance type.[31] Take for example the performance of 'making a guess' versus 'making a judgement'; these are different

[31] For discussion of how performances are individuated by their aims, see Sosa (2010a), and the essays in (ed.) Vargas (2016).

truth-directed performances; but why? A typical answer[32] adverts to a difference in the level of risk one aims at taking on as a price for a chance at truth in each case. *What it is* to make a guess is to aim at truth via affirmation in a way that tolerates at least an unusually high level of risk; were one to aim at truth via affirming *without* aiming at tolerating whatever level of risk is distinctive of guessing, then one is longer guessing. The idea is that the same goes for trusting in so far as by trusting we aim at something in a way that essentially renders us vulnerable. Monitoring a trustee (by intentionally aiming to immunise oneself from such vulnerability) alters this aim distinctive of trust, changing the performance in a way that is broadly analogous to how (through the process of collecting more evidence) one is no longer guessing, but believing.

7.4 Concluding remarks

The principal objective here has been to defend new answers to the Vulnerability Question and the Monitoring Question, answers shown to fare better than the competition. But in doing so, I've also tried to uncover an important but unnoticed way in which these questions are connected to each other, with the Vulnerability Question the more fundamental of the two. On the view defended, which views both questions through the lens of performance norms, monitoring a trustee is incompatible with trusting to the extent that, through monitoring, one intentionally aims at invulnerability to risks of betrayal that, by trusting, one essentially subjects oneself to (§III). But which risks are these? It is at this point that our answer to the Vulnerability Question kicks in: trusting essentially involves rendering oneself vulnerable to betrayal in the sense that it essentially involves subjecting oneself to risk of betrayal that is not merely *de minimis* within the relevant cooperative practice (§II.a). And – putting these ideas together – the fuller answer to the Monitoring Question, framed in terms of our answer to the Vulnerability Question, is that monitoring is incompatible with trusting insofar as one intentionally aims at invulnerability to not merely *de mimimis* risks of betrayal, viz., to not *merely* those risks

[32] See, e.g., Sosa (2015, Ch. 3).

to which, by trusting, one essentially renders herself vulnerable.

Chapter 8

Therapeutic Trust

8.1 Introduction

Suppose you are leaving town for the weekend and need someone to watch your house, feed your pets and water your plants. Now imagine two choices you might make. You might trust a reliable friend who has an established track record of responsibility and loyalty. But, you might instead trust your 16-year-old nephew with no such track record to speak of. In the latter kind of case, suppose trust is undertaken with the intended aim of bringing about (or increasing) trustworthiness.[1] Philosophers of trust often use the term 'therapeutic trust' to refer to this latter species of trust, in order to distinguish it from more standard cases of (non-therapeutic) interpersonal trust.

The matter of how exactly to characterise the relationship between non-therapeutic and therapeutic trust is contested.[2] Here is the problem in a nutshell. Philosophical accounts of the nature of trust attempt to say what trusting someone with something essentially involves[3], typically by

[1] See, for example, Horsburgh (1960) and Jones (2004).

[2] See, e.g., Horsburgh (1960) and McGeer (2008). For an overview, see McLeod (2020).

[3] The kind of trust that is principally at issue in debates about therapeutic trust is *three-*

focusing on how exactly to characterise the kind of trusting attitude one has towards her trustee. Once such accounts are made precise, it looks like therapeutic trust – given how the attitude we have in such cases about the trustee's reliability is usually much *less* optimistic than in non-therapeutic cases – either (i) simply doesn't get 'ruled in' as genuine trust on the account, or (ii) the account gets modified – perhaps stretched quite thin – in order to fit therapeutic trust in.

Here is the plan. In §§8.2-4, I discuss three notable ways philosophers of trust have attempted to deal with the tricky issue of therapeutic trust and its relationship with ordinary non-therapeutic trust: (2) Pamela Hieronymi's (2008) pure/impure approach; (3) Karen Frost-Arnold's (2014) 'unity' approach; and (4) Karen Jones's (2004) 'normative difference' approach. Each is shown to be problematic. §§8.5-8 then develops a new way of thinking about therapeutic trust which avoids the problems facing the other three views while at the same time offering its own additional advantages. §8.9 concludes by canvassing some potential objections and replies.

8.2 Hieronymi on pure/impure trust

According to Pamela Hieronymi (2008), therapeutic trust is not 'pure' or 'full-fledged' trust. Trust is full-fledged (alternatively: pure) only if one actually *believes* that the person in question will do the thing in question.[4]

place trust, e.g., with an infinitival component (schematically: A trusts B to X). As Baier (1986, 236) puts it, philosophers of trust—and not just those interested in therapeutic trust—are concerned centrally with 'one person trusting another with some valued thing'; likewise, as Katherine Hawley (2014, 2) puts it: trust is 'primarily a three-place relation, involving two people and a task'. On one popular way of thinking about the relationship between three-place trust and two-place trust, latter can be explained in terms of the former, which is the comparatively more fundamental notion. For discussion, see e.g., Ch. 1.

[4]For further support of the idea that non-therapeutic trust requires belief that the trustee will prove trustworthy, see Adler (1994), Keren (2014) and McMyler (2011). For criticism, see Jones (1996), McLeod (2002), McGeer (2008), Faulkner (2007, 2011), and Baker (1987).

One can risk betrayal by entrusting something to someone without believing they'll[5] actually do what they've been entrusted to do. But this is not 'pure' trust.

In support of this way of thinking about therapeutic trust, Hieronymi offers the following case-pair involving the betrayal of a secret.

> SECRETS: Consider two cases. In one, I fully believe you are trustworthy; in the other, I have doubts about your trustworthiness, but, for other reasons (perhaps to build trust in our relationship, perhaps because I think friends should trust one another, or perhaps simply because I have no better alternative), I decide to tell you my secret. Suppose that, in both cases, you spill the beans, and that you do so in the same circumstances, for the same reasons (2008, 230).

According to Hieronymi, once we thus hold fixed both (i) the 'importance of the good entrusted' (2008, 230); and (ii) 'the wrongness of the violation (2008, 230),' then:

> [...] it seems plausible that one's degree of vulnerability to betrayal tracks one's degree of trusting belief ... further, this seems to be because, in the second case, there was less trust to betray (2008, 230–1).

There are, however, two problems with this diagnosis of SECRETS. The first is that it's not at all clear that one's degree of vulnerability to betrayal really tracks one's degree of trusting belief, *even when* the importance of the good entrusted and the wrongness of the violation are held fixed. To see why, just suppose we run a variation on Hieronymi's SECRETS case-pair where, in the first case, my belief that you are trustworthy is full (stipulate: credence 1) but, at the same time, completely *irrational*. By Hieronymi's reasoning, the betrayal is greater by degree in the first case simply because of the irrationally ratcheted up belief. But it's not. I might, due to having this strong albeit irrational credence that you are trustwor-

[5] The singular pronouns 'they' and 'them' are used throughout whenever gender is unknown or irrelevant.

thy, be even more inclined than otherwise to *think* that the betrayal is serious. But it wouldn't in fact be a worse betrayal simply *on account of* the ratcheted up irrational credence.

But suppose, for the sake of argument, we grant that one's degree of vulnerability to betrayal tracks one's degree of trusting belief. It is worth noting that *even if* this were true, it could be explained without recourse to the idea that there is – in the therapeutic case – "less trust" to betray. For example, it might be that vulnerability to betrayal is one among various features of trust, and it is a feature that lines up with (e.g., by closely tracking) trust's doxastic component, whereas other features of trust (for example, whatever features makes it resilient to certain kinds of monitoring[6]) might track some *non-doxastic* component of trust.

If something like this were right, then we couldn't move simply from the idea that one's degree of vulnerability to betrayal tracks one's degree of belief to the conclusion that there is less trust in cases with less belief. After all, such cases might feature more prominently some other aspect of trust, and in virtue of the presence – perhaps, surfeit – of which there is not, on the whole, "less trust."

But Hieronymi has a second argument for relegating therapeutic trust the 'impure' category. The second argument has to do not with vulnerability but with the legitimacy of certain kinds of complaints. This second line of reasoning goes as follows. People can *legitimately complain* about not being trusted fully when they are trusted in the absence of belief, which occurs only when other people lack confidence in them but trust them nonetheless (2008, 230). For example, imagine the 16-year-old from our opening case saying: "But you don't *really* trust me", upon finding out that the rationale for the trust was largely trust-building, in the absence of a belief that they'd prove trustworthy. The felicitousness of such a complaint is, for Hieronymi, meant to support the idea that therapeutic trust is not pure or full-fledged trust.

[6]For discussion on this point, see, e.g., Baier (1986) and Wanderer and Townsend (2013).

This reasoning is also problematic, though, in so far as it's supposed to motivate relegating therapeutic trust to a 'second tier'. Just as the teenager could complain in this scenario, they could also felicitously praise or thank you for trusting them *despite* lacking confidence. "Wow, you trusted me without believing – you must have *really* trusted me!"[7] This is not to say that Hieronymi's example complaint is *not* felicitous, nor that praise or gratitude for trusting despite lacking confidence is any more felicitous than is complaining that one has trusted in the absence of belief. Rather, the point is that it is not clear that complaining about trust in the absence of belief is at all *more* felicitous than praising or thanking a trustor who trusts one in the absence of it.[8]

In sum, Hieronymi's arguments from vulnerability and complaint legitimacy don't give us good reason to think that the difference between non-therapeutic trust and therapeutic trust is a difference in 'purity' of trust.

8.3 Frost-Arnold's wide account

Let's look now at an attempt – due to Karen Frost-Arnold (2014) – to 'broaden' an account of trust so that it is wide enough to rule both in both varieties of trust. On Frost-Arnold's proposal, A trusts B to φ iff the proposition that B will φ is part of A's 'adjusted cognitive background (2014, 1963–4),' where one's adjusted cognitive background includes all and only those propositions that one *accepts* for the purposes of practical reasoning – where acceptance does not entail positive belief[9] (e.g., posi-

[7] For a defence of the idea that trust's most pure form involves the absence of belief, see Mollering (2006).

[8] Moreover, the felicitousness of such a reply gains some support in the literature on the psychology of gratitude see, Emmons and McCullough (2004); for instance, gratitude is a predictable response by one to another who has placed a kind of 'faith' in their good will or competence.

[9] The idea that accepting a proposition, understood as taking it for granted in one's practical deliberations—alternatively: acting 'as if' the proposition is true—does not entail that one believes the proposition to be true has been defended variously in epistemology, the philosophy of science and elsewhere. For some representative discussions of how belief and acceptance come apart, see Cohen (1989), Bratman (1992), and Buckar-

tive belief of the sort that is generally lacked in therapeutic cases, even if often present in non-therapeutic cases). This kind of 'unity' view does not relegate therapeutic trust to a second-tier, as Hieronymi's proposal does, but rather 'brings it in to the first tier' – viz., by subsuming it within a wider account of trust *simpliciter*.

There are two main problems with Frost-Arnold's 'unity'-style approach. The first is that the acceptance requirement needn't be satisfied in all cases of therapeutic trust. Suppose you trust your teenager to drive your car for the weekend and return it safely. Suppose further that, upon doing this, you purchase additional insurance, just in case. By purchasing this additional insurance, you are not accepting the proposition that the teenager will return the car safely – viz., to do what you'd trusted them to do – in the course of your practical reasoning. You act instead on the proposition that they might realistically enough not do so.[10]

But the existence of this mitigating back-up plan doesn't preclude the case from having been a case of therapeutic trust in the first place. That is, you don't suddenly cease to be therapeutically trusting the teenager with whom you aspire to build trust once you buy the insurance. It's not as though the vulnerability to betrayal one subjects oneself to is eliminated by one taking any steps whatsoever to mitigate damages against the risk occurring.[11]

Some philosophers of trust have pressed back on this point. For example, Arnon Keren (2019) holds that trusting involves declining to take precautions against the trustee's failing to come through.[12] This idea seems

eff (2010).

[10] This is the case, to be clear, even though we needn't suppose that you positively believe that the teenager *won't* return the car as entrusted.

[11] This point, it is worth noting, is compatible with the widely accepted idea that *monitoring* is incompatible with trusting, either of a non-therapeutic or therapeutic variety. See, e.g., Baier (1986, 260).

[12] Keren (2019) actually formulates his position as follows, with the qualifier 'every': 'If you rely on a person to ϕ but take *every* precaution against the possibility that she might not ϕ—by seeking evidence that might indicate that she might fail to ϕ and by acting in order to minimize the harm caused in case she fails to ϕ—then you do not trust

prima facie plausible in the epistemic case, specifically, where what the trustor trusts the trustee to do is to tell them the truth. What 'taking precautions against the trustee's failing to come through' would amount to in this case would be finding additional evidence that bears on the truth of the proposition. But *then*, having sought such evidence, it doesn't look as though you are trusting the person's word at all.

To the extent that trusting (therapeutic or otherwise) *does* involve declining to take precautions against the trustee's coming through, this might very well be idiosyncratic to the epistemic case where there is a constitutive tension between relying on one's word and acquiring the kind of evidence one would acquire by taking precautions. Crucially, we find no such tension though outside the epistemic case, at least when we hold fixed that the precautions are (as in the example of the insurance policy) precautions that are solely designed to mitigate damages *if* the trustee does not come through. Compare: this is importantly different from, and does not imply, taking precautions designed to lower the likelihood that the trustee will *fail* to come through — as one might do by hiring a team to accompany the teenage driver (see Chapter Seven). Accordingly, the attempt to reply to the objection raised to Frost-Arnold by way of appealing to Keren's insights about epistemic trust looks to come up short.

---

her to ϕ. You might rely on her to ϕ, but you do not trust her to do so' (2019, 121). As formulated, this is not controversial, as this is tantamount to the statement that trusting is incompatible with certain kinds of monitoring. What is at issue in the example I am discussing above, involving an insurance policy, is rather whether mitigating at all against the risks of the damage that would be incurred by the trustee's betrayal would be compatible with nonetheless therapeutically trusting that person. My contention that it is thus compatible with granting Keren's point that some kinds of monitoring—i.e., such as those that involve taking every precaution against the possibility the trustee won't come through—are incompatible with trust (therapeutic or otherwise). That said, Keren also makes claims about taking precautions, in the specific case of epistemic trust, which appear to go beyond the statement of his view noted above, and which appear to imply that it is essential to epistemically trusting someone that you decline entirely from taking precautions against their not coming through. Because this thesis, at least if applied generally and not just in the epistemic case, is in tension with my assessment of the insurance policy case, I focus in the main text on it as opposed to on Keren's less contentious formulation quoted above.

The upshot is that Frost-Arnold's unified account of trust which frames trust in terms of acceptance is still too narrow to do what she wants it to do, which is to rule in all cases of trust, non-therapeutic and therapeutic alike.

Even more, the proposal faces a second problem. The second problem concerns the evaluative normativity of trusting.[13] The worry is that the view lacks the resources to account for why reasonable therapeutic trust isn't just *bad* as an instance of trusting.

Continuing with the teenager car case: let's suppose you have no trust-building objectives in mind, and simply want someone dependable to drive your car for the weekend and bring it back safely. Foolishly, you choose the teenager with a record you know is patchy at best. This looks like *bad* (viz., incompetent) trust, even if it would *not* be so with therapeutic purposes in play.[14] But it's hard to see how we'd explain this normative difference on Frost-Arnold's unity-style account. One might try to begin to tell such a story by appealing to the 'epistemic constraint' that Frost-Arnold places on the kind of acceptance that matters for a proposition's being ruled-in the adjusted cognitive background. But the epistemic constraint she places on acceptance is really a very minimal one. It precludes just one thing: positive belief that the person will *not* do the thing in question. This kind of constraint won't help us in any way to adjudicate the normative question we're interested in – viz., how to

[13] Recall from Chapters One and Two that evaluative norms – unlike prescriptive norms, which prescribe conduct – regulate what it takes for a token of a particular type of thing to be good or bad with regard to its type, where the 'goodness' or 'badness' here is *attributive* in Geach's (1956) sense – viz., the sense in which a sharp knife is a good knife, *qua* knife, regardless of whether it is good or bad *simpliciter*. (Likewise, in this sense, a known belief is a good belief, regardless of whether it would be good or bad *simpliciter* – viz., as it would be were the content of the knowledge instructions for igniting a terrible bomb.For a helpful overview of the prescriptive/evaluative norm distinction, with reference to attributive as opposed to predicative goodness, see McHugh (2012, 22) and, as this distinction applies to belief specifically, Simion et al. (2016, 384–86).

[14] One sense in which the trust here is bad is that it is not likely to be *successful*, in that trusting a teenager involves incurring a relatively higher risk of betrayal than normal. For discussion on this point, see Carter (2020b).

distinguish at least some cases of good therapeutic trust from plain old bad trust.

There's another thread to this point. Just as picking out an unreliable person is 'bad' trusting (likely to lead to one's trust being betrayed[15]) even if trusting that unreliable person could have been reasonable were therapeutic purposes suitably in play, it doesn't follow that *simply stipulating* a therapeutic purpose suffices to make any instance of therapeutic trust *good* therapeutic trust. One can surely be *better or worse at therapeutic trusting*, just as one could be better or worse at trusting more generally – and indeed, very plausibly in light of different kinds of skill sets. None of this looks explicable (at least, in any straightforward way) if we embrace a unity-style view like Frost-Arnold's.

8.4 Jones on the normativity of trust

The foregoing discussion suggests that what's needed is an account of therapeutic trust which explains clearly how it features some kind of *normative difference* with respect to ordinary, non-therapeutic trust. This is exactly what Karen Jones's (2004) account of therapeutic trust purports to offer. Unfortunately, as I will argue, identifies the *wrong kind* of normative difference.

According to Jones, therapeutic trust involves the normative attitude that the trustee *ought* to do what one trusts them to do, rather than optimism that they will do it. With reference to our opening case pair: when you trust the reliable friend to watch your house for the weekend, you are optimistic that they will do this as you've entrusted them. While you're not optimistic that the teenager will do the same when you trust them with the task, you nonetheless think in trusting them that they *ought* to do what you've entrusted them to do.

There are three main problems with this proposal. First, the normative attitude that the trustee ought to do what one trusts them to do is not nec-

[15] See Carter (2020b) for a defence of this way of thinking about bad trust.

essary for therapeutic trust. Consider a case where a CEO, with the aim of striking up a romantic relationship with a low-level employee, entrusts that employee with an inappropriately enormous responsibility – hoping that doing so will help generate a trusting relationship between them as a precursor to such a romance. If the CEO is not blind to their exploitative reasons underlying the trust they are placing in this inexperienced employee, then they will not have the view that the trustee *ought* to actually do what they are entrusted to do. Quite the contrary, the CEO might well know that that the employee's succeeding in doing what they've been entrusted with is beyond reasonable expectations. But this is plausibly therapeutic trust nonetheless, in that it is – albeit for morally dubious reasons[16] – aiming to bring about and strengthen a trust relationship.[17]

A second objection to Jones's proposal – in so far as it purports to distinguish therapeutic from non-therapeutic trust – is that some cases of non-therapeutic trust involve not only optimism that the trustee will do what they are entrusted to do, but *also* the normative attitude that the trustee ought to do what they are entrusted to do. Suppose someone is drowning. I am nearby with a life vest – my expensive life vest – but my arm hurts, and so I can't throw it off the boat to help. I trust my able-bodied and reliable

[16]No assumption is being made, to be clear, that the norms of trust are moral norms. I have suggested (in responding to Frost-Arnold) that we should expect an account of therapeutic trust to be reconcilable with plausible claims about the evaluative normativity of trusting; but this commitment is a very general one—viz., to there being norms (however we best articulate them) that regulate what it takes for a token of trusting to be good or bad with regard to its type. This is at most a commitment to attributive (rather than predicative) goodness of trusting in certain cases.

[17]The same kind of point can be made with reference to a more paradigmatic kind of 'teenager trust' case. Suppose a mafioso with a conscience but a weak will accepts a hit job and, rather than to do it himself, entrusts his unreliable teenager to carry out the hit—hoping that doing so will build trust. Assume the target of the hit is known by the mafioso to be completely innocent. It is entirely plausible here that the mafioso, bearing this in mind, will appreciate that what he's entrusted the teenager to do is not something the teenager ought to do. Yet, this fact (as in the CEO case) does little to change the fact that the trust here is of a therapeutic variety. Granted, this—as well as the CEO case—relies on a weak assumption in moral psychology, which is that one can desire to bring about some state of affairs while acknowledging that its being brought about would violate one (or more) prescriptive norm. For discussion, see Stocker (1979).

friend to throw it. In this case, where I trust my friend to throw my vest, the trust isn't therapeutic in any interesting sense.[18] And yet, I have the (strong) normative attitude that the trustee *ought* all-things-considered to do what I've entrusted them to do. Thus, believing that the trustee ought to do what one entrusts them to do isn't distinctive of therapeutic trust but not ordinary non-therapeutic trust.

A third objection to Jones's proposal is that some cases of therapeutic trust positively *do* involve optimism that the trustee will do what they are entrusted to do, even if this optimism persists along with some serious doubts. To see why, it will be instructive to first consider how optimism comes apart from belief in both directions. In the literature on the psychology of optimism,[19] an optimistic attitude, with respect to some situation X, is often characterised in terms of a kind of attention profile directed at favourable features of that situation. For example, if my car breaks down and I'm stranded on a highway, then an optimistic attitude might lead me to focus my attention on how doing certain things under my control (e.g., walking to the nearest petrol station) could better my situation.

Coming back to therapeutic trust: one can distribute one's attention patterns in ways that line up with optimism (with respect to a trustee proving trustworthy) *without* having any positive belief that the trustee will prove trustworthy. (Compare: I can be optimistic when stranded without actually having the *belief* that I will be saved). For example, being optimistic that the teenager will look after the house properly or return the car might involve focusing on the teenager's good traits, feeling pride in remembering past times they've exceeded expectations, etc. This is all compatible with a lack of belief that they will in fact bring the car back. Note, furthermore, that belief and optimism come apart in the other direction as well. You could believe someone will bring a car back without being optimistic simply because your attentional profile does not line up with what you be-

[18] As we'll see in 7, though, there is plausibly an uninteresting kind of 'default' therapeutic trust in play here that is implicated by most cases of non-therapeutic trust.

[19] E.g., Carver, Scheier, and Segerstrom (2010).

lieve. You might be irrationally paranoid, given a pessimistic perspective that does not line up with your belief that the trustee will prove trustworthy. These considerations in favour of the idea that optimism can float freely of one's doxastic attitudes supports that (*contra* Jones) the kind of doubts one has in the case of therapeutic trust aren't doubts that, as such, would preclude optimism that the trustee will prove trustworthy.

In sum, Jones is mistaken that therapeutic trust involves the normative attitude that the trustee ought to do what one trusts them to do, rather than optimism that they will do it. This is because it neither requires the normative attitude that the trustee ought to do what one trusts them to do – as per the CEO case – *nor* does it preclude optimism that they will do what they are entrusted to do, at least in so far as optimistic attitudes are plausibly demarcated by their attentional profiles.

8.5 Interlude: the way forward

So far, we've seen that prominent extant accounts of therapeutic trust run into various kinds of problems. A presupposition common to each of the three views considered is that therapeutic trust is a univocal kind, and this is a presupposition we'd be better off rejecting.

There are two importantly different species of therapeutic trust – *default therapeutic trust* and *overriding therapeutic trust*. Each species of therapeutic trust interacts with ordinary (non-therapeutic) trust differently. And, each is *normatively constrained* differently from each other. Appreciating how this is so, we can – in addition to avoiding the kinds of problems considered – make sense of something other views can't, which is what makes therapeutic trust (of a philosophically interesting sort) *good* when it is.

8.6 Default therapeutic trust

As we saw in Chapter Two, we can straightforwardly situate paradigmatic non-therapeutic trust within a telic, performance-theoretic framework that is familiar in other areas of philosophy. With non-therapeutic

trust repositioned in this way, we have already a helpful vantage point to theorise about *therapeutic* trust – and in particular, about what I'll call *overriding* therapeutic trust – which is the most philosophically interesting variety of therapeutic trust.

But first it is worth making the following explicit: there is a kind of *default* therapeutic trust – viz., therapeutic insofar as it (trivially) aims at trust-building – that is *implicit* in paradigmatic cases where one trusts with the aim that the trustee take care of things as entrusted. We take for granted in trusting that trust will – apart from whatever else it does – play its normal social functions, functions that plausibly include the social function of strengthening trust relations.[20] This is so even when we trust – as we do when we seek out someone reliable and trustworthy – with the basic aim that the trustee take care of things as entrusted.

In this respect, there is a minimal and trivial kind of therapeutic trust that is going to be implicit in garden variety (non-therapeutic) trusting, and this is so even in the absence of any explicit *intention* – the kind of intention that is explicit in teenager-style cases – to satisfy the aim of building trustworthiness.[21]

Moreover, the 'implicit' kind of therapeutic trust that accompanies ordinary trust (both implicit and deliberative) as a default does not have its own constitutive aim. This is because default therapeutic trust is just *implicated by* normal, non-therapeutic trusting, which itself constitutively aims at the trustee's taking care of things as entrusted (or, in the deliberative case, at one attaining that aim by trusting aptly). In this respect, the

[20] For some representative defences of the role of trusting in trust-building, see Faulkner (2011, Ch. 1), Alfano (2016), Hall (2005), and Solomon and Flores (2003).

[21] Consider, by way of analogy, one of the plausible social-epistemic functions of *assertion*, which is to generate knowledge in the hearer e.g., Kelp (2018) and Simion (2019); cf., Williamson (2002). On the assumption that assertion has such an aim, constitutively, it's easy to see how asserters *implicitly*, in asserting, aim at other things (even if not intentionally), namely, whatever generating knowledge in a hearer generally involves, including playing roles that knowledge normally plays for the hearer. For example, one role that knowledge plausibly plays for a hearer who acquires it is that of being a possible premise in the hearer's practical reasoning e.g., Hawthorne and Stanley (2008).

competences that are relevant – trivially – to default therapeutic trust are just those that matter for non-therapeutic trust that implicates it.

8.7 *Overriding* therapeutic trust

The most interesting kind of therapeutic trust is not default therapeutic trust, but *overriding therapeutic trust*. This occurs when, as in our paradigmatic teenager cases, the aim of successful trust – given perceived vulnerabilities – isn't itself enough to *motivate* one to risk trusting. Simply wanting your house to be watched over responsibly wouldn't, from the perspective of a trustor, favour entrusting such a task to the teenager, as opposed to someone regarded to be more reliable; rather, the opposite would be the case. Necessary for bringing about overriding therapeutic trust is thus an 'overriding' and intentional aim- – *the aim of building or strengthening trust* – that is distinct from the constitutive aim of ordinary non-therapeutic trust.[22]

Unlike default therapeutic trust that is implicit in most normal trusting, *overriding* therapeutic trust is a distinct kind of performance from normal (non-therapeutic) trust, with its own constituent normativity. The constitutive aim of overriding therapeutic trust is not *merely* to trust successfully (viz., as captured by ESNT in Chapter Two). But nor, it should be emphasised, is it *merely* to build trust. It is – and this is the key idea – to *build trust through trusting successfully*.

Consider that just as ordinary (non-therapeutic) trust is defective when it misses its internal aim (that the trustee take care of things as entrusted), your choosing to trust your teenager to watch over the house has missed *its* mark if either (i) trust is not built (e.g., if a result of this trusting is not a strengthened trust relationship) *or* if (ii) trust is not successful (e.g., if the

[22]Nothing much hangs on whether the relevant contrast between therapeutic and non-therapeutic trust is between therapeutic and implicit non-therapeutic trust or between therapeutic and deliberative non-therapeutic trust. Thus, I will be discussing trust loosely here in the non-therapeutic case, given that nothing of substance turns on the implicit/deliberative distinction.

teenager throws a party, during which items from the house are stolen). *Even more*, though, your trust will have missed its mark even if the trust serves to build trust *and* the trust is successful, but (iii) if the trust built is not built *through* the successful trust, but for some reason disconnected with the therapeutic trust placed in them. This might be the case, for example, if the teenager watches over the house successfully, though – unaffected entirely by the trust you've placed in them – comes to trust you more nonetheless due to having, while watching over the house, spent some time reading false accounts of sacrifices you've made for them in the past, and only on this basis, develops toward you a stronger bond of trust.

Question: if overriding therapeutic trust constitutively aims not at mere successful trust, nor at the mere building of trust, but at building trust *through* successful trust, then what do (i) competent and (ii) apt overriding therapeutic trust consist in?

Competent overriding therapeutic trust, on the telic model we are working with, will derive from a competence to attain the constitutive aim *of overriding therapeutic trust* reliably, which is the aim of building trust *through* successful trust. Given that competences are indexed to performance conditions, a clear view of the kind of competence that matters for overriding therapeutic trust requires an understanding of the conditions under which reliable performance matters for this *particular* kind of trusting. These conditions include (at least) the satisfaction of what I'll call an *openness condition* and a *reciprocity condition*.

To appreciate the former condition, consider the following case:

> DIANE: You need someone to babysit on short notice. There are a number of people you could ask, however, you choose a local teenager, Diane, whose parents you know. You have heard that Diane is troubled, and you have had a standing desire to take Diane under your wing in hopes of having a positive influence on her. A first step toward having such a positive influence, you think, will be to establish a bond of trust, a bond you hope to develop by entrusting her with the babysitting task despite

> her reputation. Unfortunately, and unbeknownst to you, Diane recently experienced a highly traumatic event, to which she has responded by closing off the possibility of developing a trusting relationship with *anyone*, at least until she has worked through this trauma. She succeeds in the task of babysitting, though at no point was she in a position where her being entrusted with this would have changed her distrusting stance of others.

In DIANE, the conditions for successful overriding therapeutic trust are simply not in place *ex ante* – and this is so *even though* the conditions in DIANE do not preclude her in any way from taking care of things as entrusted. With respect to the aim you have of of building trust through successful trust, Diane is 'closed'. She is not in a position where your trusting her with the task that you do could – even when that trust is fulfilled by her – contribute to *building* trust on account of that fulfilment.

Now consider a twist on this case:

> DIANE*: You need someone to babysit on short notice. There are a number of people you could ask, however, you choose a local teenager, Diane*, whose parents you know. You have heard that Diane* is troubled, and you have had a standing desire to take Diane* under your wing in hopes of having a positive influence on her. A first step toward having such a positive influence, you think, will be to establish a bond of trust, a bond you hope to develop by entrusting her with the babysitting task despite her reputation. Diane* is open in principle to building trust with someone who would entrust her with this kind of task. Unfortunately, and unbeknownst to you, Diane* bears a deep-seated grudge against you. Though she babysits the kids successfully (suppose, she needs the money) – and though her doing so successfully in fact contributes to making her more trustworthy *generally* speaking – it plays no role in establishing or strengthening any trust between

you and her.

Diane* is *not* closed to building trust through successful trust, as Diane is, generally. However, the conditions in DIANE* are such that they prevent building her trust *with you*, the trustor, through successful trust. This is not to say that Diane*'s grudge would never subside so as to open up such a possibility later. The point is that the situation in which you encounter Diane* is not one in which, were she to come to establish and build trust with you, this could be achieved in the way you're attempting to do so here – viz., through facilitating successful trust via the babysitting task.

There are two interrelated points to draw from the DIANE and DIANE* cases. The first is that it doesn't count against one's overriding therapeutic trust competence, viz., one's disposition to attain the aim of overriding therapeutic trust reliably enough, were one to be *unreliable* at attaining this aim in cases like DIANE or DIANE*, where the conditions are, for different reasons, not suitably conducive to building trust through successful trust. Secondly, and relatedly: the kind of competence that *matters* for overriding therapeutic trust is, accordingly, a disposition to build trust through successful trust reliably enough when one is in conditions that *are* appropriate to doing so, conditions that include at least that openness and reciprocity are satisfied, as they are not in DIANE and DIANE*, respectively.

A further point is that *when* these conditions are met, some are disposed to achieve the aim of overriding therapeutic trust *more* reliably than others. And that is just to say that, when it comes to overriding therapeutic trust, some are *more competent* than others, some of whom simply lack this competence by not being suitably reliable in conditions that are favourable to this kind of trust.

What makes the difference? One factor that's worth noting explicitly is that we vary in the capacities we have to reliably assess trust-building payoffs. For example, recall our case of the CEO (4) who entrusted the low-level employee with a *disproportionately* large task, one which not easily the employee would have managed. The overriding therapeutic trust is unlikely to payoff here simply given that the difficulty of the task choice will

make unlikely the building trust through *successful* trust. Conversely, entrusting *too small* a task, with therapeutic aims, is likewise unlikely to payoff, though, for a different reason. (Compare: suppose you were to, with trust-building aims, entrust a teenager not with looking over the house or the kids, but with looking after a small cactus for the weekend). The task is not certainly too difficult to undermine the likelihood that the trustee will take care of things *as entrusted*, but it is *so easy* that it undermines the likelihood that, through being undertaken successfully, it will play a (non-negligible) role in increasing any kind of trust bond with the trustee. In short: (i) a propensity to miss the mark too often in *either* direction will undermine one's reliability at attaining the aim of building trust through successful trust, and so (ii) a competence to hit this aim reliably (when appropriately situated to do so[23]) requires a capacity for the kind of risk assessment that's needed to prevent one from too often 'over' or 'under' trusting (as in the CEO and cactus cases, respectively).

Let's return now to our telic evaluative norms for overriding therapeutic trust. *Accurate* or successful overriding therapeutic trust occurs when overriding therapeutic trust hits its constitutive aim, which is the aim of building trust through successful trust. *Competent* overriding therapeutic trust issues from a *competence* to hit this aim reliably enough when one trusts with a therapeutic aim whilst appropriately situated – where being appropriately situated for this kind of trust requires at least the satisfaction of the openness and reciprocity conditions. *Apt* overriding therapeutic trust can now be defined in terms of accurate and adroit therapeutic trust – viz., apt overriding therapeutic trust is overriding therapeutic trust that is *accurate because adroit*, viz., when one's building trust through trusting successfully manifests one's competence to therapeutically trust successfully reliably enough in appropriate conditions.

Apt overriding therapeutic trust is a kind of *achievement*, just like any kind of aim attained through skill rather than by other means.[24] In this respect,

[23] That is: when the trust environment is such that the openness and reciprocity conditions described in this section are met.

[24] For some representative discussions of the value of achievements understood as having a success-through-ability structure, see, e.g., Bradford (2013, 2015a), Sosa (2010b),

apt overriding therapeutic trust stands to *mere* successful overriding trust as knowledge to lucky true belief, and to an archer's successful shot attained through skill to the same success attained any old way. However, as we've seen, the achievement of apt overriding therapeutic trust is a *different* achievement than the achievement of apt (non-therapeutic) trust, one that involves the attaining of a different aim through the manifestation of a different sort of competence.

8.8 Summing up

We began with a puzzle about therapeutic trust and its relationship to ordinary non-therapeutic trust. Three prominent attempts to address this puzzle were considered, and each was shown to be problematic for different reasons. One notable problem common to each of the three accounts was that none was well-suited to explain – given what each maintains, respectively, about therapeutic trust and how it differs from non-therapeutic trust – in virtue of *what* good therapeutic trust differs from plain old bad trust, including incompetent trust that just so happens to result in the building of trust, as well as successful *and* competent trust that builds trust for reasons that have nothing to do with the trust placed.

The account I've proposed has a number of advantages over these accounts. First, it avoids the traps that these other accounts were shown to fall into given their specific commitments. The key move proposed which helps to get things right involves the recognition of two kinds of therapeutic trust. There is a philosophically uninteresting species of therapeutic trust that is implicit in ordinary trusting – what I called *default therapeutic trust*. While default therapeutic trust (trivially) aims at building trust, it does so only because building trust is among the normal social functions of ordinary non-therapeutic trust, which aims constitutively at the trustee taking care of things as entrusted. *Overriding therapeutic trust*, by contrast, has its own constituent normativity – with

Carter, Pritchard, and Turri (2018), Greco (2014), Miracchi (2015), Pritchard (2009b), and Zagzebski (1996).

reference to which we can normatively assess this kind of trust differently from how we normatively assess standard trust. In doing so, we can say *why* each kind of trust is good when it is, without reducing the goodness of either kind of trust to the goodness of the other. Moreover, the view can help us to make sense of how the the skills needed for reliable therapeutic trust come apart from the skills needed to be good at trusting well more generally; the 'SSS' profiles of competent trust and competent therapeutic trust differ in clear ways. Finally, by distinguishing between ordinary apt trust and apt (overriding) therapeutic trust on the model proposed, we have a perspective from which to appreciate two different *achievements* in trusting and why neither of these achievements reduces to the other.

8.9 Objections and replies

8.9.1 (Objection 1).

On the view proposed, the constitutive aim of overriding therapeutic trust is meant to be distinct from the constitutive aim of standard (i.e., non-therapeutic) trust in that: (i) the aim of the former is that the trustee take care of things as entrusted; whereas, (ii) the aim of the latter is to build (or strengthen) trust *through* successful trust, viz., through the trustee's taking care of things as entrusted.

However, the suggestion that these aims are distinct is not so clear given that the view *also* holds that building trust is among the normal social functions that is played by (successful) ordinary trust. But if *that's* right, then isn't it the case that standard trust constitutively aims not *merely* at the trustee's taking care of things as entrusted, but also, at this fact playing the social function of strengthening trust? If so, then it looks like the claimed difference between the constitutive aims of standard trust and overriding therapeutic trust collapses.

Reply: The fact that the constitutive aim of ordinary trust – viz., that the trustee take care of things as entrusted – is such that when this aim is met, it's doing so has a characteristic social function, *X*, does not imply that its

actually playing that function, *X*, is thereby included *as part of the constitutive aim*. The aim would still be met even if that social function characteristic of attaining that aim were *not* played.[25] (Compare: the aim of archery – hitting the target – is attained even if your hitting the target does not play any of the roles that attaining this aim would characteristically play, e.g., to build confidence, solidify social standing with peers, signal competence, etc.). Likewise, if you trust a reliable colleague to deliver an envelope to your boss without reading the message inside, and the colleague successfully does so without taking a peek, there is a clear sense in which your trust placed in your colleague on this occasion has attained *its* aim – no matter what further social functions your trust plays or does not play, including social functions you might reasonably *expect* it to play.

8.9.2 (Objection 2).

The competences involved in ordinary (non-therapeutic) trust and overriding therapeutic trust are claimed to be *different* competences. But is this really so? Here is a reason to think the answer is 'no'. Adroit overriding therapeutic trust issues from a competence to reliably enough build trust through successful trust when one attempts to do so while appropriately situated. But then – being reliable at *this* was said to require a capacity for the kind of risk assessment that's needed to prevent one from too often 'over' or 'under' trusting (as in the CEO and cactus cases, re-

[25] There is a precedent for this kind of thinking about aims and defective functioning found in Burge (2003). According to Burge, evidence that something is or is not operating defectively offers us insight into what it's aim (or, for Burge, function) is (or is not). For example, if we did not regard the heart as defective if it failed to pump blood, then this would cast doubt on the idea that the heart is normatively constrained by the aim of pumping blood. By parity of reasoning: my suggestion is that—in both archery and in ordinary trust—we would not regard a shot as defective if it hit the target but did not inspire confidence (a normal social function which, suppose, hitting a target plays) nor (ordinary) trust as defective if the trustee took care of things as entrusted despite this fact not going on to build further trust. This—with reference to the kind of reasoning we find in Burge—counts against the aim of archery being 'hitting the target *and inspiring confidence*', which is surely the right result, and likewise against the aim of ordinary trust as being 'that the trustee take care of things as entrusted *in a manner than builds trust*'.

spectively). But even if that's right – and here's the worry – doesn't being reliable at attaining the aim of *ordinary* trust *also* require a capacity for this very kind of risk assessment? That is: a competence to attain the aim of ordinary trust reliably enough (when appropriately situated) surely requires a capacity to evaluate risks of *betrayal*, including risks of betrayal generated by, e.g., incentives the trustee has to betray, the difficulty of the task relative to the trustee's perceived abilities, etc. But once these points are granted, the distinction between the substance of the competences relevant to (i) ordinary trust versus (ii) overriding therapeutic becomes blurred.

Reply: In short, the answer is the kind of competence that matters for overriding therapeutic trust *asymmetrically entails* the kind of competence that matters for ordinary (non-therapeutic trust). While risk assessment is undeniably important to both kinds of competences, and thus to both adroit overriding therapeutic trust as well as adroit ordinary trust, the kind of risk assessment that competent overriding therapeutic demands is more sophisticated, and accordingly more demanding, than the kind of risk assessment that competent ordinary trust demands. Given that the constitutive aim of ordinary therapeutic trust (that the trustee take care of things as entrusted) is a *component* of the constitutive aim of overriding therapeutic trust (that trust is built through successful trust – viz., through the trustee taking care of things as entrusted), reliably attaining the latter will require the very same kind of risk assessment needed to reliably secure the former, *plus* the capacity to assess *additional* risks – risks specifically to the non-obtaining of *trust built through successful trust*. Ordinary trust competence doesn't demand one have the capacity to assess these further risks.

The above illuminates an interesting wider point about the difference between ordinary and overriding therapeutic trust, which is that the latter is, in short, *more difficult* to do well. Being competent at overriding therapeutic trust requires all the skills required to be competent at ordinary trust, plus others which the latter doesn't require. A corollary is that the achievement of *apt* overriding therapeutic trust is more substantial, ar-

guably more valuable[26], than the achievement of apt ordinary trust, in that the former issues from a comparatively more sophisticated and demanding kind of competence to acquire and exercise.

8.9.3 (Objection 3).

Certain kinds of 'forced-choice' cases seem like they would work as counterexamples to the proposed account of overriding therapeutic trust. Consider the following:

> FORCED CHOICE: You've just moved to an apartment building in a new city, where the only person you know is teenager who lives in the flat below you. You need to leave town for the weekend – suppose your job depends on it – and need someone to feed, water and walk your dog (you've tried kennels, etc., and all are fully booked). Your hand forced, you trust the teenager who lives in the flat below you with this task – someone whom, had you had a better range of options – you wouldn't have chosen, as they've not established any track-record yet of responsibility with you, and your dog's welfare is important to you.

Two things seem, *prima facie*, to be true in FORCED CHOICE. First, (i) it looks like a case of therapeutic trust of a philosophically interesting sort (you are, after all, placing trust in a teenager to whom you wouldn't ordinarily trust a task like this); but, second, (ii) it doesn't get ruled in on the account proposed. This is because in FORCED CHOICE, it is *not* the case that the aim of ordinary trust is *not* sufficient to lead the truster to risk trusting. That aim *is* sufficient, *ex hypothesi*.

Reply: My response to FORCED CHOICE is to accept (ii) and press back against (i). It is a mistake to think that all cases in which one trusts a non-ideally suited trustee (e.g., by selecting someone regarded as being less reliable than would be preferred) are, in virtue of this, 'therapeutic' in

[26]For some notable arguments that difficulty adds value to achievement, see, e.g., Bradford (2013, 2015a) and Pritchard (2009a).

some interesting sense. On the view I've proposed, therapeutic trust of a philosophically interesting sort misses its mark – viz., is defective – even if the trust is successful, provided the trust fails to build or strengthen through this successful trust. FORCED CHOICE, however, is a case where the trust placed in the teenager *succeeds perfectly well* so long as the the teenager takes care of the dog as entrusted. This is so in a way that is not interestingly different than were the teenager perceived to have been much more reliable than they are actually perceived to be. The situation is, however, very different if we suppose that the basic aim of successful trust *weren't* enough (as it is in FORCED CHOICE) to motivate you to risk trusting the teenager – viz., as would be the case when you trust the teenager, e.g., rather than someone you think has a better track record, with the aim of using successful trust to building trust. The above diagnosis not only explains why we would be wrong to, in short, lump all high-risk cases together – but it also helps to highlight the important sense in which therapeutic trust of the philosophically interesting kind *uses* trust in a way that ordinary trust does not.

8.10 Concluding remarks

The aim of this chapter has been to reconcile the telic theory of trust with the vexed topic of *therapeutic* trust. To this end, I've shown how other attempts in the literature to make sense of the relationship between therapeutic and non-therapeutic trust have come up short. I've then used the core ideas are the heart of the telic theory of trust to offer a more promising account of this relationship. On the view defended here, therapeutic trust is a genus with two species, *default therapeutic trust* and *overriding therapeutic trust*, where the latter – which aims constitutively at *building trust through trusting successfully* – is of principal interest philosophically. I've shown how the evaluative normativity of overriding therapeutic trust relates to the evaluative normativity of standard (non-therapeutic) trust of the sort outlined in Chapter Two, and I've shown that the account given holds up against anticipated objections.

By this point in the book, the telic theory of trust been thoroughly de-

veloped, far beyond the core evaluative norms we began with in Chapter Two. But there is an important piece of the puzzle still missing.

We've been focusing almost exclusively on, to put it bluntly, what is going on on the *trustor's side* – and to such an extent that we've yet to ask much about the *trustee*. Consequently, we've given a theory of trust without yet articulating the relationship between trust and *trustworthiness*. The final chapter will focus entirely on this relationship, and it will show that our telic theory offers the best account of it yet.

Chapter 9

Trust and Trustworthiness

9.1 Introduction

What is the relationship between trust and trustworthiness? The question is fraught one, not least because philosophers of *trust* have tended to focus on three-place trust (S trusts X with ϕ), whereas philosophers of *trustworthiness* have been primarily concerned with two-place trustworthiness – viz., (S is trustworthy).[1]

This mismatch in focus presents challenges for those who want their accounts of trust and trustworthiness to be mutually illuminating.[2] And it also prompts deeper methodological questions, such as whether we ought to be trying to understand trust in terms of trustworthiness (as some philosophers have[3]) or trustworthiness in terms of trust (as others

[1] As Hardin (1996) notes, a further complication is that 'Many discussions of trust run trust and trustworthiness together, with claims about trust that might well apply to trustworthiness but that seem off the mark for trust' (1996, 28).

[2] For example, to control for the mismatch in focus, should we try to 'translate' three-place trust in to two-place trust in order to best connect trust with (two-place) trustworthiness? Or, would it be better to do things the other way around, to 'translate' two-place trustworthienss in to three-place trustworthiness in order to connect trustworthiness with (three-place) trust?

[3] See, e.g., O'Neill (2018, 293), Ashraf et al. (2006, 204), McLeod (2020, sec. 1),

have)[4]?

Though there is little consensus here, a widespread underlying assumption is that central phenomenon of interest on the trustee's side is *dispositional* (viz., trustworthiness) whilst the central phenomenon of interest on the trustor's side is *non-dispositional* (viz., trust). A byproduct of this assumption is that the evaluative norms of principal interest on the trustor's side regulate trusting attitudes and performances whereas those on the trustee's side regulate dispositions to respond to trust.

The aim here will be to highlight some unnoticed problems with this asymmetrical picture and to show that a symmetrical, 'achievement-first' picture has important advantages. The view I develop is guided by a structural analogy with practical reasoning. Just as practical reasoning is working as it should only when there is realisation (knowledge and action) of states (belief and intention) with reverse directions of fits (mind-to-world and world-to-mind), likewise, cooperation between trustor and trustee is functioning as it should only when there is an analogous kind of realisation on both sides of the cooperative exchange – viz., when the trustor 'matches' her achievement in trusting (an achievement in *fitting reliance to reciprocity*) with the trustee's achievement in responding to trust (an achievement in *fitting reciprocity to reliance*). An upshot of viewing cooperation between trustor and trustee as exhibiting achievement-theoretic structure is that we will be better positioned to subsume trustworthiness (and its cognates on the trustee's side), like trust, under a wider suite of evaluative norms that regulate attempts, dispositions, and achievements symmetrically on both sides of a cooperative exchange, with 'matching achievements' as the gold standard.

Here is the plan. §1 clarifies and criticises the kind of asymmetric picture that is embraced in the philosophy of trust and trustworthiness, which privileges performances (and norms regulating them) on the trustor's side

[4]For example, according to Wright (2010), trustworthiness requires that the trustee 'acknowledges the value of the trust that is invested in them ... and to use[sic.] this to help rationally decide how to act' (2010, sec. 3.b.). Other accounts of trustworthiness in terms of trust are found in Williams (2000) and Potter (2002, 205).

and dispositions (and norms regulating them) on the trustee's side. §9.2 develops an analogy between practical reasoning and cooperation in order to motivate an alternative picture, on which trusting and trustworthiness are better understood as having achievement-theoretic structure with reverse directions of fit. §9.3 builds on this picture in order to defend symmetrical evaluative norms – norms of *success*, *competence*, and *aptness* – on both sides.

9.2 Trust and Trustworthiness: doing versus being?

The distinction between trust and trustworthiness is almost invariably described as distinction between doing something (i.e., trusting) and being a certain way (i.e., being trustworthy) on account of having a dispositional property.[5]

For example, according to a family of views defended by Annette Baier (1986), Karen Jones (2012), and Zac Cogley (2012), trustworthiness is to be identified with a disposition to fulfil commitments, in conditions under which one has those commitments, and *in virtue of* goodwill towards the trustor. For Diego Gambetta (1988), the trustworthy person needn't be disposed to fulfill the commitments they have out of good will; they simply must be disposed to fulfil their commitments, whatever they are, 'willingly'. More minimalistically, Christoph Kelp and Mona Simion (2021) identify trustworthiness with the disposition to fulfill one's commitments *simpliciter*, and not necessarily through any distinctive motivation or accompanying attitude. More weakly, for Katherine Hawley (2019), the relevant disposition 'trustworthiness' refers to is best framed

[5]Possessing a dispositional property is not itself a matter of being in a mental state or behaving some way. Rather, dispositional properties 'provide the possibility of some further specific state or behaviour' (see, e.g., Mumford 2016) in certain conditions. Accordingly, the various 'accounts of trustworthiness' on the market are not aiming to give an account of any mental state, attitude, or behaviour; they are aiming to characterise accurately the dispositional property that they take trustworthy people but not others to possess.

negatively – viz., as a disposition to *avoid unfulfilled* commitments. By contrast with all of these views, Nancy Potter (2002), insists that the relevant disposition lining up with trustworthiness should be understood as a full-fledged moral virtue – one that consists in being disposed to respond to trust in appropriate ways.[6]

Despite their differences, these accounts all retain the fundamental idea trustworthiness and trust stand to each other as a (mere) 'being' to a 'doing.'[7] And thus, to the extent that these accounts are conducive to theorising about trustworthiness and trust in terms of each other, it will be as a being illuminates a doing, (or vice versa).

While there is no doubt that being trustworthy corresponds with possessing *some* disposition, so likewise does *being a competent trustor*, e.g., being one who trusts in ways that don't too often lead to betrayed trust.[8] And by the same token, just as trusting is itself not a disposition but an activity or performance, so likewise is the trustee's *manifestation of trustworthiness* when taking care of things as entrusted[9], viz., when actually reciprocating the trust placed in her (as opposed to merely 'being the sort of person' who would take care of things as entrusted).

Is there any good reason that would justify the asymmetrical focus on the dispositional property of trustworthiness and not on the trustee's performance of manifesting trustworthiness through reciprocity?

[6] According to Potter (2002), trustworthy persons "[...]give signs and assurances of their trustworthiness" and "They take their epistemic responsibilities seriously' (2002, 174–5; Cf., for criticism, Jones 2012, 75–76) and Kelp and Simion (2021). Note, however, that Jones (2012) is more sympathetic to the idea of trustworthiness as a virtue in the special case of what she calls 'rich' rather than 'basic' trustworthiness. See, e.g., (2012, 79).

[7] Or, alternatively, with reference to Vendler/Kenny classes – as an occurence (trust) to a state (trustworthiness). See, e.g., Verkuyl (1989).

[8] For a defence of this way of thinking about competent trust, see Carter (2020b).

[9] The locution 'as entrusted' is meant to encompass views on which the trustee counts as taking care of things as entrusted only if doing so in a particular way, including, e.g., out of goodwill (Baier 1986; Jones 1996) or in conjunction with a belief that one is so committed (e.g., Hawley 2014). The present discussion – which is theoretically neutral on this point – is compatible with opting for either such kind of gloss.

There would be if the disposition of the trustee (rather than any performance on the trustee's part) features essentially in a plausible specification of what one aims at in trusting, and thus, in explaining when trust is successful.

Such a line of thought is implicit in what Carolyn McLeod (2020) takes to be a platitude about trust, which is that 'Trust is an attitude that we have towards people whom we hope will *be trustworthy*, where trustworthiness is a property not an attitude'[10]. Variations of this idea are seen in Elizabeth Fricker's (2018) claim that '[...] One is not really trusting unless one adopts an attitude of optimism to the proposition that the trustee *is trustworthy*' (2018, 6). Likewise, as Russell Hardin (2002) puts it, 'trusting someone in some context is simply to be explained as merely the expectation that the person will most likely *be trustworthy*' (2002, 31). And perhaps most directly, proponents of doxastic accounts of trust (Hieronymi 2008; McMyler 2011) straightforwardly identify trust with a belief that the trustee *is trustworthy*.[11]

Of course, we seek out a trustworthy person when initially deciding *whether* to trust or forbear from trusting; in this respect, Onora O'Neill is right that 'where we aim [...] to place and refuse trust intelligently *we must link trust to trustworthiness*' (2018, 293). But when we actually trust someone, the relevance of the trustee's simply 'being a certain way' independent of their actually performing in a way that manifests how we perhaps hope or believe they are (i.e., trustworthy) – is not clear at all.

When I trust you to pay back the loan, I rely on you to pay it back, making myself vulnerable to your betrayal.[12] Suppose you *do* then pay it back. Is

[10] See McLeod (2020, sec. 1, my italics).

[11] For a related though less standard kind of doxastic account, see, e.g., Keren (2014, 2019).

[12] For various expressions of the idea that trust essentially involves subjecting oneself to risk of betrayal, see, along with Hardin (1992), e.g., Baier (1986, 244), McLeod (2020, sec. 1), Nickel and Vaesen (2012, 861–2), Carter (2020b, 2301, 2318–9), Carter and Simion (2020, sec. 1.a), Becker (1996, 45, 49). Dasgupta (1988, 67–68), Dormandy (2020, 241–2), Kirton (forthcoming), O'Neill (2017, 70–72), Potter (2020, 244), and Hinchman (2017). Cf., Pettit (1995, 208).

my trust successful? Not necessarily, says the proponent of the idea that trustworthiness is of special interest in understanding trust. In trusting, I aim *not just that you* take care of things any way, but take care of things *as entrusted*, which (on this line of thought) involves your 'being trustworthy'.

This is partly right. But it gets an important thing wrong. Just as my trust isn't thereby successful if you merely take care of things *any* old way (e.g., by attempting to betray me, but in doing so accidentally pay back the loan – then only my *reliance* would be successful), likewise, my trust misses the mark if you simply *are* trustworthy but (perhaps due to bad luck) *don't* pay back the loan. But crucially – even more – there is a sense in which my trust *still* misses the mark if you (i) pay back the loan; (ii) are trustworthy; but (iii) your paying back the loan doesn't manifest your trustworthiness (e.g., perhaps despite being trustworthy you pay back the loan on this occasion under threat or through some kind of manipulation by a third party.[13])

This suggests a revisionist picture of the original assumption.[14] Trust aims not at the trustee merely *being* a certain way – or even at the trustee doing a certain thing while at the same time being a certain way – but at the trustee *achieving* a certain thing, viz., succeeding in taking care of things *through* their trustworthiness.

[13]Coercion isn't essential to making this kind of point; for example, the trustworthy person might be such that her success (in taking care of things as entrusted) doesn't manifest her trustworthiness not because she lacked the opportunity to do so (as would be the case if she were coerced) but rather due to the abnormal presence of luck in accounting for the success. The underlying idea here – one that has been defended variously by Sosa (2007), Greco (2010, Ch. 5), Pritchard (2012), and Zagzebski (1996) – is that a success doesn't manifest one's reliable disposition (construed as an ability, virtue, or competence) if that success is unusually due to luck. How to unpack 'unusually' (alternatively: abnormally) is a contested point, one that features centrally in discussions in virtue epistemology of achievement, luck, and credit. See, e.g., Turri et al. (2019, sec. 5, §7).

[14]That is, it suggests we revise the wisely shared assumption in the philosophy of trust that the theoretical focus on *trustworthiness* qua disposition (as opposed to, e.g., focusing on performances of trustworthiness) is justified in light of the importance of this disposition to understanding trust.

This insight offers us a new vantage point for revisiting the relationship between trust and trustworthiness, and to appreciate some important performance-theoretic symmetries between the two that the characteristic focus on trustworthiness as a mere disposition has so far obscured.

9.3 Structural analogies with practical reasoning

Whereas mere *reliance* is successful just in case the person relied on takes care of things *any way*, the success conditions for trust are more demanding. Where we've got to so far is that *trust* as opposed to mere reliance is successful just in case the trustee *manifests her trustworthiness* in successfully taking care of what the trustor relies on her to do. And this involves, on the part of the trustee, a kind of *success through trustworthiness* – viz., an *achievement* in trustworthiness.

Let's now take this working idea – that the trustor aims in trusting at the trustee's *achievement* in trustworthiness – even further. Just consider that when the trustor *herself* attains this aim (i.e., the aim that trustee's taking care of things manifests her trustworthiness) – this might on some occasions of cooperation be down to dumb luck; they might foolishly trust the one trustworthy person in the village of tricksters, but this lone trustworthy person might then manifest her trustworthiness full well in taking care of things.[15]

Trust is *successful* here. *And* the trustee exhibits an achievement in ful-

[15] In this example, we are to suppose that the trustor is not manifesting any trusting competence here, but rather, simply and naively trusting and just happens to be lucky. The structure of the case is importantly different from a case where the trustee *does* manifest competence, in a normal environment, but could have easily trusted someone who was mistaken in that environment. In the latter case, the structure is analogous to that of a 'fake barn case' – and that is a case which, at least within performance-theoretic epistemology, there is no barrier to attributing the success to the ability and thus to attributing achievement. For discussion, see, e.g., Sosa (2007, Ch. 2), Littlejohn (2014), Carter (2016a), Pritchard (2009a), Jarvis (2013), and Kallestrup and Pritchard (2014).

filling the trust placed in her through trustworthiness. But, in this situation, there is no symmetrical achievement (success that manifests a trusting competence) on the *trustor's* side, even though trust is successful. And in this respect, there is an important sense in which the cooperation itself still falls short; the cooperation does not match 'achievement to achievement', but matches merely success (by the trustor) to achievement (by the trustee).

Of course, the symmetry can be regained if we simply shore up the performance on the trustor's side. Suppose it is *not* simply through good fortune but *through the trustor's competence* (to trust *successfully* reliably enough) that the trustor trusts successfully. In such a case, the relevant trust is not just successful but 'apt' in that the successful trust manifests the trustor's competence to bring that success about reliably. This *apt* (and not merely successful) trust, an achievement of trusting, on the trustor's side would then match the trustee's achievement in trustworthienss. And now *cooperation* is functioning well in that the cooperation between the two *falls short on neither side* of the cooperative exchange.

An analogy is useful here between (i) the symmetrical picture just described of cooperation working well; and (ii) Tim Williamson's (2017) view of practical reasoning working well. According to Williamson, a practical reasoning system is working well when and only when one acts on what one knows.[16] One is in a position to act on what one knows only if one 'realises' two kinds of states, with *reverse directions of fit* (mind-to-world and world-to-mind). Accordingly, on Williamson's picture, practical reasoning is not functioning as it should if there is a defect on either on the mind-to-world side (i.e, mere belief rather than knowledge) or on

[16]This idea, originating in Williamson (2002), is given a sustained defence in his (2017) with some further updates in (2021). Whereas Williamson encourages us to view the idea that practical reasoning's working well is a matter of acting on knowledge in the service of a wider criticism of the centrality of belief-desire psychology as explanatorily central, the core normative idea that, in practical reasoning, one should act only on what one knows has received defences by (along with Williamson) Hawthorne and Stanley (2008), Stanley (2005), Fantl and McGrath (2002). For an overview, see Benton (2014, sec. 2.a).

the world-to-mind side (i.e., mere intention rather than action).[17]

The working analogy so far is this: just as practical reasoning is functioning well only when we have symmetrical realisation (knowledge and action) of states (belief and intention) with reverse directions of fits, likewise, cooperation between trustor and trustee is functioning well only when we have an analogous kind of *symmetrical realisation on both sides of the cooperative exchange* – viz., when the trustor 'matches' her achievement in trusting with the trustee's achievement in responding to trust.

This working analogy can be extended further, by considering how the trustor's and trustee's matching achievements, when cooperation is working well, are themselves (like knowledge and action) *realisations of attempts with reverse direction of fit*.[18] To a first approximation: whereas

[17] I say 'intention' here rather than 'desire' as standing in for botched action to reflect Williamson's updated (2017) structural analogies. In *Knowledge and its Limits* (2002), practical reason's working well was also understood in terms of acting on knowledge. This picture was meant to replace belief-desire psychology as the centre of intelligent life. However, the original (2002) version of the analogy maintained that belief stood to knowledge (on the mind-to-world side) as desire stood to action (on the world-to-mind side). The updated picture assimilates desire to belief – i.e., belief about what is good – e.g., (see, e.g., Lewis 1988; Price 1989) – and replaces 'desire' with 'intention' in the analogy. Thus, the updated picture holds that belief stands to knowledge as intention to action. For critical discussion, see Miracchi and Carter (2021).

[18] The language of 'direction of fit' is originally usually credited to Anscombe (1957), as a way of characterising a distinction between theoretical and practical intentional mental states. Theoretical mental states aim at representing things as they are (e.g., beliefs) and practical mental states aim at getting things done (i.e., desires). Realisation (i.e., success), for a cognitive (or theoretical) intentional mental state involves fitting mind-to-world; realisation for a practical mental state (e.g., desire, intention, etc.) involves fitting world-to-mind. A central 'lesson' direction-of-fit theorists (e.g., Smith 1994; D. Velleman 2000b; cf. Frost 2014) have taken from Anscombe's initial discussion is that intentional mental states are characterisable along the mind-to-world or world-to-mind faultline. However, the kinds of things to which direction of fit talk is applicable are not limited to intentional mental states. For example, according to Searle (1979), statements and predictions have a *word-to-world* direction of fit, whereas commands and promises have a *world-to-word* direction of fit. It's worth noting further that the very thought that things other than mental states can have directions of fit is actually perfectly compatible with Anscombe's initial idea, which is that what *makes* intentional states like beliefs and

the trustor aims not just to rely, but to *fit her reliance to the trustee's reciprocity*, the trustee (as such) aims to *fit her reciprocity to the trustor's reliance*.[19]

When the trustor attempts, but fails, to fit her reliance to reciprocity, what is residual is a kind of botched trust. When the trustee attempts, but fails, to fit her reciprocity to reliance, what is residual is a kind of botched reciprocity. (Compare with Williamsons's suggestion that belief is a kind of 'botched knowledge' and mere intention 'botched' action).

On this wider picture, then, in any two-way cooperative system, trust stands to apt trust as reciprocity to apt reciprocity (reciprocity that succeeds through trustworthiness) in a way that is broadly analogous to how (in a practical reasoning system, for Williamson) belief stands to knowledge (viz., *apt* belief[20]) as intention to action (i.e., *apt* intention[21]). And, further, just as practical reasoning's working well requires a match between not merely belief and intention but between knowledge and

---

desires have the directions of fit is that they have normative realisation conditions; beliefs *aim* (constitutively, not intentionally) at a certain kind of realisation, the same for desires. A similar normative reading is due to Platts (1980; for discussion see Frost 2014, 449–50). What this suggests, then, is that – at least in so far as we follow progenitors of DOF theory such as Anscombe and Searle, there is no barrier to viewing attempts more generally (including, e.g., trust and its reciprocation) with constitutive aims as admitting of directions of fit in so far as they have specifiable normative realisation conditions.

[19] I say 'reverse' direction of fit for ease of presentation, given that 'reliance-to-reciprocity' and 'reciprocity-to-reliance' are ostensibly reverse directions of fit. That said, it might have been more precise to describe the kind of direction of fit here as lining up even better with what Searle (1979) calls 'double direction of fit'. The reason here is that – in the unique case of cooperation – the realisation of one entails the realisation of the other.

[20] The idea that knowledge is type-identical with apt belief has advantages in epistemology; see Sosa (2007, 2010a), Greco (2010, Chs. 5-6) and Zagzebski (1996) for some notable defences of this position. Although I find this view plausible, the identification of knowledge with apt belief – while if fits snugly with the proposal developed here – isn't essential to it. For some criticism of the identification of knowledge with apt belief, see, e.g., Lackey (2007b), Pritchard (2007), Kelp (2013), Kornblith (2004), and Kallestrup and Pritchard (2014).

[21] For defences of the view that action is fruitfully understood as apt intention, see Sosa (2015, Ch. 1) and Miracchi and Carter (2021).

action; cooperation working well requires a match between not *mere* but *apt* trust and reciprocity.

The tables below represent these key analogies:

Table 9.1: Practical reasoning: realisations and attempts

Practical reasoning	**Fitting mind-to-world**	**Fitting world-to-mind**
Functioning well	knowledge (realisation) belief (attempt)	action (realisation) intention (attempt)

Table 9.2: Cooperation: realisations and attempts

Cooperation	**Fitting reliance-to-reciprocity**	**Fitting reciprocity-to-reliance**
Functioning well	trustor's apt trust (realisation) trust (attempt)	trustee's apt reciprocity (realisation) reciprocity (attempt)

One important feature of Williamson's 'knowledge-action' centric picture of practical reasoning is that it is meant to contrast with a competing picture (see, e.g., Humberstone 1992) that takes attempts – belief and desire – *rather than their realisations* as the core explanatory mental attitudes at the centre of intelligent life.[22] Attempts at knowledge and attempts at action retain a place in this picture, but it is their realisations,

[22] On the kind of view embraced by Humberstone (1992), it is also possible to accept the structural analogy on which belief stands to knowledge as desire to action. However, such a structural analogy would (on the belief-desire centred picture) begin with belief and desire as 'direction of fit mirror images', from which we would then 'solve upward' in the analogy to get the result that belief stands to knowledge as desire to action. Resisting this picture is the central argument in Williamson (2017), who suggests we begin with knowledge and action as direction of fit mirror images and then solve 'downward', filling in the relevant attempts. For a criticism of the role of 'mirrors' in both Williamson and Humberstone's approaches, see Miracchi and Carter (2021).

rather than the attempts themselves, that are of comparative theoretical importance.

The picture of cooperation suggested here likewise gives primacy to realisations over their attempts. That is, the present picture rejects the the trustee's performance (trust), a mere attempt at realisation by fitting reliance to reciprocity, and the trustee's disposition (trustworthiness) to fit reciprocity to reliance are the most theoretically important notions in a wider picture of cooperation. Rather, we should think of the importance of the trustor and trustee's matching achievements of trust and trustworthienss in cooperation as broadly analogous to the importance of action and knowledge (as opposed to mere belief and desire) in practical reasoning.

9.4 Symmetric evaluative normativity: trustor and trustee

In the good case where cooperation is working as it should, the trustor matches her achievement in trusting with the trustee's achievement in responding to trust. In both of these achievements (of apt trust and apt reciprocity) *competence* is manifested in success.

Cooperation doesn't always go so well. It falls short – at least to some extent – if we have anything short of achievement on either the trustor or trustee's side. In some cases, cooperation doesn't fall short by much, as when the trustor matches successful and competent but inapt (i.e., Gettiered) success to the trustee's achievement.[23] The trustor could do far worse. Successful but incompetent trust falls short of Gettiered trust on the trustor's side, as does competent but unsuccessful trust.[24] On the bot-

[23] Performances that are successful and competent but inapt have a 'Gettier' structure, where the success is disconnected from the good method used. For discussion, see Sosa (2007, Ch. 2, 2010a, 467, 474–5) and Greco (2009, 19–21, 2010, 73–76, 94–99). Cf., Pritchard (2012, 251, 264–8).

[24] The performance-theoretic analogy with virtue epistemology holds that successful but incompetent trust and competent but unsuccessful trust fall short of apt trust in a

tom rung on the trustor's side, we have trust that is neither competent nor successful, e.g., the betrayal of the gullible.

Likewise, on the trustee's side, falling just short of achievement is a kind of *Gettiered reciprocity*; suppose the trustee manifests her trustworthiness in assiduously entering the correct bank details online to pay back the loan she was entrusted to pay back, but succeeds only because a fortuitous electronic glitch (good luck) *corrects for an initial glitch* (bad luck) that would have diverted the funds to the wrong account.[25]

The trustee could do far worse. For one thing, she could have *not* manifested trustworthiness in responding to the trust placed in her, but succeeded just by luck. In such a case, suppose she *intends* to wire the money to the wrong account but only accidentally wires it to the right one.

Whereas the first loan case is a case of Gettiered reciprocity, the second is successful but incompetent reciprocity. Two remaining categories, lower down the rung on the trustee's side are: unsuccessful and competent reciprocity (i.e., exactly like the Gettiered reciprocity case *without* the second stroke of good luck), and – at the very bottom rung – incompetent and unsuccessful reciprocity (e.g., the trustee intends to wire money to the wrong account, and – failing in reciprocity – succeeds in betrayal.)

The above picture shows not only the many ways that cooperation can be defective (by less or greater degree) by matching anywhere from *just* less to *much* less than achievement on either the trustor's or trustee's side. But it also reveals an important *normative symmetry* on both sides.

By 'normative symmetry' what I mean is that the relevant *attempts* on each side (fitting reliance to reciprocity on the trustor's side, fitting reciprocity to reliance on the trustee's side) are such that we can evaluate each for the very same three things: (i) *success*; (ii) *competence*; and (iii) *apt-*

way that is analogous to how unjustified true beliefs and justified false beliefs both fall short of knowledge. See, for discussion, Sosa (2007, 2010a, 2015).

[25] For discussion of this kind of 'double luck' structure in relation to Gettier cases, see, e.g., Zagzebski (1994); see also Pritchard (2007) on what he calls 'intervening' veritic luck.

ness. And, moreover, it is specifically by *failing* to satisfy combinations of these norms that performances on the trustor and trustee's side fall short of achievement to whatever extent that they do.

The symmetrical picture of evaluative norms on each side is accordingly as follows:

	On the truster's side	**On the trustee's side**
Direction of fit	Reliance-to-reciprocity	Reciprocity-to-reliance
attempt	(trust) by means of reliance	reciprocity (by means of responding to trust)
success norm	*S*'s trusting *X* with ϕ is better if successfully reciprocated than if not; *S*'s trusting *X* with ϕ is successfully reciprocated iff takes care of ϕ as entrusted.	*X*'s reciprocating *S*'s trust with ϕ is better if *X* successfully reciprocates S's trust with ϕ than if not; *X* successfully reciprocates *S*'s trust with ϕ iff *X* takes care of ϕ as entrusted.
competence norm	*S*'s trusting *X* with ϕ is better if *S* trusts *X* with ϕ competently than if *S* does not.	*X*'s reciprocating *S*'s' trusting *X* with ϕ is better if *X* reciprocates *S*'s trust with ϕ competently than if *X* does not.
aptness norm	*S*'s trusting *X* with ϕ is better if *S* trusts *X* with ϕ aptly than if *S* does not.	*X*'s reciprocating *S*'s trusting *X* with ϕ is better if *X* reciprocates *S*'s trust with ϕ aptly than if *X* does not.

This symmetrical picture offers us a number of advantages. For one thing, our guiding idea that cooperation between trustor and trustee is working as it should when both sides match achievement to achievement can now be restated as an *aptness norm on cooperation*, one that is formulated *in terms of* trustor and trustee satisfying respective evaluative norms of aptness: a cooperative trust exchange E between trustor and trustee is better than it would be otherwise if E is apt; E is apt iff trustor and trustee satisfy their respective aptness norms.[26]

Secondly – and this bring us back to where we started – it should now be even more evident why focusing principally on a disposition (trustworthiness) on the trustee's side but not on the trustor's side (and vice versa for performance) is going to be arbitrary. From a wider view that takes in and evaluates the trust exchange in full, neither has any special status, even though both are essential to cooperation going well. They are, in a bit more detail, essential to cooperation going well in a way that is roughly analogous to how our beliefs and intentions (or: dispositions to form intentions) are important to practical reasoning going well. Both deserve attention, but should be appreciated as attempts *at* realisations, where the realisations of those attempts are what's needed in good practical reasoning as well as (*mutatis mutandis*) in good cooperation.

Thirdly, by transitioning to a symmetrical picture of the evaluative normativity of trust – with achievement matching achievement as the gold standard – we are better positioned to see the importance of questions that have been so far obscured. Perhaps most conspicuously here are questions about the competence norm of trust. After all, we have a grip on *apt trust* only by understanding competent trust, and this involves a clear view

[26] The idea that cooperation itself admits of an aptness norm suggests that cooperation is a kind of multi-agent performance itself. A natural way of thinking of this is as an irreducibly collective property of cooperators engaged jointly in a trust exchange. While I am sympathetic to this kind of gloss, I want to stress that we needn't be committed to it. The crux of the idea – viz., that cooperation is apt iff its individual cooperators perform aptly – is also compatible with a 'summativst' gloss, on which the cooperation has the relevant property (i.e., aptness) iff all its individual members have that property. For relevant recent discussion of these points, see Lackey (2021) and Broncano-Berrocal and Carter (2021). For a discussion of aptness as applicable to groups, see Kallestrup (2016).

of those dispositions of the trustor that lead them to trust *successfully* reliably enough. Other questions invited by the symmetrical picture involve the evaluative normativity of cooperation generally. Even if 'aptness on both sides' of the trustor/trustee divide implies that the cooperative exchange itself is apt, it remains an open question how to evaluate certain cooperation permutations that involve at least one norm violation on one side. For example: is cooperation working better if the trustor matches success without competence to the trustee's achievement or competence without success to the trustee's achievement?

Fourthly, given that *distrust* no less than trust can be successful, competent and apt, the normative symmetry we find on the trustor and trustee's side invites us to consider – analogically – what stands to distrust as as trust to reciprocity, and to consider how to best characterise parallel symmetrical norms that would regulate – symmetrically with successful, competent and apt distrust (on the side of the trustor) – also forbearance on the side of the trustee.

9.5 Concluding Remarks

The aim here has been to motivate and defend a new way of theorising about trust and trustworthiness – and their relationship to each other – by locating both within a broader picture that captures largely overlooked symmetries on both the trustor's and trustee's side of a cooperative exchange. The view I've defended here takes good cooperation as a theoretical starting point; on the view proposed, cooperation between trustor and trustee is working well when achievements in trust and responding to trust are matched on both sides of the trust exchange. In a bit more detail, the trustor 'matches' her achievement in trusting (an achievement in fitting reliance to reciprocity) with the trustee's achievement in responding to trust (an achievement in fitting reciprocity to reliance). From this starting point, we can then appreciate *symmetrical* ways that the trustor and trustee can (respectively) fall short, by violating what I've shown are symmetrical evaluative norms – of success, competence and aptness – that regulate the attempts made by both trustor and trustee. The overall pic-

ture was shown to have important advantages over the received way of theorising about how trust stands to trustworthiness, and it clears the way – by identifying key questions that have been obscured – to making further progress.

References

Adler, Jonathan E. 1994. "Testimony, Trust, Knowing." *The Journal of Philosophy* 91 (5): 264–75.

———. 1999. "The Ethics of Belief: Off the Wrong Track." *Midwest Studies in Philosophy* 23: 267–85.

Ahmed, Wasim, Josep Vidal-Alaball, Joseph Downing, and Francesc López Seguí. 2020. "COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data." *Journal of Medical Internet Research* 22 (5): 19–45.

Alfano, Mark. 2016. "The Topology of Communities of Trust." 2016. https://philarchive.org.

Alifirov, A. I., I. V. Mikhaylova, and A. S. Makhov. 2017. "Sport-Specific Diet Contribution to Mental Hygiene of Chess Player." *Theory and Practice of Physical Culture*, no. 4: 30–30.

Alston, William P. 1989. *Epistemic Justification: Essays in the Theory of Knowledge*. Cornell University Press.

———. 2006. *Beyond "Justification": Dimensions of Epistemic Evaluation*. Cornell University Press.

Alvarez, Maria. 2017. "Reasons for Action: Justification, Motivation, Explanation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/re

asons-just-vs-expl/.

———. 2009. "How Many Kinds of Reasons?" *Philosophical Explorations* 12 (2): 181–93. https://doi.org/10.1080/13869790902838514.

Anscombe, G. 1957. *Intention*. Harvard University Press.

Ariss, Sonny S. 2002. "Computer Monitoring: Benefits and Pitfalls Facing Management." *Information & Management* 39 (7): 553–58.

Ashraf, Nava, Iris Bohnet, and Nikita Piankov. 2006. "Decomposing Trust and Trustworthiness." *Experimental Economics* 9 (3): 193–208.

Baghramian, Maria, and Michel Croce. forthcoming. "Experts, Public Policy and the Question of Trust." In *Routledge Handbook of Political Epistemology*, edited by Michael Hannon and Jeroen De Ridder. London, UK: Routledge.

Baier, Annette. 1986. "Trust and Antitrust." *Ethics* 96 (2): 231–60.

Baker, Judith. 1987. "Trust and Rationality." *Pacific Philosophical Quarterly* 68 (1): 1–13.

Ballantyne, Nathan. 2012. "Luck and Interests." *Synthese* 185 (3): 319–34.

Bauman, Zygmunt. 2013. *Liquid Fear*. John Wiley & Sons.

Beck, Ulrich, and Brian Wynne. 1992. *Risk Society: Towards a New Modernity*. Vol. 17. sage.

Becker, Lawrence C. 1996. "Trust as Noncognitive Security About Motives." *Ethics* 107 (1): 43–61.

Beddor, Bob. 2020a. "New Work for Certainty." *Philosophers' Imprint* 20 (8).

———. 2020b. "Certainty in Action." *The Philosophical Quarterly* 70 (281): 711–37. https://doi.org/10.1093/pq/pqaa006.

Benton, Matthew A. 2014. "Knowledge Norms." *Internet Encyclopedia of Philosophy*.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10 (1): 122–42. https://doi.org/10.1006/game.1995.1027.

Bernecker, Sven. 2010. *Memory: A Philosophical Study*. Oxford University Press.

Björklund, Fredrik, Gunnar Björnsson, John Eriksson, Ragnar Francén Olinder, and Caj Strandberg. 2012. "Recent Work on Motivational Internalism." *Analysis* 72 (1): 124–37. https://doi.org/10.1093/analys/anr118.

Bond Jr, Charles F., and Bella M. DePaulo. 2006. "Accuracy of Deception Judgments." *Personality and Social Psychology Review* 10 (3): 214–34.

Bondy, Patrick, and J. Adam Carter. 2019. "Well-Founded Belief: An Introduction."

BonJour, Laurence. 1980. "Externalist Theories of Empirical Knowledge." *Midwest Studies in Philosophy* 5: 53–73.

Booth, Anthony Robert. 2007. "The Two Faces of Evidentialism." *Erkenntnis* 67 (3): 401–17.

Bovens, Luc. 1999. "The Value of Hope." *Philosophy and Phenomenological Research* 59 (3): 667–81.

Bradford, Gwen. 2013. "The Value of Achievements." *Pacific Philosophical Quarterly* 94 (2): 204–24.

———. 2015a. *Achievement*. Oxford University Press, USA.

———. 2015b. "Knowledge, Achievement, and Manifestation." *Erkenntnis* 80 (1): 97–116. https://doi.org/10.1007/s10670-014-9614-0.

Bratman, Michael E. 1992. "Practical Reasoning and Acceptance in a Context." *Mind* 101 (401): 1–15.

Braynov, Sviatoslav, and Tuomas Sandholm. 2002. "Contracting with Uncertain Level of Trust." *Computational Intelligence* 18 (4): 501–14. https://doi.org/10.1111/1467-8640.00200.

Broncano-Berrocal, Fernando, and J. Adam Carter. 2021. *The Philosophy of Group Polarization: Epistemology, Metaphysics, Psychology*. Routledge.

Buckareff, Andrei. 2010. "Acceptance Does Not Entail Belief." *International Journal of Philosophical Studies* 18 (2): 255–61.

Burge, Tyler. 1998. "Reason and the First Person." *Knowing Our Own Minds*, 243–70.

———. 2003. "Perceptual Entitlement." *Philosophy and Phenomenological Research* 67 (3): 503–48.

Burns, Calvin, Kathryn Mearns, and Peter McGeorge. 2006. "Explicit and Implicit Trust Within Safety Culture." *Risk Analysis* 26 (5): 1139–50.

Calhoun, Cheshire. 1984. "Cognitive Emotions?" *What Is an Emotion*, 327–42.

Carter, J. Adam. 2011. "Kvanvig on Pointless Truths and the Cognitive Ideal." *Acta Analytica* 26 (3): 285–93.

———. 2016a. "Robust Virtue Epistemology as Anti-Luck Epistemology: A New Solution." *Pacific Philosophical Quarterly* 97 (1): 140–55.

———. 2018. "On Behalf of Controversial View Agnosticism." *European Journal of Philosophy* 26 (4): 1358–70. https://doi.org/10.1111/ejop.12333.

———. 2019. "Exercising Abilities." *Synthese*, 1–15.

———. 2020a. "De Minimis Normativism: A New Theory of Full Aptness." *The Philosophical Quarterly*.

———. 2022. "Trust as Performance." *Philosophical Issues: A Supple-*

ment to Noûs.

———. 2016b. "Sosa on Knowledge, Judgment and Guessing." *Synthese*, August.

———. 2020b. "On Behalf of a Bi-Level Account of Trust." *Philosophical Studies* 177 (8): 2299–2322. https://doi.org/10.1007/s11098-019-01311-2.

Carter, J. Adam, Benjamin W. Jarvis, and Katherine Rubin. 2015. "Varieties of Cognitive Achievement." *Philosophical Studies* 172 (6): 1603–23.

Carter, J. Adam, Duncan Pritchard, and John Turri. 2018. "The Value of Knowledge." In *Stanford Encyclopedia of Philosophy*.

Carter, J. Adam, and Mona Simion. 2020. "The Ethics and Epistemology of Trust." *Internet Encyclopedia of Philosophy*.

Carver, Charles S., Michael F. Scheier, and Suzanne C. Segerstrom. 2010. "Optimism." *Clinical Psychology Review* 30 (7): 879–89.

Chan, Timothy, ed. 2013. *The Aim of Belief*. Oxford University Press.

Chrisman, Matthew. 2012. "The Normative Evaluation of Belief and the Aspectual Classification of Belief and Knowledge Attributions." *The Journal of Philosophy* 109 (10): 588–612.

Cogley, Zac. 2012. "Trust and the Trickster Problem."

Cohen, L. Jonathan. 1989. "Belief and Acceptance." *Mind* 98 (391): 367–89.

Coleman, James S. 1990. "Relations of Trust." *Foundations of Social Theory, Cambridge, London*, 91–116.

Comesaña, Juan. 2005. "Unsafe Knowledge." *Synthese* 146 (3): 395–404.

Cook, Karen S., Russell Hardin, and Margaret Levi. 2005. *Cooperation*

Without Trust? Russell Sage Foundation.

Cottingham, John. 2002. "Descartes and the Voluntariness of Belief." *The Monist* 85 (3): 343–60.

Crisp, Roger. 2014. "II—Roger Crisp: Moral Testimony Pessimism: A Defence." *Aristotelian Society Supplementary Volume* 88 (1): 129–43. https://doi.org/10.1111/j.1467-8349.2014.00236.x.

Dancy, Jonathan. 2004. *Ethics Without Principles*. Oxford University Press on Demand.

Dasgupta, Partha. 1988. "Trust as a Commodity." In *Trust: Making and Breaking Cooperative Relations*, edited by Diego Gambetta, 49–72. Blackwell.

D'cruz, Jason. 2015. "Trust, Trustworthiness, and the Moral Consequence of Consistency." *Journal of the American Philosophical Association* 1 (3): 467–84.

Díaz, Rodrigo, and Manuel Almagro. 2019. "You Are Just Being Emotional! Testimonial Injustice and Folk-Psychological Attributions." *Synthese*, 1–22.

Domenicucci, Jacopo, and Richard Holton. 2017. "Trust as a Two-Place Relation." *The Philosophy of Trust*, 149–60.

Dormandy, Katherine. 2020. "Exploitative Epistemic Trust." In *Trust in Epistemology*, edited by Katherine Dormandy, 241–64.

Dougherty, Tom. 2013. "Sex, Lies, and Consent." *Ethics* 123 (4): 717–44.

———. 2019. "Consent, Communication, and Abandonment." *Law and Philosophy* 38 (4): 387–405. https://doi.org/10.1007/s10982-019-09355-5.

Echeverri, Santiago. 2020. "Guarantee and Reflexivity." *Journal of Philosophy* 117 (9): 473–500.

Elster, Jon. 2015. *Explaining Social Behavior*. Cambridge University Press.

Emmons, Robert A., and Michael E. McCullough. 2004. *The Psychology of Gratitude*. Oxford University Press.

Engel, Mylan. 1992. "Is Epistemic Luck Compatible with Knowledge?" *The Southern Journal of Philosophy* 30 (2): 59–75.

Enoch, David. 2014. "A Defense of Moral Deference." *The Journal of Philosophy* 111 (5): 229–58.

Fantl, Jeremy, and Matthew McGrath. 2002. "Evidence, Pragmatics, and Justification." *The Philosophical Review* 111 (1): 67–94.

Faulkner, Paul. 2007. "A Genealogy of Trust." *Episteme: A Journal of Social Epistemology* 4 (3): 305–21.

———. 2010. "Norms of Trust." In *Social Epistemology*, edited by Adrian Haddock, Alan Millar, and Duncan Pritchard. Oxford University Press.

———. 2011. *Knowledge on Trust*. Oxford University Press.

———. 2014. "The Moral Obligations of Trust." *Philosophical Explorations* 17 (3): 332–45.

———. 2015. "The Attitude of Trust Is Basic." *Analysis* 75 (3): 424–29. https://doi.org/10.1093/analys/anv037.

Faulkner, Paul, and Thomas Simpson. 2017. *The Philosophy of Trust*. Oxford University Press.

Feldman, Richard. 2000. "The Ethics of Belief." *Philosophy and Phenomenological Research* 60 (3): 667–95.

Finlay, Stephen, and Mark Schroeder. 2017. "Reasons for Action: Internal Vs. External." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/rea

sons-internal-external/.

Fischer, John Martin, and Mark Ravizza. 2000. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge university press.

Frankfurt, Harry G. 1969. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66 (23): 829–39.

Frederiksen, Morten. 2014. "Trust in the Face of Uncertainty: A Qualitative Study of Intersubjective Trust and Risk." *International Review of Sociology* 24: 130–44.

Fricker, Elizabeth. 2018. "Trust and Testimonial Justification." *Manuscript*.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Fridland, Ellen, and Carlotta Pavese. forthcoming. *Routledge Handbook of Philosophy of Skill and Expertise*. Routledge.

Friedman, Jane. 2013. "Suspended Judgment." *Philosophical Studies* 162 (2): 165–81.

———. 2020. "The Epistemic and the Zetetic." *Philosophical Review* 129 (4): 501–36.

Frost, Kim. 2014. "On the Very Idea of Direction of Fit." *The Philosophical Review* 123 (4): 429–84. https://doi.org/10.1215/00318108-2749720.

Frost-Arnold, Karen. 2014. "The Cognitive Attitude of Rational Trust." *Synthese* 191 (9). https://doi.org/10.1007/s11229-012-0151-6.

Gambetta, Diego. 1988. *Trust: Making and Breaking Cooperative Relations*. Blackwell.

Geach, Peter T. 1956. "Good and Evil." *Analysis* 17 (2): 33–42.

Gibbons, John. 2013. *The Norm of Belief*. Oxford University Press.

Giddens, Anthony. 2013. *The Consequences of Modernity*. John Wiley & Sons.

Glüer, Kathrin, and \AAsa Wikforss. 2009. "Against Content Normativity." *Mind* 118 (469): 31–70. https://doi.org/10.1093/mind/fzn154.

Goldberg, Sanford. 2015. "What Is the Subject-Matter of the Theory of Epistemic Justification?" *Epistemic Evaluation: Purposeful Epistemology*, 205–23.

Graham, Peter J. 2016. "Testimonial Knowledge: A Unified Account." *Philosophical Issues* 26 (1): 172–86.

Greco, John. 2009. "Knowledge and Success from Ability." *Philosophical Studies* 142 (1): 17–26.

———. 2010. *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge University Press.

———. 2013. "Knowledge, Testimony, and Action." In *Knowledge, Virtue, and Action*, edited by T. Henning and D. Schweikard, 15–29. London: Routledge.

———. 2014. "Episteme: Knowledge and Understanding." *Virtues and Their Vices*, 285–302.

———. 2019. "4 the Role of Trust in Testimonial Knowledge." *Trust in Epistemology*.

———. 2020a. *The Transmission of Knowledge*. Cambridge University Press.

———. 2020b. "The Transmission of Knowledge and Garbage." *Synthese* 197 (7): 2867–78.

Guo, Guibing, Jie Zhang, Daniel Thalmann, Anirban Basu, and Neil Yorke-Smith. 2014. "From Ratings to Trust: An Empirical Study of Implicit Trust in Recommender Systems." In *Proceedings of the 29th*

Annual Acm Symposium on Applied Computing, 248–53.

Hall, Mark A. 2005. "The Importance of Trust for Ethics, Law, and Public Policy." *Cambridge Q. Healthcare Ethics* 14: 156.

Hansson, Sven Ove. 2004. "Philosophical Perspectives on Risk." *Techné: Research in Philosophy and Technology* 8 (1): 10–35.

———. 2018. "Risk." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018. Metaphysics Research Lab, Stanford University.

Hardin, Russell. 1992. "The Street-Level Epistemology of Trust." *Analyse & Kritik* 14 (2): 152–76.

———. 1996. "Trustworthiness." *Ethics* 107 (1): 26–42.

———. 2002. *Trust and Trustworthiness*. Russell Sage Foundation.

Hardwig, John. 1991. "The Role of Trust in Knowledge." *The Journal of Philosophy* 88 (12): 693–708.

Hawley, Katherine. 2014. "Trust, Distrust and Commitment." *Noûs* 48 (1): 1–20.

———. 2019. *How to Be Trustworthy*. Oxford University Press, USA.

Hawthorne, John, and Jason Stanley. 2008. "Knowledge and Action." *The Journal of Philosophy* 105 (10): 571–90.

Heil, John. 1983. "Doxastic Agency." *Philosophical Studies* 43 (3): 355–64.

Hetherington, Stephen. 2013. "Knowledge Can Be Lucky." *Contemporary Debates in Epistemology* 2: 164–76.

Hieronymi, Pamela. 2008. "The Reasons of Trust." *Australasian Journal of Philosophy* 86 (2): 213–36.

Hills, Alison. 2009. "Moral Testimony and Moral Epistemology." *Ethics* 120 (1): 94–127. https://doi.org/10.1086/648610.

Hinchman, Edward. 2017. "On the Risks of Resting Assured: An Assurance Theory of Trust." In *The Philosophy of Trust*, edited by Paul Faulkner and Thomas W. Simpson. Oxford: Oxford University Press.

———. 2020. "Trust and Will." In *Routledge Handbook on Trust and Philosophy*, edited by Judith Simon. New York: Routledge.

Holroyd, Jules, Robin Scaife, and Tom Stafford. 2017. "What Is Implicit Bias?" *Philosophy Compass* 12 (10): e12437.

Holton, Richard. 1994. "Deciding to Trust, Coming to Believe." *Australasian Journal of Philosophy* 72 (1): 63–76.

Honoré, Anthony M. 1964. "Can and Can't." *Mind* 73 (292): 463–79.

Hooker, Brad. 2002. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford University Press.

Horsburgh, H. J. N. 1960. "The Ethics of Trust." *The Philosophical Quarterly* 10 (41): 343–54.

Humberstone, I. Lloyd. 1992. "Direction of Fit." *Mind* 101 (401): 59–83.

Hume, David. (1739) 2003. *A Treatise of Human Nature*. Courier Corporation.

Jarvis, Benjamin. 2013. "Knowledge, Cognitive Achievement, and Environmental Luck." *Pacific Philosophical Quarterly* 94 (4): 529–51.

Jones, Karen. 1996. "Trust as an Affective Attitude." *Ethics* 107 (1): 4–25.

———. 2004. "Trust and Terror." In *Moral Psychology: Feminist Ethics and Social Theory*, edited by Peggy DesAutels and Margaret Urban Walker, 3–18. Rowman & Littlefield.

———. 2012. "Trustworthiness." *Ethics* 123 (1): 61–85.

Kallestrup, Jesper. 2016. "Group Virtue Epistemology." *Synthese*, 1–19.

Kallestrup, Jesper, and Duncan Pritchard. 2014. "Virtue Epistemology and Epistemic Twin Earth." *European Journal of Philosophy* 22 (3): 335–57.

Kaplan, David. 1979. "On the Logic of Demonstratives." *Journal of Philosophical Logic* 8 (1): 81–98.

Kelp, Christoph. 2013. "Knowledge: The Safe-Apt View." *Australasian Journal of Philosophy* 91 (2): 265–78.

———. 2018. "Assertion: A Function First Account." *Noûs* 52 (2): 411–42.

———. 2020a. "The Epistemology of Ernest Sosa: An Introduction." *Synthese*, 1–8.

———. 2020b. "Theory of Inquiry." *Philosophy and Phenomenological Research*.

———. Forthcoming. *Inquiry, Knowledge, and Understanding*. Oxford: Oxford University Press.

Kelp, Christoph, Cameron Boult, Fernando Broncano-Berrocal, Paul Dimmock, Harmen Ghijsen, and Mona Simion. 2017. "Hoops and Barns: A New Dilemma for Sosa." *Synthese*, 1–16.

Kelp, Christoph, and Mona Simion. 2021. "What Is Trustworthiness?" *Manuscript*.

Kenny, A. J. P. 1976. *Will, Freedom and Power*. Blackwell.

Keren, Arnon. 2014. "Trust and Belief: A Preemptive Reasons Account." *Synthese* 191 (12): 2593–2615.

———. 2019. "Trust, Preemption, and Knowledge." *Trust in Epistemology*.

———. 2020. "Trust and Belief." In *The Routledge Handbook of Trust and Philosophy*, edited by Judith Simon, 109–20.

Kirton, Andrew. forthcoming. "Matters of Trust as Matters of Attachment Security." *International Journal of Philosophical Studies*, 1–20.

Korcz, Keith Allen. 2019. "The Epistemic Basing Relation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/basing-epistemic/.

Kornblith, Hilary. 1983. "Justified Belief and Epistemically Responsible Action." *The Philosophical Review* 92 (1): 33–48.

———. 2004. "Sosa on Human and Animal Knowledge." *Ernest Sosa and His Critics*, 126–34.

Kramer, Roderick M. 1999. "Trust and Distrust in Organizations: Emerging Perspectives, Enduring Questions." *Annual Review of Psychology* 50 (1): 569–98.

Kraut, Robert. 1980. "Humans as Lie Detectors." *Journal of Communication* 30 (4): 209–18.

Krishnan, Rekha, Xavier Martin, and Niels G. Noorderhaven. 2006. "When Does Trust Matter to Alliance Performance?" *The Academy of Management Journal* 49 (5): 894–917.

Kvanvig, Jonathan. 2008. "Pointless Truth." *Midwest Studies in Philosophy* 32 (1): 199–212.

Kvanvig, Jonathan L. 2016. "The Idea of Faith as Trust." *Reason and Faith: Themes from Richard Swinburne*, 4–26.

Lackey, Jennifer. 2007a. "Norms of Assertion." *Noûs* 41 (4): 594–626.

———. 2007b. "Why We Don't Deserve Credit for Everything We Know." *Synthese* 158 (3): 345–61.

———. 2021. *The Epistemology of Groups*. Oxford University Press, USA.

Lagerspetz, Olli. 1998. *Trust: The Tacit Demand*. Vol. 1. Springer Sci-

ence & Business Media.

Lahno, Bernd. 2001. "On the Emotional Character of Trust." *Ethical Theory and Moral Practice* 4 (2): 171–89. https://doi.org/10.1023/A:1011425102875.

———. 2004. "Three Aspects of Interpersonal Trust." *Analyse & Kritik* 26 (1): 30–47.

Lewis, David. 1988. "Desire as Belief." *Mind* 97 (387): 323–32.

Littlejohn, Clayton. 2014. "Fake Barns and False Dilemmas."

Luper-Foy, Steven. 1984. "The Epistemic Predicament: Knowledge, Nozickian Tracking, and Scepticism." *Australasian Journal of Philosophy* 62 (1): 26–49.

Madison, Brent JC. 2011. "Combating Anti Anti-Luck Epistemology." *Australasian Journal of Philosophy* 89 (1): 47–58.

Marušić, Berislav. 2017. "Trust, Reliance and the Participant Stance." *Philosopher's Imprint* 17 (17). http://hdl.handle.net/2027/spo.3521354.0017.017.

McGeer, Victoria. 2004. "The Art of Good Hope." *The Annals of the American Academy of Political and Social Science* 592 (1): 100–127.

———. 2008. "Trust, Hope and Empowerment." *Australasian Journal of Philosophy* 86 (2): 237–54. https://doi.org/10.1080/00048400801886413.

McGrath, Sarah. 2011. "Skepticism About Moral Expertise as a Puzzle for Moral Realism." The Journal of Philosophy. August 10, 2011. https://doi.org/10.5840/jphil201110837.

McHugh, Conor. 2012. "The Truth Norm of Belief." *Pacific Philosophical Quarterly* 93 (1): 8–30.

McLeod, Carolyn. 2002. *Self-Trust and Reproductive Autonomy*. MIT Press.

———. 2020. "Trust." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2020. Metaphysics Research Lab, Stanford University.

McMyler, Benjamin. 2011. *Testimony, Trust, and Authority*. OUP USA.

McNarry, Gareth, Jacquelyn Allen-Collinson, and Adam B. Evans. 2020. "'You Always Wanna Be Sore, Because Then You Are Seeing Results': Exploring Positive Pain in Competitive Swimming." *Sociology of Sport Journal* 1 (aop): 1–9.

Mele, Alfred R. 2001. *Autonomous Agents: From Self-Control to Autonomy*. Oxford University Press on Demand.

———. 2003a. "Agents' Abilities." *Noûs* 37 (3): 447–70. https://doi.org/10.1111/1468-0068.00446.

———. 2003b. *Motivation and Agency*. Oxford University Press.

Millar, Alan. 2009. "What Is It That Cognitive Abilities Are Abilities to Do?" *Acta Analytica* 24 (4): 223.

Miracchi, Lisa. 2015. "Knowledge Is All You Need." *Philosophical Issues* 25 (1): 353–78.

Miracchi, Lisa, and J. Adam Carter. 2021. "Refitting the Mirrors: On Structural Analogies in Epistemology and Action Theory." *Manuscript*.

Mollering, Guido. 2006. *Trust: Reason, Routine, Reflexivity*. Emerald Group Publishing.

Möllering, Guido. 2001. "The Nature of Trust: From Georg Simmel to a Theory of Expectation, Interpretation and Suspension." *Sociology* 35 (2): 403–20.

Mumford, Stephen. 2016. "Dispositions." In *Routledge Encyclopedia of Philosophy*, 1st ed. London: Routledge. https://doi.org/10.4324/9780415249126-N116-1.

Mumpower, Jeryl. 1986. "An Analysis of the de Minimis Strategy for Risk Management." *Risk Analysis* 6 (4): 437–46.

Munroe, Wade. 2016. "Testimonial Injustice and Prescriptive Credibility Deficits." *Canadian Journal of Philosophy* 46 (6): 924–47.

Neta, Ram. 2019. "The Basing Relation." *The Philosophical Review* 128 (2): 179–217. https://doi.org/10.1215/00318108-7374945.

Nickel, Philip J. 2015. "Design for the Value of Trust." *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 551–67.

Nickel, Philip J., and Krist Vaesen. 2012. "Risk and Trust." In *Handbook of Risk Theory*, edited by Sabine Roeser, Rafaela Hillerbrand, Martin Peterson, and Per Sandin. Springer.

Niebuhr, Helmut Richard. 1991. *Faith on Earth: An Inquiry into the Structure of Human Faith*. Yale University Press.

Niker, Fay, and Laura Specker Sullivan. 2018. "Trusting Relationships and the Ethics of Interpersonal Action." *International Journal of Philosophical Studies* 26 (2): 173–86. https://doi.org/10.1080/09672559.2018.1450081.

O'Neil, Collin. 2017. "Betraying Trust." In *The Philosophy of Trust*, edited by Paul Faulkner and Thomas W. Simpson, 70–89. Oxford, UK: Oxford University Press.

O'neil, Collin. 2012. "Lying, Trust, and Gratitude." *Philosophy & Public Affairs* 40 (4): 301–33.

O'Neill, Onora. 2018. "Linking Trust to Trustworthiness." *International Journal of Philosophical Studies* 26 (2): 293–300. https://doi.org/10.1080/09672559.2018.1454637.

Papineau, David. 2013. "There Are No Norms of Belief." In *The Aim of Belief*, edited by Timothy Chan. Oxford University Press.

Pavese, Carlotta. 2016. "Skill in Epistemology II: Skill and Know How." *Philosophy Compass* 11 (11): 650–60. https://doi.org/10.1111/phc3.12364.

Peterson, Martin. 2002. "What Is a de Minimis Risk?" *Risk Management* 4 (2): 47–55.

Pettit, Philip. 1995. "The Cunning of Trust." *Philosophy and Public Affairs* 24 (3): 202–25.

Plato. (385BC) 2011. *Plato's Meno*. Edited by Richard Stanley Bluck. Cambridge University Press.

Platts, Mark. 1980. "Ways of Meaning." *Mind* 89 (355): 454–56.

Potter, Nancy Nyquist. 2002. *How Can I Be Trusted?: A Virtue Theory of Trustworthiness*. Rowman & Littlefield.

———. 2020. "Interpersonal Trust." In *The Routledge Handbook of Trust and Philosophy*, edited by Judith Simon, 243–55. Routledge.

Price, Huw. 1989. "Defending Desire-as-Belief." *Mind* 98 (389): 119–27. https://www.jstor.org/stable/2255064.

Pritchard, Duncan. 2005. *Epistemic Luck*. Clarendon Press.

———. 2007. "Anti-Luck Epistemology." *Synthese* 158 (3): 277–97.

———. 2009a. "Knowledge, Understanding and Epistemic Value." *Royal Institute of Philosophy Supplement* 64: 19–43. https://doi.org/10.1017/s1358246109000046.

———. 2009b. "The Value of Knowledge." *The Harvard Review of Philosophy* 16 (1): 86–103. https://doi.org/10.5840/harvardreview20091616.

———. 2012. "Anti-Luck Virtue Epistemology." *The Journal of Philosophy* 109 (3): 247–79.

———. 2015. "Risk." *Metaphilosophy* 46 (3): 436–61.

———. 2016. "Epistemic Risk." *The Journal of Philosophy* 113 (11): 550–71.

Rabinowitz, Dani. 2011. "The Safety Condition for Knowledge." *Internet Encyclopedia of Philosophy*.

Railton, Peter. 2014. "Reliance, Trust, and Belief." *Inquiry* 57 (1): 122–50.

Rawls, John. 1955. "Two Concepts of Rules." *The Philosophical Review* 64 (1): 3–32.

Raz, Joseph. 2009. "Reasons: Explanatory and Normative." In *New Essays on the Explanation of Action*, edited by Constantine Sandis, 184–202. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230582972_11.

Rhodes, Rosamond, Jody Azzouni, Stefan Bernard Baumrin, Keith Benkov, Martin J. Blaser, Barbara Brenner, Joseph W. Dauben, William J. Earle, Lily Frank, and Nada Gligorov. 2011. *De Minimis Risk: A Proposal for a New Category of Research Risk*. Taylor & Francis.

Rorty, Amélie Oksenberg. 1980. *Explaining Emotions*. Vol. 232. Univ of California Press.

Rosati, Connie S. 2016. "Moral Motivation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/moral-motivation/.

Rulis, Alan M. 1986. "De Minimis and the Threshold of Regulation." *Food Protection Technology*, 29–37.

Sandin, Per. 2005. "Naturalness and de Minimis Risk." *Environmental Ethics* 27 (2): 191–200.

Saul, Jennifer. 2013. "Scepticism and Implicit Bias." *Disputatio* 5 (37): 243–63.

———. 2017. "Implicit Bias, Stereotype Threat, and Epistemic Injustice." *The Routledge Handbook of Epistemic Injustice*, 235–42.

Schechter, Joshua. 2019. "Aiming at Aptness." *Episteme* 16 (4): 438–52.

Schroeder, Mark. 2011. "Ought, Agents, and Actions." *The Philosophical Review* 120 (1): 1–41. https://doi.org/10.1215/00318108-2010-017.

Scott-Kakures, Dion. 1994. "On Belief and the Captivity of the Will." *Philosophy and Phenomenological Research* 54 (1): 77–103.

Searle, J. R. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts Cambridge University Press*. Cambridge.

Shafer-Landau, Russ. 2003. *Moral Realism: A Defence*. Oxford University Press on Demand.

Shah, Nishi. 2003. "How Truth Governs Belief." *The Philosophical Review* 112 (4): 447–82.

———. 2008. "How Action Governs Intention." *Philosophers' Imprint* 8: 1–19.

Shah, Nishi, and J. David Velleman. 2005. "Doxastic Deliberation." *Philosophical Review* 114 (4): 497–534. https://doi.org/10.1215/00318108-114-4-497.

Shoemaker, Sydney. 1995. "Moore's Paradox and Self-Knowledge." *Philosophical Studies* 77 (2-3): 211–28.

Silva, Paul. 2015. "On Doxastic Justification and Properly Basing One's Beliefs." *Erkenntnis* 80 (5): 945–55.

Simion, Mona. 2019. "Knowledge-First Functionalism." *Philosophical Issues* 29 (1): 254–67.

Simion, Mona, and Christoph Kelp. 2018. "How to Be an Anti-Reductionist." *Synthese*, 1–18.

Simion, Mona, Christoph Kelp, and Harmen Ghijsen. 2016. "Norms of Belief." *Philosophical Issues* 26 (1): 374–92.

Simpson, Thomas W. 2012. "What Is Trust?" *Pacific Philosophical Quarterly* 93 (4): 550–69.

Sjöberg, Lennart. 2000. "The Methodology of Risk Perception Research." *Quality and Quantity* 34 (4): 407–18.

Sjöberg, Lennart, Bjørg-Elin Moen, and Torbjørn Rundmo. 2004. "Explaining Risk Perception." *An Evaluation of the Psychometric Paradigm in Risk Perception Research* 10 (2): 665–12.

Slovic, Paul. 1987. "Perception of Risk." *Science* 236 (4799): 280–85.

———. 1988. "Risk Perception." In *Carcinogen Risk Assessment*, 171–81. Springer.

Smith, Michael. 1994. *The Moral Problem*. Blackwell.

Snyder, Charles R., Cheri Harris, John R. Anderson, Sharon A. Holleran, Lori M. Irving, Sandra T. Sigmon, Lauren Yoshinobu, June Gibb, Charyle Langelle, and Pat Harney. 1991. "The Will and the Ways: Development and Validation of an Individual-Differences Measure of Hope." *Journal of Personality and Social Psychology* 60 (4): 570.

Snyder, C. Rick. 1995. "Conceptualizing, Measuring, and Nurturing Hope." *Journal of Counseling & Development* 73 (3): 355–60.

Solomon, Robert C., and Fernando Flores. 2003. *Building Trust: In Business, Politics, Relationships, and Life*. Oxford University Press.

Sosa, Ernest. 1999. "How to Defeat Opposition to Moore." *Noûs* 33 (s13): 141–53. https://doi.org/10.1111/0029-4624.33.s13.7.

———. 2007. *A Virtue Epistemology: Apt Belief and Reflective Knowledge, Volume I*. Oxford University Press.

———. 2009. *Reflective Knowledge: Apt Belief and Reflective Knowledge, Volume II*. Oxford University Press.

———. 2010a. "How Competence Matters in Epistemology." *Philosophical Perspectives* 24 (1): 465–75.

———. 2010b. "Value Matters in Epistemology." *The Journal of Philosophy* 107 (4): 167–90.

———. 2015. *Judgment & Agency*. Oxford University Press UK.

———. 2017. *Epistemology*. Princeton: Princeton University Press.

———. 2019. "Animal Versus Reflective Orders of Epistemic Competence." In *Thinking About Oneself: The Place and Value of Reflection in Philosophy and Psychology*, edited by Waldomiro J. Silva-Filho and Luca Tateo, 21–32. Philosophical Studies Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-18266-3_2.

———. 2020. "Default Assumptions and Pure Thought." *Manuscript*.

———. 2021. *Epistemic Explanations: A Theory of Telic Normativity, and What It Explains*. Oxford University Press.

Sousa, Ronald de. 1987. *The Rationality of Emotion*. MIT Press.

Stanley, Jason. 2005. *Knowledge and Practical Interests*. Oxford University Press.

———. 2008. "Knowledge and Certainty." *Philosophical Issues* 18: 35–57.

Stanley, Jason, and Timothy Williamson. 2017. "Skill."

Starr, William B. 2020. "A Preference Semantics for Imperatives." *Semantics and Pragmatics* 13: 6.

Steglich-Petersen, Asbjørn. 2006. "No Norm Needed: On the Aim of Belief." *The Philosophical Quarterly* 56 (225): 499–516.

Stocker, Michael. 1979. "Desiring the Bad: An Essay in Moral Psychology." *The Journal of Philosophy* 76 (12): 738–53.

Sutton, Jonathan. 2007. *Without Justification*. MIT press.

Treanor, Nick. 2014. "Trivial Truths and the Aim of Inquiry." *Philosophy and Phenomenological Research* 89 (3): 552–59.

Tsai, George. 2018. "The Virtue of Being Supportive." *Pacific Philosophical Quarterly* 99 (2): 317–42.

Turri, John. 2010. "On the Relationship Between Propositional and Doxastic Justification." *Philosophy and Phenomenological Research* 80 (2): 312–26.

———. 2012. "A Puzzle About Withholding." *The Philosophical Quarterly* 62 (247): 355–64. https://doi.org/10.1111/j.1467-9213.2011.00043.x.

———. 2017. "Sustaining Rules: A Model and Application." In *Knowledge First: Approaches in Epistemology and Mind*, edited by J. Adam Carter, Emma C. Gordon, and Benjamin W. Jarvis, 259–77. Oxford: Oxford University Press.

———. 2011. "Manifest Failure: The Gettier Problem Solved." *Philosopher's Imprint* 11 (8). http://hdl.handle.net/2027/spo.3521354.0011.008.

Turri, John, Mark Alfano, and John Greco. 2019. "Virtue Epistemology." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/epistemology-virtue/.

Tversky, Amos, and Daniel Kahneman. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5 (2): 207–32.

Unger, Peter. 1975. *Ignorance: A Case for Scepticism*. Oxford University Press.

Unkelbach, Christian, Joseph P. Forgas, and Thomas F. Denson. 2008. "The Turban Effect: The Influence of Muslim Headgear and Induced

Affect on Aggressive Responses in the Shooter Bias Paradigm." *Journal of Experimental Social Psychology* 44 (5): 1409–13.

Vargas, Miguel Ángel Fernández. 2016. *Performance Epistemology: Foundations and Applications*. Oxford University Press.

Vaughn, Leigh Ann. 1999. "Effects of Uncertainty on Use of the Availability of Heuristic for Self-Efficacy Judgments." *European Journal of Social Psychology* 29 (2-3): 407–10.

Velleman, David. 2000a. "On the Aim of Belief." *The Possibility of Practical Reason* 244.

———. 2000b. *The Possibility of Practical Reason*. Oxford University Press.

Verkuyl, Henk J. 1989. "Aspectual Classes and Aspectual Composition." *Linguistics and Philosophy* 12 (1): 39–94.

Vitz, Rico. 2008. "Doxastic Voluntarism." In *Internet Encyclopedia of Philosophy*, edited by Unknown Unknown.

———. 2010. "Descartes and the Question of Direct Doxastic Voluntarism." 2010. https://philpapers.org/rec/VITDAT.

Vrij, Aldert. 2000. *Detecting Lies and Deceit: The Psychology of Lying and Implications for Professional Practice*. Wiley.

Wanderer, Jeremy, and Leo Townsend. 2013. "Is It Rational to Trust?" *Philosophy Compass* 8 (1): 1–14.

Watson, Sean, and Anthony Moran. 2005. *Trust, Risk, and Uncertainty*. Palgrave-Macmillan.

Wedgwood, Ralph. 2002. "The Aim of Belief." *Philosophical Perspectives* 16: 267–97. https://doi.org/10.1111/1468-0068.36.s16.10.

Whipple, Chris. 2012. *De Minimis Risk*. Vol. 2. Springer Science & Business Media.

Whiting, Daniel. 2013a. "Nothing but the Truth: On the Norms and Aims of Belief." In *The Aim of Belief*, edited by Timothy Chan. Oxford University Press.

———. 2013b. "Truth: The Aim and Norm of Belief." *Teorema: Revista Internacional de Filosofía*, 121–35.

Williams, Bernard. 1970. "Deciding to Believe." In *Problems of the Self*, edited by Bernard Williams, 136–51. Cambridge University Press.

———. 1979. "Internal and External Reasons." In *Rational Action*, edited by Ross Harrison, 101–13. Cambridge University Press.

———. 2000. "Formal Structures and Social Reality." *Trust: Making and Breaking Cooperative Relations* 1: 3–13.

Williamson, Timothy. 2002. *Knowledge and Its Limits*. Oxford University Press on Demand.

———. 2013. "Knowledge First." In *Contemporary Debates in Epistemology*, edited by Matthias Steup John Turri, 1–10. Blackwell.

———. 2016. "Justifications, Excuses, and Sceptical Scenarios." *The New Evil Demon. Oxford University Press, Oxford*.

———. 2017. "Acting on Knowledge." In *Knowledge First: Approaches in Epistemology and Mind*, edited by J. Adam Carter, Emma C. Gordon, and Benjamin W. Jarvis, 163–81. Oxford: Oxford University Press.

———. 2021. "Epistemological Ambivalence." In *Epistemic Dilemmas*, edited by Nick Hughes. Oxford: Oxford University Press.

Wolff, Kurt H. 1950. *The Sociology of Georg Simmel*. Glencoe, Ill: Free Press.

Wong, David B. 2006. "Moral Reasons: Internal and External." *Philosophy and Phenomenological Research* 72 (3): 536–58. https://doi.org/ppr200672338.

Wright, Stephen. 2010. "Trust and Trustworthiness." *Philosophia* 38 (3):

615–27. https://doi.org/10.1007/s11406-009-9218-0.

Zagzebski, Linda. 1994. "The Inescapability of Gettier Problems." *The Philosophical Quarterly (1950-)* 44 (174): 65–73.

Zagzebski, Linda T. 1996. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge University Press.