THE CONDITIONS OF TRUST

Michael Pope

A dissertation

submitted to the Faculty of

the department of Philosophy

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Boston College
Morrissey College of Arts and Sciences
Graduate School

April, 2024

THE CONDITIONS OF TRUST

Michael Pope

Advisors: Richard Atkins and Daniel McKaughan

Abstract: Trust is vital for much of what we know and do. Yet, standard accounts of trust face a problem. Either they analyze trust in terms of necessary and sufficient conditions such that they face counterexamples, or they fail to explain trust's role in social practices. To address this problem, I develop and defend a view that I call *pragmatic pluralism*. Pluralism is the view that trust comes in many forms. I show how pluralism can address counterexamples and preserve the insights of standard theories of trust. However, pluralism neglects to explain how the diverse interests of different parties coalesce in cooperative social practices. In turn, pragmatism provides an explanatory strategy for uniting various forms of trust according to their function. Specifically, I examine trust's role in disposing parties to rely on each other to achieve their goals. This shared, dispositional function explains how various forms of trust can facilitate cooperative social practices. I argue that pragmatic pluralism is plausible given developments in empirical trust research. I then apply insights from pragmatic pluralism to disputes about values in science and trust in artificial intelligence. I contend that *well-placed* trust in each case requires a normative view about the appropriate conditions of trust. While pragmatic pluralism is a descriptive account of trust, I conclude that it provides resources for inquiring about the normatively appropriate conditions of trust—those conditions according to which a trustee is worthy of trust.

TABLE OF CONTENTS

For Kate

ACKNOWLEDGMENTS

A number of people have made this project possible. Foremost among them is my wife, to whom this is dedicated. Her love, support, and patience have afforded me the opportunity to pursue this project. My parents, Mike and Laura, and brother, Jon, have provided invaluable support and encouragement over the years, as have my parents-in-law, Vinny and Jen. I have also had the tremendous benefit of two kind and generous advisors, Richard Atkins and Dan McKaughan, without whose individual and collective support I would not be the philosopher and person I am. Boston College and my colleagues in the Philosophy Department have provided an ideal environment for pursuing my research. I am also grateful to Kevin Elliott and Tom Simpson for their support of the project. It is a mark of my abundant blessings that there are too many people to mention here. I wish I could thank all those colleagues, teachers, collaborators, and friends who have contributed to this dissertation in some way, but I am especially grateful to the following: Sam Hall, Greta Turnbull, Isaac Handley-Miner, Liane Young, Katherine McAuliffe, Micah Lott, Katie Harster, Jorge Garcia, William Cochran, Jeff Behrends, Pete Blum, Bill Haynes, and Todd Meadows. I have also received insightful and encouraging feedback from participants at conferences in Indiana, Illinois, Texas, Chicago, Tennessee, Hannover (Germany), and Dublin (Ireland). Despite this help, I know there is much work to do in the future, so any errors in what follows are my own.

LIST OF FIGURES

INTRODUCTION

Trust plays a crucial role in much of what we know and do. As Annette Baier (1994, 98) puts it, trust is like the air we breathe—something both essential and taken for granted in living, until it becomes polluted. Yet, as Onora O'Neill remarks, "[p]lacing and refusing trust intelligently is demanding, and leaves opportunities for those who seek to push false claims, to fail in their commitments, or to pretend to competence and expertise that they lack" (2013, 161). When well-placed, trust promises to enliven and economize our social, political, and intellectual lives. For this reason, it is important not only to understand trust's nature and function, but also to consider the conditions under which a trustee is trustworthy. Research on trust over the past four decades reveals a host of factors relevant to our reliance on others, including epistemic, practical, moral, social, and political factors. Despite this progress, consensus about how best to understand trust remains elusive. In this dissertation, I develop an account that can preserve the insights of existing views while resolving two critical problems that they face. I call the view *pragmatic pluralism* about trust.

I distinguish between two broad approaches to trust. First, *monism* is the view that trust has a unified and paradigmatic form. In Chapter One, I show how influential existing accounts of trust develop monist views of trust by analyzing trust in terms of necessary and sufficient conditions. There are specific problems facing monist views, but I argue that there are two general problems for monism. First, monist analyses face counterexamples, often generated by other plausible monist accounts of trust. Second, addressing counterexamples presents the monist with an explanatory dilemma. Here's the dilemma. Either a monist view handles difficult cases implausibly, or it fails to explain how the interests of different parties result in cooperative social practices. The upshot of these two problems, I argue, is that monism fails, suggesting the need for an alternative approach.

Second, *pluralism* about trust is the thesis that trust comes in many forms. I argue in Chapter Two that pluralism addresses the counterexample problem for monism, while leaving the explanatory problem unresolved. To this end, I develop *pragmatism* about trust. Pragmatism is a methodological approach for identifying, what Matthieu Queloz calls, "internally diverse" phenomena *by* focusing on their function (2021, 25); in a slogan: function-first. I argue that the common function of plural forms of trust is in

disposing trustors to rely on trustees. That is, according to pragmatic pluralism, what forms of trust share is a disposition to rely, while those forms are distinguishable according to different conditions.

In Chapter Three, I argue that pragmatic pluralism is plausible given how trust is conceptualized and operationalized in empirical trust research. First, social scientists conceptualize trust as a psychological state indicative of a willingness or tendency to rely on trustees. Further, this psychological state is sensitive to different dimensions (or condions) of trust. I argue that trust's multidimensionality, as it is called in empirical literature, suggests pluralism about trust. Second, to understand the impact of various dimensions on trust, researchers develop models. I argue that two influential models are consistent with pragmatism. That is, the models orient trust as a psychological state that disposes trustors to act according to salient dimensions (or conditions) of trust. To conclude Chapter Three, I argue that pragmatic pluralism is both aided by and beneficial to empirical trust research, providing opportunities for collaborations between philosophers of trust and empirical trust researchers.

In Chapter Four, I consider a problem for non-expert trust in science. The problem emerges from two points. First, many philosophers of science accept that science is value-laden. Second, empirical evidence suggests that values similarity between parties directly affects trust and, in turn, perceptions of risks and benefits. The problem is that when values diverge (i.e., when there is value *dis*similarity), it can be rational to reduce or suspend trust. I call the resulting problem *value divergence*. I examine five strategies for addressing value divergence and argue that each involves higher-order value judgments. Yet, for the purposes of grounding trust, each strategy alerts us to possible conditions on trust. I conclude by arguing that trust in science is itself value-laden and attending to conditions on trust contributes to a norm-based approach to values in science.

In Chapter Five, I engage with recent views denying that trust is possible in AI. I argue that trust in AI is possible. The pressing question for consideration is not whether we *can* trust AI, but whether we *should*. I connect this argument to approaches to trustworthy AI, arguing that pragmatic pluralism helps to clarify considerations in identifying normatively appropriate trust in AI—that is, AI worthy of our trust.

Nevertheless, as a descriptive theory of trust, there are limits to pragmatic pluralism. The view reveals how trust can lead people to rely under myriad conditions and for multiple ends. Some cases will rightly meet with disapproval, as in sexist, racist, or abusive forms of trust. Yet, pragmatic pluralism explains how these are indeed forms of trust precisely in the sense that certain conditions (appropriately or not) dispose trustors to rely on trustee. Of itself, the view offers limited normative input for adjudicating when trust is appropriate or not. What it provides, however, is a means for approaching questions about what should dispose one to rely on another in relevant

circumstances. In this way, pragmatic pluralism can play an important part in developing an ethics of trust.

1. 0 BEYOND MONISM ABOUT TRUST

Consider Aesop's fable "The Lion and The Eagle."

> An Eagle stayed his flight and entreated a Lion to make an alliance with him to their mutual advantage. The Lion replied, 'I have no objection, but you must excuse me for requiring you to find surety for your good faith; for how can I trust any one as a friend, who is able to fly away from his bargain whenever he pleases?'[1]

Trust is widely believed to be essential for cooperation and mutual advantage. Personal relationships, communities, economies, and governments depend on it. Yet, we are vulnerable in trusting, for trustees may abandon us when it pleases them. So, one must not trust indiscriminately. The Lion might ask: under what conditions should I trust the Eagle? To act in good faith, as the Lion requires, should the Eagle be trusted as a friend? Further, both Lion and Eagle have their limits and competencies that affect their reliability. Are there forms of trust, some for friends and others for colleagues or acquaintances—even rivals? And for what can the Lion trust the Eagle?

In consideration of these questions, this chapter examines influential philosophical analyses of trust. Views that defend what I call *monism* about trust hold that all instances of trust are instances of the same thing. To identify trust's nature, monist views attempt to set out necessary and sufficient conditions for trusting others. A consequence of—and sometimes a motivation for—monism is that certain putative cases of trusting are ruled out as trust. As such, monism also involves an evaluative component: Some putative cases of trusting turn out not to be cases of trust at all. Monist views face two problems, I argue. First, they face counterexamples, often emerging from other

[1] From Mondschein (2023, 39).

plausible views. Second, monism faces a dilemma. Either the view handles difficult cases implausibly, or it fails to explain how the interests of different parties result in cooperative social practices. The upshot of this chapter is that monism about trust fails.

I proceed as follows. First, I provide preliminary distinctions for analyzing trust. Second, I critically examine influential monist approaches to trust. Third, while I contend that monism fails, I also argue that the plausibility of many analyses of trust indexed to specific situations suggests that trust is a pluralistic attitude. I develop this pluralistic insight in Chapter 2.

1.1 PRELIMINARIES

Before proceeding, a few preliminaries are in order.

1.1.1 TRUST AND RELIANCE

First, while there is widespread agreement that trust is related to reliance, their relationship deserves careful attention. Most philosophers regard trust as a subspecies of reliance.[2] That is, trust is viewed as mere reliance *plus* a set of necessary and sufficient conditions for a relationship to be one of trust. For example, in her seminal work on trust, Baier defines trust as a type of confident reliance on the goodwill of a trustee (1986, 1994). Is reliance, including *confident* reliance, an attitude, an action, or a combination of both? I contend we should view reliance as a type of action, particularly a *strategic, non-basic* form of action.

[2] For a view of trust without reliance, see Thompson (2017). For a reply, see Hawley (2019, 7).

Richard Holton (1994) distinguishes two senses in which our relying is strategic. We can rely on something happening. For example, I might rely on prisoners remaining in prison, without relying on the prisoners to remain in prison. Alternatively, we can rely on someone to do something. "If I rely on you to stay in prison," Holton remarks, "I do not simply plan on the supposition that you will stay there; I plan on the supposition that you will stay there because you are motivated to stay there, and not just because you have no choice" (1994, 3). In other words, when *A* relies on *B* to *φ*, party *A* incorporates party *B*'s *φ*-ing into their plans with respect to *φ* or a goal of which *φ* is a part. Moreover, as Berislav Marušić (2017, 3) argues, to rely "is to act in a way in which the success of your action—its achieving its end—depends on what or who you rely on." Of course, beliefs, desires, and intentions are often crucial to planning. The point here is that those attitudes are not inherent to relying.[3] More specifically, reliance is non-basic form of action. As Daniel Howard-Snyder and Daniel McKaughan (manuscript) argue, we rely on others *by* doing something. For example, I rely on my car to get to work *by* driving it to work—by doing something—such that my arriving at work is contingent on my car. So, when *A* relies on *B* to *φ*, *A* plans in light of *B*'s *φ*-ing *by* doing or abstaining from doing other activities.

Most attempts to distinguish mere reliance from trust focus on the relevant attitudes involved. Cognitive approaches to trust, for example, take trust to be mere

[3] My focus in this chapter is on trust, so I must set aside a fuller discussion of reliance to connect reliance to trust. But there are two alternatives that cast reliance as a type of attitude worth mentioning. Matthew Smith (2010) draws a distinction between internal and external reliance, where the former involves a credal-conative attitude or suite of attitudes that accompanies the latter type of reliance. Marušić (2017) is right to question the need for an internal attitude that must accompany our acts of relying and one's beliefs, desires, or intentions about that reliance. Alonso (2009, 2014, 2016) offers a subtler view, according to which reliance is a practical attitude that can guide our actions. In other words, reliance could be a type of acceptance.

reliance plus a belief about the motivations or interests of the trustee.[4] Normative views, in contrast, emphasize reactive attitudes, such as feelings of betrayal or gratitude, as the hallmarks of trust. As I argue in §2.1.2, investigating the relevant attitudes that accompany trusting is often useful for identifying conditions on trust in a particular context, but it is a misleading method for understanding what trust *is*. For example, two leading and distinct approaches to trust, affective approaches and predictive approaches respectively, both take trust to entail a belief. What the belief is *about* is non-trivially different, with the former focusing on care and reciprocity and the latter emphasizing information about shared interests and a trustee's reliable behavior. While belief focuses our attention to what is relied upon when trusting, there are other instances of trust for which the relevant belief is unnecessary.

Instead, I investigate trust's role or function in facilitating reliance. I concur with Holton that trust is "a distinctive kind of attitude involving a distinctive state of mind" (1994, 1). Or, as Baier describes it, trust is an "intentional mental phenomenon" (1994, 100). In Chapter 2, I argue that trust is a psychological state that *disposes* one to rely on another party for some aim or target. For the present, what is important is to see that reliance, as an action, involves depending on something happening or someone doing something, while trust is one's attitude toward this dependence.

1.1.2 Trust and Trustworthiness

The second preliminary is the relationship of trust to trustworthiness. A natural place to begin an analysis of trust is with cases where trust is well-placed or, in other words, when

[4] For examples, see Hardin (2002; 2006), Keren (2014), and Simpson (2017; 2018).

one trusts the trustworthy. For example, O'Neill contends that "Trust is valuable only when directed to agents and activities that are trustworthy" (2019, 159). Moreover, Russell Hardin, whose view I examine in the next section, argues that many analyses of trust are in fact analyses of trustworthiness (2006). By linking trust to trustworthiness, we could better position ourselves to assess what trust is and when it is warranted.

I have two reasons for hesitation about this approach. First, to say that someone or something is trustworthy is to say that that person or thing is worthy of trust, suggesting that trust and appraisals of trustworthiness are interrelated. So, we might expect disputes about trust to resurface in identifying the hallmarks of trustworthiness. Second, belief about the trustworthiness of a trustee does not entail trust. One could think that a potential trustee is trustworthy without thereby coming to trust. Consider O'Neill's conditions for assessing trustworthiness, namely evidence of a trustee's honesty, competence, and reliability. When receiving medical advice, I might judge that a physician is generally honest, competent, and reliable (trustworthy as a physician in general) without thereby trusting the physician, say if the stakes are especially high or if I think the physician disdains me. In Baier's view, trust in its most genuine form is sensitive to a trustee's goodwill in addition to their competence. One could reply, as O'Neill (2002) does, that the patient *should* trust the physician, but this is to dispute what the conditions on trust *ought* to be rather than what they may possibly be.

1.1.3. Conditions of and for Trust

Considering the possible conditions on trust, a third preliminary is to distinguish between the conditions *of* trust and the conditions *for* trust. The conditions *of* trust are the

necessary conditions inherent to trusting. The conditions of trust are manifest in the expectations of trustors and trustees within a domain. If conditions go unmet, one must either revise one's conception of trust or the trust relationship is suspended. In the physician case, these are the conditions under which the patient is prepared to rely on the physician. The conditions of trust are those by which a trustor, the patient, judges the trustworthiness of the trustee, the physician.

The conditions *for* trust, in contrast, are the socio-historical features of the context in which trust relationships are formed and maintained. Early sociological investigations of trust, for example, examine how differences in social organization impacted trust and related attitudes. Niklas Luhamnn (1979) discusses at length differences in interpersonal trust and confidence in abstract systems and institutions.[5] What emerges is a complex picture of cooperation, calculation, and reciprocity in social living. Such investigations reveal ways that conditions *of* and *for* trust can interact in interesting ways. For instance, supposing a trustee's goodwill is a necessary condition of trust, we should expect trust relationships to be rarer in authoritarian contexts, where fear and threat can erode social cohesion, in contrast to conditions of expression and association in freer societies. This is because conditions of goodwill are significantly constrained in the former case, unlike the latter. Yet, given the plurality of conditions that could possibly dispose one to rely, I will argue below that it is possible for one to trust on the basis of fear or threat. We may rightly identify such forms of trust as exploitative, even immoral. Yet, they are exploitative and immoral forms of *trust*. This results in an important distinction with

[5] We can see how social organizations founded on status differ from more modern systems founded on contractual relations. This latter shift can be seen in the emphasis on promise-keeping that stretches at least as far back as Hugo Grotius and Samuel von Pufendorf, and is shared by John Locke, David Hume, and Immanuel Kant.

which I conclude this chapter, namely between descriptive and normative assessments of the conditions of trust. We should beware neglecting a shift *from* descriptive analyses of what trust is and how it functions *to* normative views about the appropriate conditions of trust.

1.2 Monism about Trust

I distinguish between two broad approaches to understanding the conditions of trust, monism and pluralism respectively. Pluralism is the view that trust can come in many forms, rendering it unamenable to analyses in terms of necessary and sufficient conditions. Next chapter, I consider pluralist approaches. In this chapter, I focus on what I call monist views of trust, which aim to restrict the conditions of trust to rule out as illusory or to demote certain instances of trust.[6] For example, Baier's inclusion of goodwill as a necessary condition of trust is intended, in part, to rule out (apparent) forms of trust that arise from fear, anger, and threat, as well as to render more calculative forms of trust derivative and artificial (1994, 98). Baier is not alone in this effort, however. Other views prioritize predictive information, viewing trust as necessarily involving confidence about a trustee's cooperative behavior. Russell Hardin's encapsulated interest view (2002, 2006) is paradigmatic in this vein. Normative views, in contrast, see trust as involving normative or moral expectations that a trustee will act in a particular way, for which we can hold them responsible and feel gratitude or betrayal depending on trust's

[6] It is tempting here to say that monists aim to rule out or demote *forms of trust*. However, I take it that monists hold that either "forms of trust" indicates different instances of the same thing (what I call *weak monism* below) or we should instead say "instances of trust" since anything failing the relevant monist definition is not really trust (what I call *strong monism* below).

fulfillment. Richard Holton's (1994) view of trust as involving a "participant stance" is influential on this score. Different still, more recent approaches mix elements from affective, predictive, and normative views to formulate a unique set of conditions that restrict what counts as trust, often within a particular domain. For example, Gürol Irzik and Faik Kurtulmus (2019) develop "enhanced" public trust in science. They argue for conditions of transparency and value similarity between stakeholders and scientists are necessary for grounding public trust in science.[7] It is important to see what monist views *share*, viz. an effort to elucidate what trust is by identifying necessary and sufficient conditions for trusting.

Monism can come in stronger and weaker varieties. Consider cases of "therapeutic trust." Therapeutic trust involves cases in which trust is undertaken with the aim of bringing about trustworthiness. In this way, it is a *conative* instance of trust. For example, parents might entrust money to their child for a night out with friends *not* because they think or predict that the child is trustworthy, but because they desire for the child *to become* trustworthy.[8] The strongest form of monism designates the conditions relevant to a paradigmatic case as *the* conditions of trust, such that trust under other conditions either is not trust or is confused (even immoral). For example, a strong

[7] Irzik and Kurtulmus (2019) clarify that they are interested in second-order reasons for warranted trust in science. To this end, they set aside "a socio-psychological account of why public trust or distrust scientists" (ibid., 3). However, as T.Y. Branch (2022) argues, this presents conceptual, relational, practical challenges. Indeed, I argue below, and again in the next two chapters, these socio-psychological features of trust are crucial for understanding what trust *is* and how it might be enhanced. I return to Irzik and Kurtulmus' view in Chapter 4.

[8] Therapeutic trust is hotly contested. I do not purport to address whether therapeutic trust is really trust here, although it will be clear below that I think it *is* a form of trust. Whether it is a normatively appropriate form of trust is a separate matter. For more on therapeutic trust, see Horsburgh (1960), Nickel (2007), Hieronymi (2008), McGreer (2008), Simpson (2012), McLeod (2015), Carter (2022), Pace (2021), among others. An early identification of this phenomenon, though without the label 'therapeutic trust', is Gambetta (1988, 234), who argues that trust "can generate the very behaviour which might logically seem to be its precondition."

predictive monist takes it that confidence about a trustee's performance is *always* relevant to trust. On this view, therapeutic trust simply is not trust, since it is insensitive to evidence of another's unreliability. Either 'therapeutic' trust is a case where one misattributes trustworthiness when there likely is none, or therapeutic 'trust' is more akin to hope rather than trust.[9]

Alternatively, a weaker monist view identifies a hierarchy of forms of trust descending from a paradigmatic case, creating space for artificial and suboptimal forms of trust. For example, Carolyn McLeod (2002) proposes a 'prototype' theory of trust, according to which interpersonal cases of trust provide a prototype from which other forms of trust are derived. Accordingly, trust is a more flexible concept than traditional analyses allow, while at the same time having a set of prototypical features. For the therapeutic case, the parents entrust money to their child out of a duty to raise responsible citizens. For proponents of a more tolerant affective view, it can be admitted that real trust exists in such a case, and even that this form of trust serves a particular civic good, while acknowledging that there is something nonprototypical about the case. Nonetheless, weaker forms of monism must identify prototypical features relevant for identifying cases of trust and organizing those cases into some hierarchy, returning us to the problems faced by stronger monist views.

[9] Hardin (2002, 73-75) argues to this effect. For an alternative view, where trust involves confidence sufficient for certain risks, see Pace (2021). Pace argues that it is possible to entrust something without trusting such that some cases of therapeutic cases do not count as trust. He defines entrusting as (1) assigning responsibility to someone for something and (2) placing it into their care (ibid., 11902). Clearly, one can fulfill both conditions without trusting. This is consistent with my view, since one can rely without trusting, provided that one views reliance as a type of action. Terminologically, I prefer 'reliance' for this type of action because 'entrusting' seems to imply that one acts from a position of trust. I return to entrusting in Chapter 2.

An examination of monist views is not entirely negative, however. Determining the strength of a view is often contingent on features of a particular case, such as the history and nature of the relationship between relevant parties. Seeing the apt descriptions provided by monist views reveals the conceptual flexibility and diversity of forms of trust. Attention to the insights and differences between these views points us to an alternative approach to trust, namely, pluralism. I argue that it is through the conditions and particularities revealed in monist views that we can construct the most plausible form of pluralism.

1.2.1 Affective Conditions of Trust

Recent philosophical interest in trust begins with Baier's "Trust and Antitrust" (1986, 1994). As noted above, Baier defines trust as confident reliance on another's goodwill. Baier's paper offers a rich discussion of trust relationships, in many ways setting the agenda for future inquiries. She describes trust as a ubiquitous phenomenon that appears in many forms. She writes, "Trust can come with no beginnings, with gradual as well as sudden beginnings, and with various degrees of self-consciousness, voluntariness, and expressions" (1994, 105). The breadth of trust's relevance highlights both its significance and possible complications for identifying what it is.

To begin, Baier asks: What is the difference between trust and mere reliance? She contends that it must be "reliance on their goodwill toward one, as distinct from their dependable habits, or only on their dependably exhibited fear, anger, or other motives compatible with ill will toward one, or on motives not directed on one at all" (1994, 98–99). The heart of Baier's approach is to distinguish cases of reliance conditional on

another's predictable habits from those in which a trustee is motivated by goodwill toward a trustor and care for whatever is entrusted. That is, trust involves "letting other persons (natural or artificial, such as firms, nations, etc.) take care of something the trustor cares about, where such 'caring for' involves some exercise of discretionary power" (Baier 1994, 105). To be sure, reasonably believing that a trustee bears one goodwill is difficult in many cases. For this reason, Jones (1996) clarifies that one need only have optimism about a trustee's goodwill. But what matters according to this view is that the basis of one's trust is belief in or optimism about the trustee's goodwill, rather than optimism, belief, or any other attitude about a trustee's reliable behavior. In other words, what is relied upon when trusting is the goodwill of the trustee. In this sense, affective conditions of care and goodwill are necessary for trust.

We can view Baier's argument in two phases. The first is broader and addresses the neglect of trust in moral philosophy, while the second involves a more specific argument about determining the necessary conditions of trust. Let's take each in turn. Baier's account of trust fits into her general critique of contractarian ethics, with its assumptions about power, gender, and social status. As she pointedly puts it, "[i]t takes inattention to cooperation between unequals, and between those without a common language, to keep one a contented contractarian" (1994, 106).[10] Her Humean alternative orients ethical considerations to care and reciprocity, where trust and trustworthiness play a central role. Baier argues that an overemphasis on contracts and predictive conditions

[10] With rare exception, Baier argues that modern philosophers—primarily a "collection of clerics, misogynists, and puritan bachelors"—"managed to relegate to the mental background the web of trust tying most moral agents to one another and to focus their philosophical attention so single-mindedly on cool, distanced relations between more or less free and equal adult strangers" (1994, 114). Only "trade fetishists" could see every social interaction as one of contract and exchange (ibid, 109).

on trust distorts trust's moral character. She remarks: "not all the things that thrive when there is trust among people, and which matter, are things that should be encouraged to thrive. Exploitation and conspiracy, as much as justice and fellowship, thrive better in an atmosphere of trust" (1994, 95). Not just any form of confident reliance will do. Distinguishing moral from immoral forms of trust motivates Baier's monistic affective approach.

The second phase of Baier's argument is what I call the 'developmental argument'. It is not that Baier denies that there are other forms of trust or that one can trust on the basis of calculation and contract. Rather, she argues that what she calls 'infant trust' reveals the necessary affective condition on trust, rendering alternative forms of trust derivative and artificial. Exhibited in relationships between parents and children, infant trust is a non-contract-based form of trust wherein trust is automatic and intimate. "Parental and filial responsibility," Baier writes, "does not rest on deals, actual or virtual, between parents and child" (1994, 110). What requires explanation is "ceasing to trust, the transfers of trust, the restriction or enlargements in the fields of what is trusted, when, and to whom, rather than any abrupt switches from distrust to trust" (1994, 111). That is to say, infant trust is the "essential seed" of other trust relationships (1994, 110). Contractual forms of trust develop from—and therefore are derivative of—a more basic form of trust, infant trust. We learn to rely on others in situations where goodwill plays no role from "the prior existence of less artificial and less voluntary forms of trust, such as trust in friends and family" (1994, 112). For this reason, Baier argues that relying on the goodwill of a trustee is a necessary condition for the most authentic and basic form of trust.

1.2.2 Limitations of Affective Views

Is goodwill necessary and sufficient for trusting? Richard Holton (1994) argues to the contrary. Recall O'Neill's physician case. Suppose that the physician is the best in her subfield, a real medical rock star. Yet, she is motivated *only* by the intellectual challenge that medical cases present, caring nothing for patients' well-being. It is possible that she feels a sense of accomplishment when a patient recovers and that she revels in the approbations and financial rewards that accompany her success. It is only that she is not motivated by goodwill toward her patients. Moreover, professional constraints and regulations ensure that there is oversight and accountability. In such a case, the fact that the physician bears *you* goodwill is not *necessary* for trusting her. Is goodwill sufficient for trusting? A conman might rely on a trustee's goodwill to exploit her. For trust among friends, as Aristotle argues for friendship, goodwill is insufficient without each party reciprocating goodwill.[11] So, goodwill alone on the part of the trustee is insufficient in many cases.

It is also unclear exactly what role goodwill should play in facilitating trust relationships. In cases where conditions of care and reciprocity are relevant to trusting, it is not clear that goodwill should be directed toward the trustor. Suppose parents hire a babysitter. They entrust the babysitter with the wellbeing of their child. Yet, they need not do so on the basis of goodwill *toward them*. Rather, they might think that it is goodwill toward what is entrusted, namely the child. Yet, defenders of affective views might respond: goodwill still plays a crucial role in such cases, similarly to the way that

[11] See *Nic. Ethics* 1156a.

one entrusts one's own well-being in cases of trust generally. So, according to affective proponents, this is less a counterexample than an illustration of the ways care and reciprocity can influence trust relationships. This leads to a deeper worry than counterexamples for affective views.

Given the ubiquity of trust that Baier identifies in social living, why take infant trust as the paradigm? To be sure, it seems conditions of care and goodwill can be central to intimate, interpersonal trust relationships, especially when the thing entrusted to another is of great value. There is a difference between (a) trust between children and parents and (b) trust between adults. In many cases, the differences lie in asymmetries of power, as Baier suggests. But is not the adult trust less-gullible and more discerning? Rather than infant trust, the mark of development is learning to curate trust relationships, growing from the milk of infant trust to the rich food of adulthood, as it were. The point is not to focus on the explicitness and contractual nature of forms of trust that rely on predictions about another's behavior. Rather, developing heuristics for assessing relevant features of another's behavior is an essential part of learning to place trust wisely. This is evidenced by findings in developmental psychology, where very young children begin to track the reliability of informants over time, even if they have a deep bond with the informant.[12] Baier seems to acknowledge this point, writing: "The trustor, who always needs good judgment to know whom to trust and how much discretion to give [to the trustee], will also have some scope for discretion in judging what should count as failing to meet trust, either through incompetence, negligence, or ill will" (1994, 103). Yet, there

[12] See Tummeltshammer, et al. (2014), Butler (2020), Corriveau and Harris (2009), Corriveau, et al. (2013), Corriveau and Kurkul (2014), Harris, et al. (2018), Harris (2012), Jaswal and Neely (2006), among others.

is a difference between someone acting from ill will and being incompetent.[13] This alternative view of trust's development, then, is that we learn to determine when goodwill and reciprocity are reliable indicators that a trustee will come through for us.

With the combination of competence and goodwill, we see both the insight and the limits of affective conditions of trust. On the one hand, in more intimate cases, considerations of care and goodwill are necessary for the continuation of trust. For instance, it would be highly unusual for romantic partners to trust only on the basis of written contracts, paying no mind to reciprocated care for each other's well-being. On the other hand, requiring affective conditions in more socially distant relationships can be inappropriate, even harmful. For while some cases might result in social awkwardness or even humor—say, when a mechanic discovers that a customer expects affective motivations in addition to competence—other cases can place trustees in compromising positions, especially in cases of power asymmetry.[14] Consider again the physician case from §1.1.2. When we trust a physician, *what* we trust the physician to do is often underdefined. We expect the physician to act conscientiously with respect to our health. Does this require the physician bear us goodwill? The fact that people can reasonably disagree does not undermine Baier's interest in the moral importance of trust. Rather, the challenge is to discern the appropriate conditions of trust across situations.

1.2.3 Predictive Conditions of Trust

[13] Indeed, we can distinguish between goodwill as a positive stance toward a trustor (or what a trustor entrusts) and goodwill as minimally involving the absence of ill will. For goodwill conceived as merely lacking ill will fails to institute a strong condition on trust, since one could reasonably trust another on the basis of their having no ill will toward her because the trustee has no knowledge of the trustor. In such a case, the trustor's only recourse is to a trustee's reliable habits.

[14] I return to this point in §2.3.3 with Dormandy's (2020) objection to ascribing obligations through trust.

In contrast to affective approaches to trust, predictive approaches view trust as a matter of rational choice. According to such views, at its most basic level, trust is a cognitive attitude that is sensitive to the reliability, competence, and interests of a trustee. In this way, when *A* trusts *B* to *φ*, as Russell Hardin remarks, *A*'s trust involves "essentially rational expectations about the self-interested behavior of the trusted" (2002, 6). In this subsection, I discuss the insights and limitations of predictive views. I begin with a more sociological perspective before turning to Russell Hardin's influential encapsulated-interest account of trust. While I argue that predictive conditions are neither necessary nor sufficient for trusting, the relevance of predictive information to many instances of trust highlights salient features for understanding trust.

1.2.4 COLEMAN ON TRUST

The sociologist James Coleman offers the most austere and straightforward of predictive views, according to which trust is a matter of rational betting behavior. His account of trust is situated within an investigation of social cohesion and declining confidence in individuals and institutions in the latter half of the twentieth century.[15] He views trust as a crucial part of "the transactions that make up social action" (1990, 91). When presented with an opportunity to trust, Coleman argues, "the elements confronting the potential trustor are nothing more or less than the considerations a rational actor applies in deciding whether to place a bet" (ibid., 99). That is, trust is a matter of interest-maximizing rational *action*. The "elements confronting the potential trustor" in *placing*

[15] In this way, Coleman's view relates to and is influential on Robert Putnam's (2000), Francis Fukuyama's (1995), and Adam Seligman's (1997). As with Coleman, trust and distrust play only a partial role in understanding declining civil engagement. The part it plays, however, is directly related to cooperation, risk tolerance, and the management of social capital. See especially Putnam (2000, 134–47).

trust are only how much she could lose, how much she could gain, and the chance that a trustee will succeed. Clearly, affective information could be factored into the likelihood that a trustee will behave in a particular way, but, according to Coleman's predictive view, this information is only relevant to a small subset of trust cases. From the perspective of predictive views, it may be that goodwill between trustor and trustee facilitates trust in the sense of allowing one to predict a trustee's behavior. As an empirical matter, it may be that relying on affective information is a less secure bet than other sources of information. What is crucial is to see that affective information is only one possible basis for trust in the predictive sense.

Coleman approaches trust to understand how systems of "social interdependence" produce certain types of resources within and across communities, especially social capital (1990, 300).[16] Different from material or financial capital, Coleman defines social capital by its function. Rather than a single entity, on this view, social capital is inherent to social organization and is crucial for the development of individuals and communities. It is manifest in social obligations and expectations, in potential information sharing, in norms and effective sanctions, in relations of authority, and in the promotion of public goods through stabilizing and limiting social organization (see ibid., 304–21). For Coleman, trust is a central source of social capital, revealing ways in which "individuals do not act independently, goals are not independently arrived at, and interests are not wholly selfish" (ibid., 301).

Coleman highlights several important functions of trust relationships. Three are worth mentioning here. First, through trust, the trustor "allows an action on the part of the

[16] For discussions of social capital in general and in relation to trust, see Loury (1977), cited by Coleman (1990, 300), and Putnam (2000).

trustee that would not have been possible otherwise" (ibid., 97). That is, while trusting makes the trustor vulnerable to another's actions, trust facilitates cooperation. Second, a trustor is better off when trust is fulfilled and worse off when trust is unfulfilled. For example, Coleman discusses a farmer whose hay crop is in jeopardy (ibid., 93). By trusting his neighbor for help, even when he does not know the potential cost of the neighbor's help, the farmer avoids losing his crop and thereby increases the likelihood of profits. In this way, trust facilitates cooperation in conditions of risk and uncertainty. Third, trust often functions without any explicit commitment on the part of the trustee. Knowing that a trustee will behave in a particular way can produce the best outcome for the trustor, irrespective of the trustee's awareness or commitment. Indeed, sometimes the only evidence we have is the performance record of a potential trustee.

1.2.5 Limitations of Coleman's View

There are two problems for Coleman's predictive view. First, we can distinguish between trust and actions that result from trust. As Simpson (2012, 553) rightly points out, viewing trust as an action—emphasizing *placing* trust—obscures the motivational character of trust. Simpson introduces the following case. With no children of her own, a rich aunt pledges to come to her nephew's aid if he should fall on hard times. As it happens, the nephew is successful, never needing to rely on his aunt's generosity. There is nothing objectionable in saying that the nephew trusts that the aunt *would* have bailed him out, if he had ever needed it. So, the nephew trusts without ever acting in a way that directly relies on the aunt.[17] That is, the nephew's trust consists in his readiness to rely on

[17] I say "directly" here because it seems plausible that knowing his aunt will aid him in times of trouble could impact, for instance, the riskiness of the investments he makes.

the aunt, rather than his actually relying on her generosity. For a predictive view, then, we need an account of how trust, as an attitude, is predictive.

Second, not just any prediction or bet about another's performance counts as trust. Imagine in the farming neighbors case that an uninvolved neighbor predicts that her neighbors will harvest the hay. To say that she trusts the farmers to harvest the hay is simply to say that she believes that they will, in contrast to the dependence operating between the other farmers. Accordingly, trust is more directly tied to one's interests than merely one's thinking or believing about another's probable behavior.

1.2.6 Hardin's Encapsulated Interest Account

In turn, Russell Hardin's "encapsulated interest" account of trust is predictive, following Coleman, but nuances how trust differs from similar attitudes. Hardin's account of trust is rationalistic inasmuch as trust is a belief about the behavior of a trustee. As he asserts, "trust is a cognitive notion, in the family of such notions as knowledge, belief, and the kind of judgment that might be called assessment" (2002, 7). That is, trust is "simply an epistemological, evidentiary matter" (ibid., 31). Yet, trust as encapsulated interest is not merely a belief that a potential trustee's and trustor's interests overlap. Overlapping interests is necessary but insufficient for trust. Trust "requires that the trusted values the continuation of the relationship with the trustor and has compatible interests at least in part for this reason" (ibid., 4). So, to say that one's interests encapsulate my own, as a trustor, is to say that the trustee has "an interest in fulfilling my trust. It is this fact that makes my trust more than merely expectations about [the trustee's] behavior" (ibid., 3). Hardin's account reveals how trust is *motivational* on the part of the trustee. "Trust is

little more than knowledge," says Hardin, "trustworthiness is a motivation or set of motivations for acting" (ibid., 31). In other words, a potential trustee is trustworthy only if that trustee is competent within the relevant domain and motivated by the fact that her interests overlap with a trustor's interests. Trust is simply a belief about this trustworthiness.

1.2.7 Limitations of Encapsulated Interest

The first question to ask of Hardin's predictive view is whether a belief about trustworthiness (in Hardin's sense) is necessary for trust. For example, according to the view, it seems that therapeutic trust is either irrational or not trust, being instead perhaps a form of hope or wishful thinking. Recall the parental case, wherein a parent entrusts money to a child to promote responsible habits. The parent can do so without any definite belief that the child will be trustworthy, rendering such a belief unnecessary to trust. To say that such cases fail to count as forms of trust risks begging the question against alternative approaches to trust, since determining whether a trustee is worthy of trust depends on what one takes trust to entail.

Although Baier similarly conceives of trust as involving a belief, that belief concerns the goodwill and competence of a trustee in a way that licenses reliance. In her view, it is the affective dimension that underscores the relational character of trust. Recall her point that trust need not always serve moral ends. To Coleman's insights about trust's social function, she adds that there are forms of trust such as "unconscious trust, as unwanted trust, as forced receipt of trust, and as trust which the trusted is unaware of" (1994, 99). Yet, she argues that a plausible measure of "*proper* trust will be that it

survives consciousness, by both parties" (ibid., my emphasis).[18] The point of Baier's polemic against contractarian ethics, at least with respect to trust, is that basing trust on predictions about common interests fails to capture the goodwill, care, and reciprocity undergirding the most intimate cases of trust. For Baier, calculative and predictive conditions are at best secondary to the care and reciprocity central to the relationship between trustee and trustor.

Additionally, while one could think that information about the competence and reliability of a trustee is necessary for trust, it might not be sufficient. For example, Richard Holton (1994) argues that there is an important disanalogy between belief and trust. Rather than being primarily epistemic, trust instantiates norms and expectations between trustor and trustee. Predictive views restrict the rich motivational role of commitments, expectations, and reactions that are characteristic of trust relationships. It is to this normative view that we turn next.

Before proceeding, I should clarify the force of my objections to predictive views. As with affective views, my argument is not that predictive information is irrelevant to trusting. Nearly every available approach to trust makes some room for confidence about a trustee's reliability and competence. Indeed, Coleman and Hardin's interest in situating trust within contexts where it functions is all to the good, in my view. Uncovering trust's social function can help determine when certain conditions—whether predictive, affective, or otherwise—are relevant to relying on others. So, my point is that predictive information is not a necessary condition for trust, while plausibly being a sufficient condition in many cases. Some cases involve *ex ante* or *ex post* irrationality. For

[18] This is at the heart of Baier's "expressibility test" for the moral decency of trust. In Chapter 5, I return to Baier's argument to consider how an ethics of trust might be developed.

example, prior to entrusting money to their child, parents can be aware of their child's dismal track record and still trust the child. While possibly foolhardy, trusting the child may prove wise or unwise, depending on how well it achieves the parents' aims in trusting. Likewise, one may find out only after trust is broken that a trustee was unworthy of her trust—perhaps owing to the trustee's ill will or incompetence. In future, such an experience may lead the trustor to be more sensitive to any one of those factors. And this is the central upshot: while not necessary, encapsulated interest is one possible condition that can influence one's trust.

1.2.8 Normative Conditions of Trust

In one sense, every monist approach to trust requires normative claims about the right basis for trust. By demarcating conditions of trust to distinguish authentic trust from misplaced trust or non-trust, a monist position about what *should* dispose one to rely on others is built into monist analyses of what trust *is*. For example, if trust is predictive, insensitivity to one's evidence could result in *irrational* trust.[19] But there is a different and more precise sense in which one can have a normative view about trust. Some theorists take trust crucially to involve normative expectations of potential trustees.[20] With her affective view of trust, Baier intends to alert us to the moral status of trust.[21] However, she views this as fundamentally a belief about the goodwill of a trustee.

[19] For the predictive monist, this may not amount to trust at all, instead being a form of gullibility.
[20] Identifying something normative about trust has been common and productive, arguably attracting more attention than affective and predictive views in recent years. For this reason, this subsection will occupy us more than the previous two. For a nice overview, see Carter and Simion (2020).
[21] Baier proposes an "expressibility test" that assesses "the moral decency of a trust relationship" by clarifying "what the other party relying on for the continuance of the trust relationship" (1994, 123). If the relationship can continue after passing the expressibility test, Baier argues, then it is morally decent—or at least we are on our way to determining its moral decency.

Likewise, proponents of predictive views conceive of trust as a fundamentally rational and epistemic attitude, most often referring to beliefs or judgments about another's predicted behavior. In contrast to both these approaches, Richard Holton (1994) argues that there is something amiss in conceiving of trust as necessarily involving belief, whether those are beliefs about another's goodwill or future behavior. Holton's point is that conceiving of trust as *believing* obscures the normative dimensions of trusting. When one relies on another to do something, she incorporates it into her plans. For trust, Holton continues, a trustor invests "that reliance with a certain attitude" (1994, 5). This attitude is trust and involves a certain stance towards a trustee and what is entrusted to a trustee. Belief, in contrast, is an attitude that is sensitive to one's evidence for the truth of a proposition. To be sure, beliefs about others' intentions and capabilities can restrict when, to what extent, and for what purposes one can rely on another. Yet, we can have beliefs about others' behavior or goodwill without thereby trusting, just as we can rely on others without trusting them. What is unique about trust, Holton argues, is that it involves a *participatory* relationship between trustor and trustee.

For Holton, what distinguishes mere reliance from trust is the appropriateness of *reactive attitudes* in the latter case but not the former.[22] Feelings of resentment when harmed or gratitude when helped by another person characterize reactive attitudes in the

[22] Holton draws from Strawson (1974) in developing an account of reactive attitudes and the participant stance. Strawson argues that we can adopt two different attitudes toward someone who commits a harmful act. To take a *participant attitude* toward someone is to treat that person as a member of the moral community and, therefore, as someone who is responsive to moral reasons and capable of goodwill toward others. When someone we take the participant attitude toward harms another, we hold that person responsible. In contrast, we take an *objective attitude* toward those we treat as non-responsible agents. These agents must be trained, incentivized, corrected, restricted, and so on in order to protect oneself from their potentially harmful behavior. From Strawson, Holton develops the participant stance from the participant attitude. For Jones, below, we can see how the objective attitude relates to predictive expectations, while normative expectations arise from the participant attitude.

relevant sense. For example, when a book shelf collapses, crushing a valued vase, one would hardly feel betrayed by the bookshelf, although one might blame oneself or others for placing the vase on the shaky shelf. This is because one relies on the shelf but does not trust it.[23] Suppose instead one lent the vase to a friend in whose care it is shattered. Even if the breaking is accidental, one might feel betrayal in the sense that the friend should have taken more care. If the breaking is intentional, then feeling betrayed would be appropriate. Alternatively, when the vase is returned intact, one might feel gratitude or relief. For Holton, these reactive attitudes are part of a broader stance that colors trust relationships, namely the *participant stance* (1994, 4). A trustor's readiness to feel reactive attitudes exhibits a participant stance toward a trustee. In this way, trust is an attitude we take toward another person for the purposes of what is entrusted to that person, where trust's distinctiveness lies in one's preparedness to feel certain reactive attitudes. According to Holton, this is a necessary condition of trust and reveals the normative nature of trust.

1.2.9 Developing Holton's View: Obligation and Expectation

Concerns abound regarding the participant stance as necessary for trust. On the one hand, the view is too strong. Philip Nickel (2007, 318) offers the following case. As an instance of therapeutic trust, suppose a businessperson allows a novice assistant to manage a minor account, predicting that he will lose the account. The risk is worthwhile, for the manager, given that the experience will help the novice become more competent and feel

[23] This example is part of Hawley's argument that one can only rely, and not trust, inanimate objects (2019, 16–17). Her argument is representative of a widely held view about trust and reliance. In Chapter 5, I return to this question, especially as it relates to nonhuman animals, artificial technologies, and digital technologies.

more like a part of the organization. For the businessperson, Nickel suggests, "the issue is not sufficiently important, and her predictive expectations of the assistant are sufficiently low, that her reaction to his nonperformance would not rise to the level of resentment or betrayal" (ibid.). While this fails to count as trust on the reactive-attitudes account, it clearly seems like she entrusts the account to the assistant. On the other hand, distinguishing trust from non-trust by virtue of the reactive attitudes involved—that is, whether one enters into the participant stance—rests on a distinction that is secondary to one's trusting. For we will only know whether one was trusting *after* one feels betrayal or gratitude in relying on another. Sometimes this is the case, as when feelings of betrayal and recognition of tacit trust emerges only in retrospect. Yet, this appeal to secondary attitudes is insufficient to explain what trust is. To do so would be, in part, to offer an explanation for why one feels betrayal or gratitude.

Jones (2004) provides a plausible route of reply to this latter objection. Jones' view aims to capture what is shared by "different kinds of three-place trusting relations…and in virtue of which it is correct to call them all *trust* relations" (2004, 5). To accomplish this, Jones argues that reactive attitudes express "normative expectations," which allow us to distinguish trust from the broad class of dependencies on others that resist generalization.[24] Consider a case of therapeutic trust. Jones describes a mother who entrusts her house to her teenage daughter over a weekend. She may lack confidence that the daughter will act in a way that is responsive to the mother's trust. Still, in hopes that her daughter will be responsive, the mother willingly accepts vulnerability to her

[24] This demonstrates a shift from Jones' (1996) affective view of trust. While related to normative expectations and responsivity, Jones argues that an affective view is too restrictive to furnish a generalizable account of trust. Concerning the first condition here, see Simpson (2017) on trust and evidence.

daughter's actions. "The mother might have no *expectation that* the daughter will look after the house well," Jones writes, "the past track record makes such *predictive* expectations unwarranted" (2004, 5; emphasis original). Nevertheless, the mother "does have *normative* expectations *of* the daughter," to which she will respond with reactive attitudes of resentment or gratitude (ibid.; emphasis original). While predictive expectations regard how one thinks someone or something is likely to act, normative expectations concern how we think they *should* act. For Jones, then, trust is "accepted vulnerability to another person's power over something that one cares about, where (1) the trustor forgoes searching (at the time) for ways to reduce such vulnerability, and (2) the trustor maintains normative expectations of the one-trusted that they not use that power to harm what is entrusted" (2004, 6). In this way, forms of reliance that lack normative expectations and the attending reactive attitudes are not trust.

There are two clarifications and a question for Jones' view. First, Jones provides a way to incorporate the insights of predictive views into a normative account of trust. In many cases, especially transactional cases, normative expectations work in tandem with predictive expectations. This is because we navigate the vulnerabilities in those relationships through role expectations. For instance, what Jones calls physician-trust will involve predictions about the competence and reliability of the physician, as well as expectations that physicians are responsive to our vulnerability. This combination of expectations is consistent with cases of therapeutic trust where predictive expectations play no immediate role in our trusting. Viewing trust in this way allows us, Jones argues, to "individuate the relevant shifts in vulnerabilities and grounds according to relevant differences in the kinds of functional virtues required to respond well to such

vulnerabilities" (2004, 6). As Jones (2012) clarifies, these functional virtues need not be *moral* virtues. This leads to a further point.

Second, a normative view about trust is compatible with amoral and immoral forms of trust, provided that normative expectations are present. What matters is that a trustee is responsive to one's vulnerability and normative expectation. In Jones' example of therapeutic trust, the normative expectations are *of* the daughter such that one feels reactive attitudes toward her when trust is fulfilled or broken. This is because the expectations are of the daughter's responsivity to the mother's vulnerability. In turn, this counters Hardin's objection to normative accounts that seem to moralize trust (2002, 75). For instance, when thieves depend on each other in a heist, there are expectations of cooperation, even in the service of immoral ends. When a thief double-crosses his partner, the betrayed thief will likely feel reactive attitudes. At this point, we should wonder what "grounds" our normative expectations of others? To be sure, trust relationships are shaped by and can institute norms and social roles.[25] But to rephrase the question, what explains our expectations and renders them necessarily normative when we trust?

Philip Nickel (2007) argues that reactive attitudes and normative expectations are inherent to the *attitude* of trust because that attitude necessarily involves ascribing obligations to others. According to his Obligation Ascription (OA) Thesis, "if one person

[25] Sociologists were quick to realize this function of trust. See Luhmann (1979), Gambetta (1988), and Sztompka (1999) for representative examples. Their work emphasizes how normative expectations can vary in scope. For instance, one can have general expectations that strangers will act ethically, avoiding unnecessary harms. Alternatively, a trustor may have expectations that arise from a trustee's social role, as with physicians or journalists. Different still, a trustor can expect something of a trustee according to norms for a specific type of relationship, such as in romantic relationships. Across a spectrum of individuals, institutions, and organizations (formal and informal), norms and corresponding expectations shape and are shaped by trust. The question for our present inquiry is whether this normative component is necessary for understanding the nature of trust.

trusts another to do something, then she takes him to be obliged to do that thing" (2007, 310). That is, what distinguishes trust from mere reliance is that we ascribe obligations in the former case but not in the latter. Nickel argues that the insights in Holton's view (and, by extension, Jones' view) reveal that trust involves the ascription of obligations, which explain the connection between reactive attitudes, normative expectations, and trust.

Nickel's account requires a specific conception of obligations. There are two parts. First, when *A* ascribes an obligation to *B* to do *X*, *A* does not regard *B*'s *X*-ing as optional. That is, an obligation to *X* is a *requirement* to *X*—one has "no other intelligible options" (2007, 311). Second, if one is obliged to *X* and does not *X*, then blame and punishment are appropriate. Of course, there may be circumstances that excuse nonperformance—even the most trustworthy can fail to fulfill one's trust due to no fault of their own. But without a "rationalizing explanation" of nonperformance, blame and punishment are appropriate responses, Nickel argues (ibid.). As a result, if trust necessarily involves obligation ascription, then "it is not possible to trust somebody to do what one thinks to be optional or supererogatory" (2007, 310). While we sometimes speak of trusting others to do what is kind, for example, and not strictly required, Nickel argues that this is to conflate "trust" with "rely upon." Instead, he argues that a normative condition on trust allows us to distinguish trust from mere reliance inasmuch as an obligation or obligation-ascription can motivate a trustee to fulfill one's trust.[26] In this

[26] Nickel argues that this is consistent that predictive views in the sense that fulfilling one's obligations can be in one's interests distinctly from other forms of interest. As I discuss below, Nickel distinguishes between the *attitude* of trust and the *grounds* of trust, where predictive information often serves as the latter without illuminating the former. In turn, as a necessary condition on trust, Nickel contends that obligation-ascription characterizes the attitude of trust, while other grounds (affective and predictive) are required for a sufficient account of trust.

way, what explains normative expectations and reactive attitudes is that trust necessarily involves ascribing obligations to a trustee.

The OA Thesis, Nickel contends, provides a necessary but not sufficient condition on trust. That is, ascribing obligations is part of the *attitude* of trust. There are many *grounds* to which the attitude might be responsive, including predictive and normative information about a potential trustee. In this way, affective and predictive conditions on trusting might be individually necessary and jointly sufficient for trusting, depending on the case. However, while seeming pluralistic, Nickel maintains that obligation ascription, as a normative condition, is necessary for trust.

1.2.10 Limitations for Normative Monist Views

There are two sources of objection to this view. First, recall Baier's account of infant trust. Is it that an infant ascribes obligations to its mother? This seems developmentally implausible and, therefore, contributes nothing to an explanation of trust between mother and child. Nickel's reply is that "the connection between trust and obligation *must not* be interpreted as requiring explicit awareness and avowal of an obligation-ascription" (2007, 314). Instead, he argues that we should look for ways that behavior demonstrates ascribed obligations, namely requirement and appropriateness of blame or punishment. Proponents of predictive views, however, can argue that this move is explanatorily unnecessary. For example, Coleman's farming neighbors need not ascribe any obligations to count as trusting, since they might come to rely on each other having seen the advantages that arise from their cooperation.

A second difficulty for Nickel's account is the potential blurring of reliance and trust. Nickel argues that obligation ascription distinguishes cases of reliance from those of trust. Indeed, he argues that it is difficult to imagine a case where someone trusts and does not ascribe some obligation(s) to the trustee. Coleman's farming neighbors seem to provide a ready example. But note, again, that there are cases of obligation ascription that do not involve trust. One such case is reliance. One can rely on another *and* ascribe obligations to the person or thing relied on *without* trusting. For example, I could rely on a physician's word because I take her to have a professional obligation to tell the truth, even if I do not trust her.[27] From this obligation, the physician is required to tell the truth according to her best medical judgment, violation of which is subject to blame and punishment. Despite meeting these conditions, however, I can rely without trusting. So, while trust can institute normative conditions via obligations in particular cases, these seem more like *grounds* for trusting, in Nickel's terms, than cleanly revealing what the *attitude* of trust is.

Approaching normative conditions of trust from the perspective of the trustee raises questions about the sufficiency of normative conditions. Our expectations of others can be presumptuous and, in extreme cases, immoral. To borrow an example from Katherine Hawley, one may be happy for their spouse to predict that she will cook dinner, "but I do not want him to develop normative expectations, to be poised to resent me if I don't" (2019, 15). For this reason, she argues that we "need a story about when

[27] This lack of trust could be because I take trust to involve an affective condition that goes unmet. However, in my view, this is not the only—and not the correct—explanation for relying and ascribing obligations without trusting. Instead, following Nickel, I argue that what distinguishes mere reliance from trust is that trust involves an attitude that disposes one to rely on another. Accordingly, I could rely on the physician and take her to be obligated in a particular way, without having a disposition to rely.

trust, distrust, or neither is objectively appropriate—what is the worldly situation to which (dis)trust is an appropriate response" (2019, 16).[28]

Moreover, Katherine Dormandy (2020) argues that normative forms of trust can be exploitative for both trustor and trustee. In cases of testimony, a speaker both trusts a hearer for recognition and accepts the hearer's trust for information. The speaker can exploit the hearer by lying or adopting unearned epistemic authority. The hearer can exploit a speaker by imposing unsolicited trust or by betraying the speaker's trust for recognition. Accordingly, ascribed obligations, as with normative expectations generally, can place demands on us that are morally objectionable. By way of reply, Nickel clarifies that his OA account does not assume that particular obligations exist, including obligations to do what one is trusted to do (2007, 312). If *A* trusts *B* to do *X*, it does not necessarily follow that *B* actually has an obligation to *X*.[29] For the account to be explanatory, however, obligations or obligation-ascriptions must motivate trusting and trustworthy behavior, in which case Hawley's and Dormandy's objections come to the fore.

Developing the limits of normative conditions on trust is of vital importance to normative conceptions of trust. Nickel's view is that we should see trust as a moral attitude and develop normative conditions in a way that can "adequately account for the moral dimensions of trust" (2007, 318). While a "moral condition" is underdeveloped in

[28] Hawley's own account counters this problem by building an account of trust based on commitment rather than motivation, where the presence of a commitment on the part of the trustee distinguishes cases of mere reliance from trust. While commitments clearly can play a role in trusting, as Hawley shows, her account encounters similar difficulties as other normative accounts—in part because she takes Holton's participant stance to be a necessary condition for trusting.

[29] Moreover, Nickel argues that obligations can exist *de novo* such that there need not be any implicit or explicit agreement before one ascribes an obligation. Once can ascribe an obligation just by thinking that a particular action is appropriate, so long as that 'thinking' meets the two conditions for obligation, namely *requirement* and *preparedness to blame and punish*.

Nickel (2007), Jones (2012) provides a start. Jones argues that trust involves normative expectations that are inextricable from 'rich trustworthiness', whereby someone signals the things for which they are trustworthy. In this way, the richly trustworthy signal "who can count on them for what and so they do not merely turn their backs on poorly placed or presumptuous trust" (ibid., 80). In this way, it could be that moral trust involves norms against objections from presumption or exploitation. Of course, *what* the richly trustworthy person communicates is not neutral. One might communicate that she is a reliable bank robber. Subsequently, an accomplice could ascribe relevant obligations for a heist. Yet, this surely fails to meet the moral condition. Accordingly, for normative accounts of trust to furnish an account of trust under a moral condition, more must be said about how moral evaluations relate to what trust *is* and cases of amoral trust. In the end, Jones argues that we *could* define trustworthiness, and thereby trust, in such a way as to exclude a "brotherhood of thieves" (2012, 85). Human finitude and practical needs for others to be responsive to one's trust militates against this moralizing move, Jones argues. The lingering problem is that the content of norms governing trusting, moral or otherwise, is underdetermined by what trust entails.

This not to deny that some forms of trust are more appropriate than others in particular situations. What is required to differentiate these forms, however, is a normative argument that certain conditions are appropriate in the sense that they *should* have influence in the relevant case(s). In other words, a focus on normative features of trust uncovers the need for an ethics of trust—for an account of what Baier calls "proper trust." For now, I contend that we should distinguish between describing the conditions

that can and could influence trust from those that should, normatively and ethically speaking.

In the end, as with affective accounts and predictive accounts, normative accounts of trust fall short in providing a satisfactory account of what trust is in terms of necessary and sufficient conditions. Clearly, reactive attitudes, predictive and normative expectations, responsiveness to dependence, and obligation reveal important features of trust and trust's social function. Likewise, we have seen ways that a trustee's motivations and different parties' interests can facilitate, prevent, or break trust. A prudent trustor is surely attentive to such conditions depending on the case. The difficulty, however, is determining the demands of prudent trust in particular cases and, if possible, in general. To conclude this section, I raise objections for monism in general, before turning to what I maintain is a way forward for understanding trust and the conditions of trust.

1.3 Two General Problems for Monism

While the preceding sections raise particular problems for different forms of monism, they reveal two general problems for monism as a strategy for analyzing trust. First, since monist approaches take certain conditions of trust to be necessary, counterexamples abound, many of which arise from rival approaches. Specifically, the availability of counterexamples undermines the monist ambition to provide a general account of trust. An immediate means of reply is to distinguish between competing views on the basis of trustworthiness. But appealing to trustworthiness or to ideal forms of trust in the relevant domain risks begging the question, since, as I have argued, one's conception of trustworthiness is connected to how one thinks about trust. I take the ease in generating

counterexamples to provide strong inductive evidence against the plausibility of analyses of trust in terms of necessary and sufficient conditions.[30]

Second, monist approaches face an explanatory problem. Following Hardin (2006), Philip Nickel (2017) proposes a two-part constraint on theorizing about trust. First, trust should be "explained as the outcome of central concerns or interests of the relevant actors" (2017, 197). Second, trust should "explain the emergence and sustenance of cooperative practices and social institutions" (ibid.). The former he calls the input condition, while the latter provides an output condition. The plausibility of the two-part constraint is that trust is of little interest if it contributes nothing to an explanation of how the interests of interacting parties impact social practices.

The two-part constraint presents a dilemma for monist views. Consider the following case. Paula is traveling in an unknown city and seeks directions from a stranger. Paula knows nothing of the stranger's reliability, nor does she know whether the stranger bears her goodwill. Yet, suppose that Paula is disposed to rely on the stranger's directions. In my view, it seems plausible that Paula trusts the stranger for directions. Here's the dilemma for understanding the explanatory role of trust, viewed in monist lights, in such a case. The first horn is to suppose that, despite a disposition to rely on the stranger, Paula does not really trust. If so, then trust plays no explanatory role in Paula and stranger's cooperation, meaning that trust's role in explaining the emergence and sustenance of cooperative practices and institutions (second part of the constraint) is quite limited. But suppose one thinks this *is* a case of trust (in the monist sense). Taking this

[30] Simpson (2012) likewise argues that counterexamples present *inductive* evidence against analyses of trust in the form of 'trust is *this*', where *this* entails providing necessary conditions. See also discussions in Nickel (2017), Keren (2020), McLeod (2015), and Goldberg (2020).

horn requires the implausible view that people in such unpredictable situations always attribute the relevant features to trustees, such as good will, predictions that the trustee is reliable, and so on.

1.4 Toward Pluralism about Trust

A possible rejoinder to the two general problems for monism is to note that weaker monist views can acknowledge different forms of trust. For instance, one might argue that affective instances of trust are more appropriate to intimate interpersonal relationships, whereas those conditions can lead to harm and manipulation in contractual or transactional contexts. Therefore, predictive conditions are more relevant to the latter case, but not the former. These weaker monist views could capture the contextual flexibility that trust demands.

I wish to highlight two issues with this response. First, monist views, even weak views, cannot adjudicate persistent disagreements about the conditions of trust *within cases.* As I argued with the physician case, different monist views can render different judgments for the same relationship or context. Accordingly, it is not enough for an account to acknowledge that the conditions of trust vary across contexts—as they clearly do—but we should also explain why the conditions of trust can vary within the same context. Without such an explanation, an analysis of trust ends in dialectical stalemate.

Second, adjudicating appropriate conditions on trust within and across situations raises a *normative* problem for theorizing about trust. Imagine a case where a patient feels that her dentist has broken her trust, say after a friend reveals that the dentist in general disdains his patients. Yet, he works tirelessly at his practice, both out of a love for

dentistry and financial need. When the dentist is alerted to feelings of broken trust, he might reasonably respond that he is trustworthy with respect to dentistry. Such conflict could initiate a negotiation between trustor and trustee about the appropriate conditions and expectations within the trust domain (see Stewart 2022). We may be able to resolve such negotiations in several ways. The patient could continue seeing the dentist without trusting, staying as a matter of mere reliance. Alternatively, she might come to trust, for instance, if the dentist expresses that he does care for his patients after all. Or the patient might seek care elsewhere. What is important to see about the disagreement, however, is that we cannot preserve trust without appealing to *ideal* or *appropriate* conditions. To determine whether trust is appropriate requires consideration of what the conditions of trust *should* be, even if this is independent of those conditions which do or could influence trust.

Baier proposes an "expressibility test" for assessing "the moral decency of a trust relationship" (1994, 123). The test is to express "what the other party [is] relying on for the continuance of the trust relationship," or in other words, to clarify the conditions of one's trust (ibid.). If the relationship can continue after expressing one's basis for trusting, Baier argues, then it is morally decent—or at least we are on our way to determining its moral decency. One way to advance this point is to follow Holton and proponents of normative views in thinking that trust inherently institutes norms and obligations. But the fact that trust relationships can and do institute obligations in some cases is not enough to override the dentist's reply in the previous example, since the nature of those obligations is not the same in every case. Rather, the obligations and conditions instituted in trust relationships are, as it were, downstream of the conditions of

trust. Accordingly, expressibility may have little to say if there is *disagreement* about the appropriate grounds for trust. Put differently, in the absence of a normative argument that dismisses rival views of trust in a particular context, both problems for monism—the counterexample problem and the explanatory problem—underscore monism's failure as a strategy for analyzing trust.

In turn, the challenge for a plausibly weak form of monism is to explain the variance in cases of trust without forfeiting the insights of stronger forms of monism. In my view, this urges us in a pluralistic direction, where we can recognize multiple forms of trust. How is trust pluralistic? Without monism, is it possible to connect various forms of trust *as* forms of the same thing? If trust differs across contexts, how might pluralism about trust help address disagreements, like the dentist case? It is to these challenges that I turn in Chapter 2.

2.0 PRAGMATIC PLURALISM

In this chapter, I develop a pluralistic account of trust in connection with the two general problems for monism discussed in chapter one. I call my approach *pragmatic pluralism*. My argument proceeds in two phases. The first phase argues for a version of pluralism about trust. As a *descriptive* view, pluralism demarcates forms of trust. In doing so, the view aims to incorporate insights from monist views, while avoiding problems arising from counterexamples. I argue, however, that pluralism fails to address the second, explanatory problem for monism. In turn, the second phase of my view develops a pragmatist approach to pluralism that focuses on trust's function within and across contexts. Pragmatism is an *explanatory* strategy for uniting the plural forms of trust.[31] My view joins with recent function-first approaches to concepts and practices. Such approaches eschew traditional, monistic analyses in terms of necessary and sufficient conditions in favor of analyses that focus on the role or function of a concept, term, or practice. This strategy is not intended to relegate traditional analyses *in toto*, but rather to provide a means for investigating cases that are, as Matthieu Queloz describes them, "internally diverse" (2021, 25). One can be a pragmatist without being a pluralist, or a pluralist without being a pragmatist. I contend that the two working in tandem provide the most plausible approach to trust.

[31] That is, 'pragmatism' here is not a view about truth, meaning, expediency, or else besides. Rather, it is a methodological approach to understanding concepts and social practices by focusing on their function.

Before proceeding, a distinction is crucial for approaching pragmatic pluralism. In his argument for obligation ascription, Nickel (2007, 312) distinguishes between the *attitude* of trust and the *ground* of that attitude. He writes: "An attitude such as trust is accounted for, minimally, by its characteristic functional role in human behavior in cognition" (2007, 312). In contrast, the ground of "an attitude consists of the reasons that characteristically support or rationalize the psychological attitude" (2007, 312). In the previous chapter, we saw various competing grounds for trust, ranging from goodwill to predictability to responsiveness to obligations and norms. I see these as the possible grounds of trust. The pluralist insight is that there are many possible grounds for trust—indeed, many more than those discussed in the previous section. What, then, is the attitude of trust? I contend that looking to trust's function supplies an answer. Specifically, I argue that trust's function in disposing trustors to rely on trustees can unite the plural grounds of trust. In other words, according to pragmatic pluralism, trust is a psychological disposition to rely on another according to, as Nickel puts it, "the reasons that characteristically support or rationalize the psychological attitude." It is by virtue of this function that well-placed trust is valuable and central to cooperation.

2.1 PLURALISM(S) ABOUT TRUST

Pluralist accounts of trust begin from the insight that trust is a family resemblance concept, which, as Simpson argues, emerges from the "domestic life of child-rearing and shared company, of exchange, and of joint, positive and negative collective action" (2012, 557). What Simpson calls "Ur-trust" indicates how reliance on the cooperative

behavior of others is central to social forms of life. Over time, Ur-trust acquires new and richer resonances, both sentimental and normative. In this way, what trust means shifts across contexts according to developments around cooperative behavior within those contexts, rendering trust resistant to monistic analyses.

For our purposes, attending to varieties of trust provides a way to avoid the counterexample problem for analyses of trust. That is, pluralism does not require that there be necessary and sufficient conditions for trusting over time and across contexts. Rather, trust can come in many forms, allowing us to acknowledge, for example, that there can be trust between business partners that differs from trust between loved ones *without* denying that either is fully trust. Yet, this does not tell us *how* to recognize various forms of trust. There are two points to make for developing pluralism.

First, at the heart of pluralism, I argue, is a distinction between describing the possible conditions of trust and a normative evaluation of those conditions. This is clearest in the drift in monist accounts from claims about what trust *is* to claims about genuine or appropriate forms of trust. For example, Baier's infant trust begins from a plausible description of trust in intimate relationships before *shifting* to an evaluation of forms of trust as more or less *normatively* legitimate. For the pluralist, however, describing the occurrent conditions of trust need not imply a view about the normatively appropriate conditions of trust. For instance, one can acknowledge that trust can exist between conmen without thereby condoning such trust. Rather, an aim of pluralism is to describe possible forms of trust wherever they may occur.

Second, philosophical and empirical literature on trust is replete with distinctions between forms of trust, including social trust, epistemic trust, general trust, public trust,

testimonial trust, organizational trust, political trust, legal trust, aesthetic trust, e-trust, and more besides. While acknowledgments of *forms* of trust are widespread, it can be challenging to determine what is inherently pluralistic about trust. So, in the next subsections, I examine five candidates for developing a pluralistic categorization of trust—I label them *attitudinal*, *agential*, *vulnerabilities, aims*, and *conditions* pluralisms respectively. I argue that the last of these is most helpful for tracking trust across contexts, drawing on distinctions and insights from monistic views. Nonetheless, alternative versions of pluralism highlight features of trust relationships whose variance across contexts can impact trust. For example, in my view, any view of trust should speak to the role of varied history, power dynamics, demographics, and geography in trust relationships. In other words, I think attending to the attitudes, agents, vulnerabilities, and aims of those in (or possibly in) trust relationships can point us to salient features that shape those relationships. The question for this section, however, is how best to distinguish forms of trust.

2.1.1 Attitudinal Pluralism

The first candidate for developing pluralism about trust I call *attitudinal pluralism*. Attitudinal pluralism distinguishes forms of trust by the attitudes involved in trusting. For example, while Baier defends an affective account of trust, she argues that trust involves a *belief* about the goodwill of a trustee. In contrast, Holton argues there are significant dissimilarities between belief and reliance such that belief is not a necessary condition for trusting. According to the attitudinal pluralist, we can distinguish the respective views by

attending to the diversity of attitudes relevant for trusting. To see how this view might work, consider Simpson (2012, 564):

> 'trust' may be felicitously used to describe mental states that result in dispositions to rely on cooperative behaviour, as well as actual instances of reliance. The sorts of mental state that may lead people to have a disposition to rely on others includes beliefs about what will lead the other party to be trustworthy. But there is no reason to suppose that it is restricted to beliefs. Loving someone may prompt a disposition to trust, and a very robust one at that, often surviving despite evidence of untrustworthiness. And this is what we observe…['trust' can] refer to cognitive, conative and affective mental states. All of these may be significant in issuing in a disposition to trust; so all of these are felicitously described as trust. Call beliefs which lead to a disposition to trust, *cognitive trust*. Call judgments, decisions, intentions and resolutions which lead to a disposition to trust, *conative trust*. Call emotional states which lead to a disposition to trust, *affective trust*.

Simpson adds to these forms of trust what he calls *predictive trust*, which "involves nothing more than a prediction of reliability" (ibid., 565).[32] He argues that "repeated use has hardened these [forms of trust] into discrete notions" such that we need not qualify them adjectivally (ibid.). And he clarifies that he does not purport to "have identified all the forms of trust; those that I have noted are merely the most obvious forms" (ibid.). In this way, there is something social and axiological at the heart of Simpson's pluralism, particularly in the way that forms of trust emerge from practices of reliance and dispositions to rely. That is, forms of trust arise in and are sensitive to what we value in social living. For pluralism, it is important to underscore here that identifying different paradigmatic attitudes allows us to distinguish forms of trust.

[32] Predictive trust describes trust where there is "no expectation that the trusted may take account of me in their action" (ibid.). Accordingly, predictive trust provides a possible way to account for trust in inanimate objects and systems. I agree with Simpson that an account of trust should explain how the term can be applied to such objects. In Chapter Four, I argue that pragmatic pluralism clarifies disputes around trust in digital technologies, especially where artificial intelligence is involved.

In large part, the allure of attitudinal pluralism arises from different approaches to trust in the philosophical literature. For example, the cognitivism debate—concerning whether belief constitutes or is entailed by trust—assumes an attitudinal basis for determining the necessary conditions of trust.[33] This allows us to distinguish different forms of trust and approaches to those forms. For instance, Arnon Keren (2020) distinguishes doxastic and nondoxastic accounts of trust. Doxastic accounts of trust, Keren suggests, maintain that trust consists in or is entailed by a belief. What he calls "pure doxastic accounts," which include Hardin's predictive view, argue that trust simply is a matter of belief (2020, 109). As Hardin remarks, "'I believe you are trustworthy' and 'I trust you' are equivalent" (2002, 10). In contrast, impure doxastic accounts hold that belief is a necessary condition for trust, though not sufficient.[34] Non-doxastic accounts of trust, on the other hand, reject the claim that trust entails belief. This is rendered most clearly by cases of therapeutic trust, wherein one's trust is not based on a belief about the trustworthiness of the trustee. Moreover, recall that disanalogies between trust and belief lie at the heart of Holton's normative view. While belief might have some role to play in explaining why someone trusts, it is a moral stance (Holton's view) or an emotive or affective attitude (Jones (1996); see also McLeod (2015)).

In the cognitivism debate, the promise of attitudinal pluralism is that it allows us to acknowledge the insights of different views about trust without facing the two problems for monism. For instance, the attitudinal pluralist recognizes that sometimes love plays an important role in grounding trust. However, we need not think this is apt in

[33] See Simpson (2018) and Keren (2020) for good overviews of this debate.
[34] See Keren (2014). For other doxastic views, see Baier (1994), Adler (1994), Fricker (2006), Hieronymi (2008), and McMyler (2011).

every context. In business transactions, for example, believing that another is honest and reliable is often sufficient for trusting. In this way, attitudinal pluralism can distinguish forms of trust, retaining insights from the diversity of views discussed in Chapter One.

2.1.2 Limitations of Attitudinal Pluralism

The principal problem with attitudinal pluralism is not that it fails to illuminate salient features of trust relationships, but whether it satisfactorily distinguishes forms of trust. Belief, acceptance, intention, desire, and many other attitudes surely have a place in an investigation of trust and trust relationships. Consider therapeutic trust again. The fact that therapeutic trust does not require a belief about the trustworthiness of a trustee is relevant for developing an account of trust that can handle therapeutic cases, as well as an account that can contrast various forms of trust. My objection to attitudinal pluralism is that it fails to track salient differences between forms of trust.

Consider Baier's and Hardin's views. Both Baier and Hardin maintain that trust involves a belief. They are cognitivists about trust. What divides their views, however, is what the belief is *about*—namely the difference between affective and predictive conditions on trust. For this reason, what distinguishes their views is not whether a trustor follows evidence of a trustee's trustworthiness, but which types of evidence are utilized in navigating trust relationships.[35] That predictive (and cognitive views generally) are dissatisfying to proponents of affective views is not simply a matter of the attitudes

[35] The inclusion of belief as a necessary condition does contribute to one's understanding of the requirements of trust, however. It might be argued that an attitude wherein one is vulnerable to another's (in)action precludes cognitive attitudes like belief. This is because when one trusts, one is not seeking evidence about the trustee's trustworthiness. However, as Simpson (2017) argues, there is an important difference between following evidence (i.e., being sensitive to one's evidence) and gathering evidence. Only the latter is inconsistent with trust.

involved. Rather, *that* certain attitudes are appropriate to trust alters what they think trust is. Moreover, for normative views, one can acquire beliefs through trust, but the psychological profile is very different—since one is unlikely to feel betrayed when the proposition one believes turns out false, unlike when someone breaks trust. For trust influences and is influenced by what one believes, hopes for, expects, intends to do, and so on. To distinguish the salient differences between forms of trust, however, we must attend to the content of those attitudes and how trust can be conditional on that content.

2.1.3 VULNERABILITIES PLURALISM

To "get any sense of the variety of forms of trust," Baier argues, we must look "both at varieties of vulnerability and at varieties of grounds for not expecting others to take advantage of it" (1994, 100). While she conceives of trust as belief, Baier explains that "[w]here one depends on another's goodwill, one is necessarily vulnerable to the limits of that goodwill" (1994, 99). Developing this point, Jones (2004, 6) distinguishes forms of trust according to the vulnerabilities that arise through relationships of dependence. For example, in friendships, trust will introduce "personal vulnerabilities of various kinds, accepted on the basis of goodwill. But other cases are covered, also–trust in business transactions, in professionals, postal carriers, and plumbers…" (ibid.). In this way, we could identify forms of trust by the vulnerabilities incurred by relying on others. I call this view *vulnerabilities pluralism*.

In Jones' argument, identifying the forms of trust links trust to virtues that render another trustworthy. She contends that we can "individuate the relevant shifts in vulnerabilities and grounds according to relevant differences in the kinds of functional

virtues required to respond well to such vulnerabilities" (ibid.). In short, the idea is that subvarieties of trust have functional virtues unique to the relevant varieties. Jones' argument is that vulnerabilities incurred in trusting allow us to classify forms of trust "according to the functional virtues required to respond appropriately to them" (ibid., 7). For example, when someone fails to fulfill role-specific trust, either through professional misconduct or incompetence, we can suppose that the trust breaker lacks certain functional virtues, such as conscientiousness and self-knowledge. As Jones argues, the virtuous trustee must "neither overestimate nor underestimate their competence lest they be…timid in exercising discretionary power or…practice outside their ability" (ibid.). In contrast, consider how vulnerabilities differ in intimate relationships. Jones argues that vulnerabilities in such cases are typically accepted "on the basis of assumptions about the other's care or love for us" (ibid.). Virtuous trust in such cases can involve proper regard for the autonomy of the dependent other, as in cases of children and parents.

I think Baier and Jones are exactly right to look for functional virtues for forms of trust. This is not to assert that such virtues exist or that trust is a virtue.[36] Rather, the point is that we should look for normative ways to distinguish possible forms of trust, especially within cases, as we saw with the dentist case at the conclusion of Chapter One. But there is an important distinction between descriptive pluralism and normative pluralism. After all, it is possible that there are many forms of trust, while virtuous trust is the same across contexts. Alternatively, it may be that a form of normative pluralism is true, where different normative requirements apply to different forms of trust. My present

[36] Although, I think virtue ethical approaches are most promising for developing an ethics of trust. For views in this vein, see Potter (2002) and Carter (forthcoming).

interest is whether vulnerabilities are the best means for "classifying varieties of trust," as Jones suggests (ibid.). I think a negative answer is correct.

2.1.4 Limitations for Vulnerabilities Pluralism

There are at least three problems facing vulnerabilities pluralism. First, as Hawley (2019, 7–8) argues, it is not clear that trust is necessarily linked to vulnerability. To be sure, relying on others can introduce vulnerability, and trust seems especially important in cases where one cannot avoid vulnerability or where one's vulnerability is especially great. But there are cases where one is not vulnerable in trusting. For example, Sam could trust Kate to bring food to a picnic, while bringing enough food such that Kate's failure to bring food has no great impact. Sam might be motivated to do this out of generosity. Does that mean he does not trust Kate to bring food? It seems to me that Sam can trust Kate to bring food, even if he is not vulnerable in so doing. The problem in this case is that vulnerabilities do not seem to capture the possible differences in Sam's trust; for instance, the difference between ascribing obligations and making a prediction about how Kate is likely to act. In the former case, it may be that Kate's obligations are more important when vulnerabilities are greater. But it's not clear why this is essential for distinguishing forms of trust.

Second, variance in vulnerability across cases leave the problem of counterexamples unaddressed. Consider again the role-specific trust example. Two people can be differentially vulnerable to the professional misconduct of an expert or supervisor. For instance, one may have the wealth to weather being fired, while the other does not. Jones mentions that one accepts vulnerability to the actions of another under

assumptions about "professional competence and about the well-functioning of institutional structures of training, accreditation, and audit" (ibid.). Even if such structures serve trustors equally (which is doubtful), it seems that different people will have different expectations on the basis of things other than vulnerabilities. For instance, one may accept vulnerability in home repairs from her plumber not on the basis of conscientiousness or sufficient self-knowledge, but because of some other feature, such as the plumber's reputation or prior relationship with the homeowner. Is such trust vicious? It could be, but this seems afield of classifying forms of trust.

Third, there can be a difference in degree of vulnerability across forms of trust. As Jones is quick to point out, "we don't want to proliferate varieties of trusts according to just any difference in vulnerability or grounds" (2004, 6). Rather, she argues, we should "individuate the relevant shifts in vulnerabilities and grounds according to relevant differences in the kinds of functional virtues required to respond well to such vulnerabilities" (ibid.). I have already noted the subtle shift from discerning possible forms of trust to adjudicating whether one trusts well in a particular case. But we should also wonder whether vulnerabilities are sufficiently similar to distinguish forms of trust. For example, within a professional context, I could be extremely vulnerable to one colleague but not another. Likewise, I could be extremely vulnerable to a business partner, while being relatively invulnerable to my best friend. In addition to degree, vulnerabilities can differ in kind. I could be physically vulnerable when depending on another, say for safe transport, whereas I could be psychologically vulnerable to a close friend or someone online.

Recall that Baier's original formulation of the strategy is a conjunction, whereby one determines the varieties of trust from (1) vulnerabilities incurred in depending on others and (2) "the varieties of grounds for not expecting others to take advantage of [one's dependence]" (1994, 100). Jones' view develops the former but not the latter in classifying forms of trust. While I argue that this fails to facilitate pluralism, vulnerabilities can help us to orient the grounds for expecting others not to take advantage of our trust. In many cases, this will concern the stakes involved in our dependence on others. If my life depends on you coming through for me, my vulnerabilities are great such that the grounds of my trust should be appropriately great. What are the appropriate grounds? I will return to this question below to argue that we should see these as the conditions of trust in a particular case. With that in mind, there is another form of pluralism derivable from Baier and Jones' approach.

2.1.5 AGENTIAL PLURALISM

What I call *agential pluralism* distinguishes forms of trust according to the agents involved, including salient information about the social standing of the agents. I use 'agents' in a broad sense, inclusive of individuals, groups, and organizations. Interpersonal trust, for example, involves trust between persons or individuals, such as between a patient and her physician, a mother and child, business partners, spouses, and so on. As Jones argues, such differences allow us to distinguish between various relational forms of trust (2004, 6). In contrast to interpersonal trust, one might have a high degree of social or general trust, thinking that others (i.e., a set of people) are trustworthy, at least for the most part. Similarly, we often talk of groups trusting

individuals, organizations, and institutions. The president's approval rating among citizens could be taken as a measure of how much the general public trusts the president (or the president's cabinet). The president could trust a subset of his constituents. Individuals as well as groups can trust agencies, such as the Federal Bureau of Investigation or the Food and Drug Administration in the United States. Similarly, it is common to speak of trust in institutions, like the family, marriage, or higher education. For the forms of monism in the previous section, we can distinguish infant trust, contractual or predictive trust, as well as the norms involved between the agents involved, to develop taxonomy of trust types.

Through this process, we can derive various social, epistemic, legal, and political forms of trust.[37] Recently, discussions of e-trust—that is, trust designed for digital environments, especially involving artificial agents—distinguish forms of trust based on the agents involved.[38] Indeed, understanding e-trust is at the forefront of human-computer interaction (HCI) research. For example, Pepijn Al (2022) investigates trust in artificial intelligence, in part, by contrasting trust between humans with trust between humans and technology.[39] These examples and developments are not exhaustive of distinctions in trust based on the agents involved. The strategy for agential pluralism, however, is that we can distinguish various forms of trust by identifying the agents involved and their relationship.

[37] These forms of trust vary considerably, including: political trust (Levi and Stoker 2000), legal trust (Bergh, Bjørnskov, Vallier 2021), aesthetic trust (Nguyen 2021), medical trust (Nickel and Frank 2020), and more besides. See Vallier and Weber (2021) for more on social trust.
[38] See Taddeo (2009).
[39] For more on trust in technology, including AI, see Nickel (2011, 2013); Ferrario, Loi, and Eleonora Viganò (2020); Ryan (2020); Nickel, Franssen, and Kroes (2010); Braun, Bleher, and Hummel (2021).

2.1.6 Limitations for Agential Pluralism

An immediate hurdle for agential pluralism is that it requires controversial assumptions about the nature and relationships of individuals to groups, organizations, and institutions.[40] This raises the following sort of questions: to trust a group, must one trust each member of the group? If not, how do representatives of groups mediate cases of trust, especially public trust? For HCI, does human-AI trust reduce to human-human trust in the sense that trusting AI amounts to trusting the designers of AI? Or is there a distinct sense in which one can trust AI but not the designers of AI? It is hard to see how answers to these questions are unnecessary for distinguishing forms of trust based on the agents involved. Worse still, it is difficult to see how such an approach is necessary for understanding the salient differences in trust between children and parents, between adults, in-group and out-group members, contractors and homeowners, citizens and governments, and people and institutions. Moreover, one's trust can differ for the same agent, depending on context—as when declining to trust your physician to repair your car, or forgoing heart surgery from a mechanic.

Nevertheless, distinguishing features of trust according to the agents involved is nontrivial. In the context of public trust in science, for example, Gabriele Contessa (2022) argues for viewing trust as a part of an epistemic division of labor. On his view, trust is directly sensitive to what he calls the "socio-epistemic infrastructure" of society (ibid., 17). The public's trust in science is not a matter of individuals trusting science as an institution or individual scientists. Rather, a society trusts science when it "collectively relies on science to inform its actions and decisions (and those of its members)" (ibid.,

[40] For example, see Lackey (2020).

16). To neglect the social environment in which trust relationships are formed and maintained is to risk missing evidence of trust's function and value, to Contessa's point. Yet, agential pluralism is ill-suited for the purposes of distinguishing plural forms of trust.

2.1.7 AIMS PLURALISM

A fourth alternative arises from one's aims when trusting, what I call *aims pluralism*. Aims pluralism distinguishes forms of trust according to what one trusts another *for*. For instance, epistemic trust involves trusting others for knowledge or a basis for reasonable belief.[41] When O'Neill's patient trusts her physician, at least in part, she trusts the physician to provide accurate information. We can trust others in order that some action or activity is accomplished. For instance, I can trust a reliable mechanic to have my car repaired correctly. Furthermore, this practical goal of my trust is distinguishable from forms of prosocial trust, which aim at building relationship and community. For example, someone might trust me to bring dessert to a dinner party because the host aims for me to feel included *by* being entrusted with dessert—and not because my desserts are especially desirable. With therapeutic trust, one can entrust something to another without any expectation that the trustee will fulfill one's trust. Rather, the parent's aim in trusting—namely, to cultivate a certain type of character in her child—distinguishes therapeutic trust from alternative forms of trust.

Consider again Jones' agential view (as distinct from her vulnerabilities view). Rather than distinguishing forms of trust according to the agents involved, we instead

[41] See Faulkner (2011).

identify the motivational element in trusting within those specific cases. For instance, distinguishing trust in a plumber from trust in a loved one is not merely that there are different agents involved or that our attitudes toward the agents are different, but instead that we have different aims in the cases; in the former, we could be aiming to have the sink fixed, whereas we might be seeking life-long companionship in the latter.[42]

So, an advantage of aims pluralism is that it directs our attention to ways that forms of trust are differentially sensitive to possible outcomes of trust relationships. Consider Baier's three-placed analysis of trust. Tracing the idea to Locke, Baier argues that an analysis of '*A* trusts *B* with *C*' reveals not only *whom* we trust but *what* we entrust to them.[43] In this way, Baier analyzes trust within a "model of entrusting," allowing us to discern "different forms of trust by the different valued goods we confidently allow others to have some control over" (1994, 101). From the point of view of aims pluralism, attention to what is entrusted to others distinguishes forms of trust through one's aims or intentions (rather than the actual goods entrusted), whether that be to acquire or preserve knowledge, money, health, or friendship.

2.1.8 Limitations for Aims Pluralism

I raise two problems for developing pluralism from one's aims in trusting. First, attention to aims can fail to distinguish seemingly different forms of trust. Consider two patients, patient_a and patient_p who meet with a physician following testing. Both patients aim to receive an expert opinion of the results. But patient_a trusts according to whether goodwill

[42] Of course, *that* these are the appropriate aims one *should* have in the relevant case is a separate matter.
[43] Baier is quick to note that accepting an entrusting analysis of trust can distort some cases, especially those where there is no clear candidate for *C*; however, she argues that the analysis "will prove more of a help than a hindrance" (1994, 101–2).

and care motivate the physician's reading of the evidence. Patient_p, in contrast, trusts on the basis of the physician's professional competence. Here, aims pluralism fails to distinguish two plausibly distinct forms of trust. At the same time, the view's incompleteness should not lead us to neglect the role of aims and motivations in trusting. In this two-patient example, for instance, understanding a person's aims can help alert us to their expectations and values in trusting.

Second, there are cases where one may have no definite aim in trusting. Although Baier, Hardin, and Holton conceive of trust as a three-placed relation (*A* trusts *B* to/for *C*), this is not uncontroversial. There are cases of two-placed trust, where simply *A* trusts *B*.[44] As Simpson explains, two-placed trust "does not entail that A does or should trust B over all potential trust relations that could arise; rather, it signifies that there is a particular kind of relationship between A and B, namely a trusting one" (2023, 85). It may be that there is nothing specifically that I trust another for, having only a vague and undefined range of aims in trusting. Rather, one way to conceive of my trust (in the two-placed sense) is that it indicates a particular depth to a relationship, whether that be interpersonal, transactional, or else besides. Settling the relationship between two-placed and three-placed trust is controversial and beyond our present purview.[45]

This second problem for aims pluralism is twofold. One the one hand, reckoning how aims distinguish two- and three-placed relationships could tie the view to

[44] Faulkner (2015) even argues that a one-placed relation—'X is trusting'—is basic.
[45] For my part, the aptness of two-placed trust is a difference in degree rather than kind. I trust my spouse deeply. Clearly, though, upon investigation, I trust her for some things and not others, even if I cannot articulate all the ways that I might trust her in the three-placed sense. That my trust can be analyzed in terms of three-placed trust does not undermine the felicity of my saying 'I trust you'. More controversially, I maintain (but do not defend here) that '*A* trusts' or '*A* trusts *B*' are, for the purposes of analysis, incomplete predicates; that is, in each case, we could determine that *A* trusts *B* to *C* (and in context *D*) for each case. For alternative views, see Simpson (2023; forthcoming), Domenicucci and Holton (2017), and Faulkner (2015).

controversial, unsettled theoretical disputes about the nature of trust. On the other hand, if we can identify what two-placed trust aims at abstractly or in general, the distinction between forms of trust seems to rest more on the nature of the relationship between parties—whether through their attitudes about each other or something about their status as agents—than on the role of aims within the relationship.

In the end, just as vulnerabilities are salient for understanding the grounds of one's trust, so too are one's aims and the impact of outcomes on trusting. One's aims in trusting can help to identify what we value in the trust relationship and, correspondingly, what grounds of trust we think are appropriate.

2.1.9 CONDITIONS PLURALISM

The final approach to pluralism for consideration, and the one I prefer, I call *conditions pluralism*. With monism, we examined how attempts to define the nature of trust by enumerating necessary and sufficient conditions fail. Despite this, each form of monism captures salient conditions for trusting in relevant cases. Put differently, for the cases that a view takes as paradigmatic, the purported conditions are plausible. The problem is that counterexamples undermine each form of monism *as* an account of what trust is across cases. The idea behind conditions pluralism is to distinguish forms of trust according to the conditions on one's trusting, or according to trust's *multi-dimensionality*.[46] Across the spectrum of situations where trust is apt, there are dimensions which help trustors identify trustworthy trustees, including considerations of competence, benevolence, integrity,

[46] The term "multidimensionality" is derived from PytlikZillig et al.'s (2016) empirical investigations of trust, to which I turn in the next chapter. While I utilize the term in a slightly more restrictive sense in this section, my aim in the next chapter is to show how conditions pluralism is both consistent with and helpful for empirical investigations of trust.

shared values, and more besides. These dimensions, also sometimes called indicators (Branch and Origgi 2022) or antecedents (Siegrist and Zingg 2014) of trust, are what I mean by "conditions" of trust.

Before proceeding, there is an important rejoinder for the forms of pluralism discussed in previous sections. My argument for conditions pluralism is not that attitudes, vulnerabilities, agents, or aims are irrelevant for discerning salient features of different forms of trust. Rather, I think each view is insufficiently attentive to important distinctions if taken independently. Instead, I think a plausible way to identify plural forms of trust is to consider how different aspects of trust relationships can serve as conditions on one's trusting or, to put it differently, as *grounds* for the attitude of trust. As the grounds or conditions vary, so too do the forms of trust. This insight, I argue below, directs us in the direction of pragmatism. In this section, I formulate conditions pluralism in two steps. First, I relate the conditions of trust to the attitude of trust and the consequences of trusting. Second, following an objection, I suggest that pluralism about trust should be theoretically (but not empirically) unrestricted.

We can distinguish the attitude of trust, the conditions upon which one is prepared to rely on another, and the consequences of trusting.[47] "In contrast to trust *per se*," Lisa PytlikZillig et al. suggest that "trustworthiness refers to beliefs, evaluations, or expectancies of the target that are often theorised to form the basis for trust" (2016, 114). Again, it is tempting to think that the basis of trust (i.e., indicators of trust*worthiness*) is invariant and explanatorily prior to the attitude of trust (or trust *per se*, in PytlikZillig et al.'s terms), but I have argued that this is mistaken. Instead, one positive upshot of the

[47] I only flag here the consequences of trust as an output of trusting another based on some conditions. I return to this point in the next section with considerations of trust's function.

examination of monism in the first chapter is that trust can be formed on multiple bases. In other words, we can view the conditions of trust as the grounds for trust inasmuch as they help one identify relevant information for coming to trust another within a context. One need not be conscious of nor have control over the conditions that influence one's dependence on others.[48] For patient$_a$ in the previous section, who trusts her physician under conditions of care and goodwill, it may be that she can recognize the reasonableness of someone who trusts on the basis of predictive information alone, while nevertheless remaining unable to trust in the absence of care and goodwill. Likewise, she might see the advantage in trusting a supremely skilled physician who cares little for his patients' well-being, while remaining unable to trust that physician. This bears similarity to those who cannot trust after trust has been broken, however advantageous trust might be.

Recall the distinction between conditions *of* and *for* trust from Chapter One. The conditions *of* trust are intimately tied to trust in the sense that they help one determine what the demands of trust are in a particular context. The conditions *for* trust, in contrast, denote features of the context in which trust emerges, persists, or breaks. The conditions of trust can be sensitive to the conditions for trust. For example, Maya Goldenberg (2020) argues that trust plays an ineliminable role for understanding public hesitancy about vaccines. As she shows, one's social standing and relationship to authorities, one's education and beliefs about vaccine technologies, and one's values and aims, such as protecting public health or preserving individual liberty, all impact public trust. Indeed, they help to explain why someone is hesitant or compliant with respect to vaccination. In

[48] Baier (1994, 99, 105) asserts this point. For a convincing argument on this score, see McMyler (2017).

turn, attention to trustor and trustee self-understanding, the social and environmental context, and the actions that follow from trusting can all influence and help to identify the conditions that distinguish forms of trust.

Another way to home in on the conditions *of* trust is to consider how one can withhold trust while thinking a possible trustee is trustworthy. Suppose one requires surgery. There are many factors that could influence one's trust, including the surgeon's competence and the environment in which the surgeon operates. Imagine a competent surgeon who finds herself in a context where crucial resources are scarce. In such a case, a patient could reasonably forgo surgery, at least in non-emergency cases, and think that the surgeon is nevertheless trustworthy. Alternatively, imagine the case of a surgeon who has developed a reputation for incompetence or bad luck. In a resource-rich context, one could reasonably think that the surgeon is untrustworthy even if he operates within a context where he could be reliable.[49] In the first case, the refusal arises not from thinking the surgeon is untrustworthy, but from features of the context which thwart the surgeon's otherwise present reliability. These are conditions for trust. In the latter case, the surgeon fails to meet a condition *of* trust, namely competence, while the conditions *for* trust are met. To be sure, the distinction between conditions for and of trust is not absolute; a condition of trust can interact with conditions for trust. For instance, oversight committees and sanctions for incompetent surgeons could institute a system wherein competence is guaranteed, rendering it a universal condition of trust for those within the system. Still, for some patients, ensuring competence may be insufficient for trust

[49] Of course, one may be incompetent in a context of scarcity or, ideally, competent in a context of plenty.

because other conditions are unmet. What is crucial is the *way* these conditions influence how one relies on another for some aim or goal.

The conditions of trust are the *psychological* dimensions that influence trust. That is, they are what trustors attend to when trusting. There may be a worry with this move that it results in a type of subjectivism or relativism about the conditions of trust. We should be careful to note what is salient in the surgeon case and what is not. To describe the different possible conditions of trust is not to suppose that those are the right conditions, in either prudential or moral senses. One can be wrong about the conditions of trust in the sense that the relevant conditions do not deliver one's aim(s) in trusting. For instance, goodwill may be an unreliable condition for receiving the best possible medical advice. One can also trust another in ways that are exploitative or otherwise morally problematic, as we saw with Dormandy in §1.2.10. We should be careful in describing the conditions of trust not to assume too quickly that a given condition is normatively appropriate. Instead, having considered those conditions that can or could influence trust, the task then is to consider whether they should.

According to conditions pluralism, then, identifying the conditions of trust allows us to distinguish different bases of trust across individuals and contexts. Counterexamples to forms of monism demonstrate ways in which conditions alter what we think trust consists in. For example, in Baier's view, trust simply is a belief about a trustee's goodwill, whereas Nickel sees it as a stance involving obligation ascription. Attending to the different conditions or bases of trust allows us to distinguish affective and normative forms of trust. Again, in relation to other forms of pluralism, one's aims, the agents involved and their histories, and other attitudes one has, including beliefs, desires,

intentions, hopes, and so on, can all render some conditions more or less relevant in a given context. That is, aims, agents, and attitudes can direct our attention to differences in conditions of trust and, therefore, different forms of trust. For example, the basis of trust between close friends will possibly involve close affective and normative considerations that are lacking when one considers trust in governments or large multinational corporations. We can disagree about the most plausible conditions on this point. But in doing so, we are either disputing an empirical question (what is the actual basis of trust for most people in the relevant circumstances?) or a *normative question* about which forms of trust are most appropriate for the relevant cases. If the latter, it seems to me, we engage in a dispute that assumes conditions pluralism. While such disputes include information about the agents involved and their aims as well as attitudes, the heart of the debate concerns what the appropriate grounds for trust are in the relevant circumstances. Conditions pluralism alerts us to variance in these grounds, facilitating debate about them.

2.1.10 Limitations for Conditions Pluralism: Toward Pragmatism

There are two objections to conditions pluralism that I should address here. First, can we enumerate the conditions of trust? In its most radical form, one might object that conditions pluralism posits as many forms of trust as there are contexts where trust is applicable. In reply, there is compelling empirical evidence that the set of dimensions influencing trust is relatively small. In the next chapter, I examine how factors of care, confidence, competence, fairness, and values similarity, as well as a measure of one's general propensity to trust, can account for a majority of the variance in most cases.

Nevertheless, theoretically speaking, my view remains open about the possible dimensions of trust. That is, there is a theoretical openness to conditions pluralism such that the answer to the enumerative question is 'no', at least in any final sense. That does not mean that we cannot identify and dispute the relevant conditions of trust in specific cases. The point is rather that, in principle, we should not expect to list all the possible conditions that might lead others to trust. This becomes more plausible, I argue, when we consider *how* the conditions of trust influence trust. To see this, we should consider an objection to pluralism as an explanatory strategy.

The second objection regards the impact of pluralism—the 'so what?' question. One might object that it is not explanatory to say that trust is simply whatever disposes one to rely on others. As Nickel remarks, this is "like saying that desire is simply whatever it is that leads on to intentional action…[which] is merely to label a phenomenon, rather than to explain it" (2017, 198–99). The problem is that pluralism alone cannot address the second problem for monism. In my view, this objection lands for any version of pluralism about trust. Pluralism is designed to address the prevalence of counterexamples for monistic analyses of trust. By itself, it fails to explain how and why it is that normative conditions, for instance, can lead someone to trust a trustee for something within a relevant domain.

Following these two objections, we should consider how pluralism might be augmented and why classifying forms of trust is useful. By distinguishing forms of trust, we identify different grounds for our trusting. These grounds are deep and motivational, often uncovering facets of how one thinks certain relationships should be organized and maintained. So, rather than attempting to enumerate all the possible forms of trust, my

view is that we should seek to discover the bases of trust indexed to agents, times, and places. In this way, discussions of trust are intimately connected to empirical investigations of trust and the consequences of trust, allowing us to learn from and to inform such inquiries only if we understand trust's role in facilitating reliance on others. What we need, therefore, is an account of how plural forms of trust function across contexts.

2.2 Pragmatism about Trust

For developing pluralism, I distinguished between the attitude of trust and the grounds for the attitude, calling the latter the conditions of trust. The conditions of trust can vary across individuals and contexts, providing a means for identifying different forms of trust. In this section, I argue that these different forms of trust share a common core, viz. that what it means for one to trust in a given context is that one is *disposed* to rely on a trustee for an aim or goal. Put differently, we can identify a common core of trust by looking at its function. I begin the argument by situating my view methodologically within recent function-first approaches to concepts and practices. I then explain what I mean by "disposition" and argue that viewing the attitude of trust as a disposition to rely facilitates a psychological model of trust. I consider five objections from the following four points: (1) whether trust has a common function and, if so, whether it concerns dispositions to rely; (2) how my view handles amoral and immoral forms of trust; (3) whether my view neglects the second-personal character of many forms of trust; and (4) how non-trust and distrust fit into the analysis. In responding to these worries, I contend

that we can address explanatory problems confronting monistic and pluralistic views of trust.

Describing trust by its role or function is a methodological strategy. Applied in epistemology, Michael Hannon argues that a function-first approach "seeks to explain the nature and value of an epistemic concept, norm, or practice by reflecting on its function or purposes. The guiding idea is that we will better understand our epistemic concepts, norms, and practices by investigating what they are *for*" (2019, 12; emphasis original). My aim is to extend this strategy to investigations of trust. "Unlike traditional analysis," Hannon describes, "the goal of function-first epistemology is not to enumerate the necessary and sufficient conditions of our epistemic concepts…[instead, it consists in] identifying what might be called the 'core' of our epistemic concepts and practices" (ibid., 18).

As I stated at the beginning of §2.0, however, my goal in adopting a function-first approach to trust is not to replace traditional analysis *in toto*. Rather, the failure of monism and the plausibility of pluralism suggests that a different method is required. In other words, as Queloz nicely states, "When conceptual practices are held together by criss-crossing relations of family-resemblance…[traditional analysis] is likely to leave us either with a definition that is too thin to be informative, or with no definition at all" (2021, 25). This is exactly the case with trust. Identifying the core of trust *through* its function promises to facilitate a model of *how* conditions differentially impact one's reliance on others, without losing hold of what is common across cases. While my method is to construct an account of trust *per se* by considering its function, I call my approach pragmatist in the following sense. Pragmatism is designed to unify trust's

pluralistic character by examining what and how trust works or functions. For this reason, hereafter, I refer only to pragmatist approaches, where 'function-first' and 'pragmatist' are taken to be equivalent.

Most recent pragmatist (or function-first) approaches adopt state-of-nature genealogies or just-so evolutionary stories to frame plausible explanations of our present use. I provide no such story. Instead, my aim is to develop a view of trust that explains various forms of trust while maintaining their relation *as* forms of the same thing. In many cases, discerning the actual conditions that influence trust is an empirical question. My task here is to develop a model that is suitable for empirical investigation, which I connect with empirical research on trust in the next chapter. As a result, since the purpose of genealogical accounts is, as Queloz aptly remarks, "to elucidate our present," I view my more empirically-oriented approach to trust as complementary to approaches that rely on genealogies (2021, 6, n13). Indeed, Hannon argues that pragmatist views "need not make any essential reference to prehistory, nor must we trace the genealogical development of our current concept…from a more primitive concept" (2019, 52–53). At the same time, genealogies can contribute to the pragmatist enterprise. For example, Simpson's (2012) genealogy of trust highlights trust's role in social living and generates reasons to think that pluralism is plausible.

But there are potential points of methodological disagreement for pluralists. For example, in conversation with Jonathan Kvanvig's (2018) account of faith, Simpson suggests that we can distinguish forms of trust according to different values they promote—in a slogan, "axiology first" (2023, 89–91). While Kvanvig and Simpson diverge on the importance of genealogy, both take axiology to be fundamental for

analyzing faith and trust respectively. Methodologically, the idea is to allow the value we find in a topic to guide our inquiring (see Kvanvig 2018, 24–25). Simpson suggests that "there are times when trust is valuable because of the kinds of relationship that it promotes, and there are other times when trust is valuable because of what it enables the trusting individual to do, or to learn" (ibid., 90). These ways in which trust is valuable can conflict, suggesting that "there are different kinds of trust—different psychological states, with different justification conditions, and which may do different things for us" (ibid.). It is genealogy's job to explain this plurality, not axiology alone. That is, genealogy allows us to compare what we value in trust.[50]

There are two means of reply to the pluralist insight, Simpson argues. One is to scale back the explanatory ambitions of an account, recognizing that the form of trust identified in a particular case is one of many. Another approach is to contend for the primacy of a certain kind of value promoted by trust, in contrast to other sorts of value. I opt for the latter strategy, but two clarifications are in order. First, attention to what we value in trusting, including the genealogical explanation that elucidates it, is compatible with a pragmatist approach. The pragmatist grounds that value in how trust functions, emphasizing what "psychological states" share across contexts. If there are psychological states that count as trusting which share nothing in common, my view faces a problem. I return to this problem below. Second, discerning the primacy of what we value in trust among competing possible values is distinct from describing how trust realizes certain values through its function. In other words, I maintain that we can square pragmatist and

[50] Moreover, I should point out that although Williams (2002) is clear that genealogical strategies are not meant as empirical investigations, Simpson's genealogy of trust is not inconsistent with those investigations. For example, see Tamasello et al. (2012).

axiological approaches to trust in a way that unifies analyses of trust while remaining open about the preeminence of certain values. This is to ground a *description* of trust's value in its function. Determining the primacy of one form of trust relative to a context is vitally important for cultivating and maintaining trust, but such an enterprise belongs to normative evaluation rather than my descriptive, pragmatist account.

There is an additional matter for how we think about trust's function. So far, I have described pragmatist approaches as applying to terms, concepts, and practices. It is not obvious that the term 'trust', one's concept of trust, and trust as it occurs in the world are approximately the same and amenable to the same types of inquiry. Indeed, Queloz underscores that he is investigating the practical origins of *concepts* and *ideas*, not the things that those concepts and ideas are about. In part, this is a natural move for pragmatist approaches, since we are looking more to a thing's role or function in a particular context than to what it actually *is*. Consider knowledge. Queloz argues that a "pragmatic genealogy of knowledge itself would be quite a different affair [than the concept of knowledge], leading us to ask why a creature would need to *have* knowledge *about* its environment rather than why it would need to become sensitive to the *presence of* knowledge *in* that environment" (2021, 132, *original emphasis*). This is an important distinction in some cases, especially those involving natural kinds. Yet, *pace* Queloz, I agree with Hannon, who argues that just "as it makes no sense to wonder whether our culture's concept of an SUV properly answers to the true nature of SUVs, it makes little sense to distinguish the attempt to become clearer about our concept of knowledge as such. There is no sharp contrast here" (2019, 32). Why is this so? For Hannon, it is because knowledge is a social kind, rather than a natural kind.

Independently of whether knowledge is a social or natural kind, trust emerges from the needs and interests of individuals and groups that manifest in social practices and institutions. In short, trust's function is essentially social.[51] When people disagree about the conditions of trust, we can felicitously say that the two people have different conceptions of trust and that they disagree about what trust *is*. While distinct in analysis, the boundary between what trust is and what one's conception of trust is blurry. To say that one has a particular conception of trust in a given context is to say that such a conception is what the person takes trust to be in that context.

2.2.1 Trust as a Disposition to Rely

Given the plausibility of pluralism about trust, the challenge in addressing the explanatory problem is to explain why trust can vary so widely and remain identifiable as trust. In this section, I contend that plural forms of trust share a common function, namely in disposing trustors to rely. The idea in pragmatism is that we can identify a thing by its function. In relation to pluralism, then, I argue that we should view trust as a disposition to rely. That is, the attitude of trust is marked by a disposition to rely, while that disposition may have many possible grounds. To trust is to be motivated to rely on another in a particular way. That is, trust *leads* one to rely on a trustee. In the physician case, a patient trusts the physician to the extent that she is disposed to heed the physician's word. One can rely in many ways, *by* believing, acting, abstaining, and so on. What the physician case shares with the therapeutic case is not *what* disposes trustors to rely or their goals in relying—both clearly differ—but *that* one is disposed to rely.

[51] For a discussion of how trust can be considered both natural and social kinds, see Nickel (2017, 209–11).

Noticing trust's dispositional role is not without precedent. Simpson (2012, 564) writes of both dispositions to rely and dispositions to trust, identifying the former with the sorts of mental states characteristic of trust. We can distinguish the latter as being disposed to trust (see §3.4). Nickel describes how "philosophers often introduce conceptual restrictions on the allowable motivations and reasons embodied *in a person's disposition to rely on another person*, if it is to count as trust" (2017, 196; *my emphasis*). And in his account of how trust can justify testimonial belief, Kappel argues that trust involves a "non-inferential disposition to believe what some individual or other source of information asserts or transmits" (2014, 2011).

The root of my account is traceable to Bernard Williams' (2002) description of trust. Trust, says Williams, "involves the willingness of one party to rely on another to act in certain ways" (ibid., 88). When *A* trusts *B* to *φ*, *A*'s willingness to rely is sensitive to *B*'s motivations. But "in its most basic sense," Williams argues that trust "does not imply that those motives have to be of some specific kind" (ibid.). What is this most basic sense? Williams suggests that it includes cases where *A* trusts *B* because *B* can expect punishment if he fails *A*. I recommend that we view the attitude of trust, in its most basic sense, as one's willingness to rely, while one may be willing to rely according to varying conditions and expectations. Yet, as we saw with Baier's account of trust (§1.2.1), willingness does not require voluntariness or awareness. In this way, we should see trust as a disposition to rely that is inclusive of cases where one seems to have more control of whether one relies as well as cases where trust is more automatic or tacit.

Williams clarifies that identifying whether *B* is trustworthy is contingent on a more "settled background" (ibid., 89). Only against this settled background do some

assurances of trustworthiness make sense. For instance, assurances against murder may only make sense in "mafioso circles" (ibid.). In other cases, we may be able to trust on the basis of a trustee's immediate self-interest. But "in better times and places," one can count on other's cooperative actions (ibid.). I contend that pluralism in the previous section provides a ready means for identifying the salient features that differentially dispose trustors to rely on trustees.

One's disposition to rely can come in degrees. I might have an easily defeasible inclination to rely on another for some goal. That is, my trust, as a disposition to rely, is relatively weak. In other cases, my disposition to rely might be so ironclad as to be resilient to strong counterevidence. In this way, trust operates at different strengths across a spectrum. One natural way to conceive of degrees of trust follows Hardin's predictive view, wherein trust is primarily an epistemic attitude that tracks one's evidence for the reliability of a trustee. This conceives of degrees of trust as akin to degrees of belief or credence. In addition to neglecting affective, normative, and conative forms of trust, such a view obscures the explanatory function of dispositions to rely. Consider affective trust. In affective cases, trust is conditional on goodwill and care between the trustor and trustee. That is, affective conditions ground one's attitude toward a trustee *through* disposing a trustor to rely on a trustee. The same goes for other conditions on trust. In this way, for cases of trust, being differentially disposed according to certain conditions explains one's reliance. To say that *A* is *trusting B*, therefore, is to say that *A*'s relying on *B* comes about from a disposition to rely.

In addition to degrees of strength, trust can come in degrees of awareness and control. One may discover that she was disposed to rely on *A* rather than *B* only after

relying on *A* instead of *B*. Likewise, despite it being advantageous to trust *B*, a trustor might find that she cannot bring herself to trust *B*, whether from past experience, bias, or some other source.[52] In other cases, such as in business transactions or professional relationships, one may have more control over whether she is disposed to rely on a potential trustee. Actually determining the degree of strength and level of control over one's dispositions I take to be an empirical matter.

What results from trust's dispositional role is an explanatory model of trust. Recall the explanatory objection to monism, where an account of trust must meet two conditions. The input condition is that trust should be explained as the outcome of agents' concerns and interests. This condition is fulfilled by the conditions that influence one's disposition to rely on a potential trustee. The output condition states that trust should explain the emergence and sustenance of social practices and institutions. This occurs when one's disposition to rely facilitates actual reliance on another for some aim or goal. Trust's role in psychologically disposing agents to rely on each other illuminates how relevant interests and concerns result in forms of cooperative practices and institutions *through reliance*. In conjunction with pluralism, *pragmatic pluralism* leaves open what might count as instances of trust so that a range of possible conditions can possibly dispose one to rely, instead of defining the set of *possible* conditions *a priori*.

[52] That one is sensitive to certain conditions can be the result of internalized norms and expectations, as Faulkner argues (see 2011, Ch. 7).

2.3 Objections and Replies

In what follows, I consider five objections. My goal in replying to these objections is to further clarify the nature of trust as a disposition to rely. Having considered these objections, I conclude by connecting pragmatism and pluralism.

2.3.1 Objection 1: What of Trust's Other Functions?

An initial objection to pragmatism concerns trust's function or functions. Why think that its primary function is to dispose one to rely? As we saw with predictive forms of monism, trust institutes forms of social capital and cooperation that are crucial for commerce and harmonious social living. In intimate interpersonal relationships, this can be by fostering certain connections that are not shared in more contractual relationships (such that, as Simpson (2023) argues, we naturally view trust as two-placed, rather than in reference to something that is entrusted to a trustee). Moreover, mutual trust can foster love and companionship in romantic relationships. In less proximal relationships, trust could facilitate senses of belonging and kinship. Of course, each of these functions can be abused and exploited, but it seems right that trust does serve such functions. My claim is that trust serves these functions by facilitating reliance between parties. The basis of that reliance and the ends that it serves may vary, just as the conditions grounding trust vary by type of relationship and according to one's aims. My view is not that trust cannot have other functions, as I argue below with cases of amoral and immoral forms of trust. But the contention of pragmatism is that each of these functions will be consistent with

analyzing trust as a disposition to rely. In this way, we can explain not only the many possible functions of trust, but also its many forms.

2.3.2 Objection 2: Monism Reborn?

Pragmatism attempts to identify a common core to trust through its function. An immediate worry for this is that it is inconsistent with pluralism. But one might ask: is it a necessary condition that trust dispose one to rely? If so, is that not a form of monism with necessary and sufficient *functional* conditions? Suppose that the upshot of pluralism is that trust has no single function. For example, given the prevalence of counterexamples to analyses of trust, Simpson argues that "[t]here is no single, fixed concept which we all, or nearly all, use for the word 'trust'" (2012, 555). He continues: "Your disagreement with my analysis simply reveals that we represent the world differently with that word. There may be no fact of the matter about which is right" (ibid.). We see this in different forms of pluralism. For example, conditions pluralism suggests how different forms of trust "represent the world differently" by what we mean by 'trust'.

This is an important objection and underscores how I see pragmatism as addressing the explanatory problem facing monist views. First, I think there are methodological differences between monism and pragmatism. While pragmatism identifies a basic sense of trust in a disposition to rely, other contextual and empirical information is vital for understanding what leads parties to pursue cooperative social practices. The idea in pragmatism is only that trust's role or function across contexts is to dispose one to rely. In this sense, being disposed to rely is necessary for counting as trusting. But that disposition may differ considerably across contexts, including in its

strength, voluntariness, consciousness, and bases. As Williams suggests, we are never disposed to rely independently of context. That is, *in situ*, a particular form of trust may entail a host of expectations, norms, and consequences. In evaluating the suitability of predictive trust for romantic relationship, for example, one might appeal to expectations that render predictive forms of trust normatively inappropriate. What predictive trust shares with alternative forms of trust is that it is disposing one to rely on a trustee in context. Pluralism helps us to identify how the cases differ. However, in my view, if *A* entirely lacks any disposition to rely on *B* to φ in relevant circumstances, then *A* does not trust *B*. To say that one trusts is to say that one is disposed to rely.

Second, as Simpson suggests, it is true that there *may* be no fact of the matter about whose conception of trust is right, where 'right' includes prudential and moral senses. But it may be that we come to regard certain forms of trust as inappropriate to a context. It seems to me, here again, that such a view requires a normative argument about proper trust, in Baier's terms, which is separable from a discussion of possible forms of trust. Pragmatism is only meant to deliver the latter, descriptive sense. It is not intended to adjudicate deeper normative disagreements or to identify what is valuable in a trust relationship, though it can help to identify what parties value in trusting.

2.3.3 Objection 3: Concerning Distrust and Non-trust

A third objection concerns whether pragmatism provides a straightforward account of non-trust and distrust. While delimiting predictive and normative forms of trust appears in virtually all recent literature on trust, Hawley notes that the "distinction between distrust and lack of reliance, however, is usually overlooked" (2019, 4). As a first reply,

since trust comes in degrees, it is tempting to think of trust and distrust as on a spectrum, with trust lying at one end and distrust at the other. As PytlikZillig et al. (2016) caution, however, there are empirical reasons to view trust and distrust as distinct attitudes, as I suggest in the next chapter. If so, pragmatism can provide a straightforward analysis of distrust and non-trust.

For non-trust, one lacks a disposition to rely or to avoid relying. Since reliance is a type of action, one may still rely without being disposed to do so. There are people I have never met and toward which I have no disposition to rely or to avoid relying. For present purposes, I am agnostic about whether non-trust is a distinctive attitude or merely the absence of trust or distrust. What is important is that *not* trusting need not amount to distrusting.

Distrust, by contrast, involves a disposition *not* to rely on a potential trustee. That is, distrust is the functional opposite of trust. For distrust, one is sensitive to certain conditions such that those conditions block one's being disposed to rely on the relevant potential trustee. Like non-trust, one can nevertheless rely despite distrusting. For example, I may distrust an administrator but, as a forced choice, nevertheless rely on him to submit a report. Alternatively, suppose that I distrust a news source. When someone reports information acquired from that source, I disregard it *because* I am so disposed. This can help to explain the impact of reluctance on the part of historically disadvantaged groups to rely on some authorities, for example. While distrust may have different grounds, what instances of distrust share is a disposition to not rely on a potential trustee.

2.3.4 Objection 4: Second-personal Trust

According to a fourth objection, my pragmatist approach to trust and distrust neglects the second-personal nature of many trust relationships. For instance, Stephen Darwall (2017) argues that "trust is a species of second-personal attitude through which we lay ourselves open to others in a way that is distinctive of personal relationship and attachment" (46). Similarly, Paul Faulkner (2011) argues that reciprocity in trust provides an opportunity for a trustee to be responsive to one's trust, allowing trust to serve as a reason for acting within the trustor/trustee relationship.[53] One limitation of Darwall's view is that he follows Holton in viewing trust as necessarily involving a participant stance, counterexamples to which we saw in Chapter One, §1.2.10.

Nevertheless, as normative views of trust reveal how trust relationships can institute norms and expectations, so too can the reciprocity, encouragement, and confidence instituted in certain forms of trust dispose one to rely. Understanding and valuing the distinct second-personal features of trust relationships is consistent with trust's core dispositional function. Pragmatic pluralism does not exhaust relevant features for understanding trust. And there may be limitations to third-personal analyses of trust. What I maintain, however, is that the core of trust is a disposition to rely, not the second-personal features of a relationship that can dispose one to rely.

2.3.5 Objection 5: Amoral and Immoral Trust

A final objection is that pragmatic pluralism fails to distinguish amoral and immoral forms of trust from virtuous or moral cases of trust. Recall that part of Baier's motivation

[53] See also McGreer and Petit (2017) on being "trust-responsive."

for raising goodwill as a necessary condition on trust is to rule out forms of trust based on "dependably exhibited fear, anger, or other motives compatible with ill will toward one, or on motives not directed on one at all" (1994, 99). Indeed, Baier contends: "Exploitation and conspiracy, as much as justice and fellowship, thrive better in an atmosphere of trust" (ibid., 95). On this issue, for the purposes of analyzing trust and its myriad forms, I admit that fear, anger, and other motives can all serve to dispose one to rely on a trustee for some aim. This is consistent with viewing such grounds for trust and aims as immoral, rendering those forms of trust inappropriate, exploitative, and unjust. But in my view, they are immoral forms of trust and, therefore, should be included in an analysis of trust.

As for amoral forms of trust, wherein one relies on motives or dependable habits "not directed on one at all," these can be cases of trust. For example, Baier says, "Kant's neighbors who counted on his regular habits as a clock for their own less automatically regular ones might be disappointed with him if he slept in one day, but not let down by him, let alone had their trust betrayed" (ibid., 99). Now, Kant's neighbors would not be right to feel betrayed if he altered his sleeping schedule—that is, their reactive attitudes would be inappropriate. While I should note that reactive attitudes are not necessary for trust, *could* Kant's neighbors feel betrayal? Yes. Suppose they mistakenly took Kant to be committed to a public time service, or thought him a nosy neighbor checking in on others' routines, or for some other reason. Whatever his reasons or theirs, it seems to me psychologically possible that Kant's neighbors could be disposed to rely on his predictable behavior for tracking the hours. If this is possible, then they can be said to

trust, even as we are prepared to condemn that trust as inappropriate. In this way, such a condemnation is not altogether distinct from Baier's views about contractarian ethics.

2.4 Concluding Remarks: The Limits of Pragmatic Pluralism

With moral and amoral forms of trust, we come to the limits of pragmatic pluralism. The view is *not* designed to resolve disagreements about what ought to dispose one to rely on another, only to illuminate such disagreements. Once more, consider the physician case. It may be that when a patient is willing to continue seeing her physician after discovering that her physician disdains his patients, the patient continues on conditions of expert judgment and competence, alongside conditions for trust, such as occupational standards for behavior. Alternatively, for the patient that seeks medical care elsewhere, her trust may fail because goodwill, alongside competence and professional integrity, disposes her to rely on a physician for medical care. The maintaining *and* breaking of trust relationships in the respective cases reveal what trustors value in their physician-patient relationships—indeed, they identify potentially valuable features of those relationships. To adjudicate between them, we must consider what we value in relationships and how those values relate to the conditions we consider necessary for trusting. In disputes about the conditions of trust, we must enter into what Elizabeth Stewart (2022) calls a negotiation about trust in the relevant domain. Such negotiations necessarily involve "individual ideals and social norms," as well as views of ideal and non-ideal conditions on trust within the relevant domain (ibid., 4, 6). This task, however, lies outside the ambitions of pragmatic pluralism.

To conclude, allow me to underscore a central upshot of this chapter. With a turn to pragmatic pluralism, trust and, by extension, trustworthiness are value-laden. That is, the conditions that dispose one to rely on another are sensitive to the values, needs, and expectations of trustors and trustees. In some cases, there may be widespread agreement about what trust consists in—for instance, the appropriateness of care and goodwill between close friends. In other cases, the appropriate conditions of trust may be hotly contested. Pragmatic pluralism allows us to describe this variability. Trust *can* lead people to rely under myriad conditions and for multiple ends. We may (rightly) disapprove of some cases, as in sexist, racist, or abusive forms of trust. Yet, these are indeed forms of trust precisely in the sense that certain conditions (appropriately or not) dispose trustors to rely on trustees.

The *normative* challenge, therefore, is to differentiate appropriate conditions of trust from those that are not. Where does pragmatic pluralism leave the problem of differentiating conditions of trust? Of itself, the view offers limited normative input for determining when trust is well-placed. The service it provides, however, is to direct our attention to considerations that are essential for determining when one's trust is normatively appropriate. By identifying the conditions that do and could dispose one to rely on another for some aim, pragmatic pluralism supplies terms from which to begin appraising what *should* dispose one to rely in the relevant circumstances.

3.0 THE EMPIRICAL PLAUSIBILITY OF PRAGMATIC PLURALISM

In Chapters One and Two, I developed and defended *pragmatic pluralism* about trust. According to that view, trust comes in many forms that share a common function, namely disposing trustors to rely on trustees. While addressing counterexamples to other theories and providing a philosophical explanation of trust, I suggested that the view is theoretically unrestricted in the sense that it does not assume *a priori* limits on what can or could dispose one to rely. However, this openness may seem too high a cost, raising questions about how useful the theory is. As Karen Jones remarks, views that count any disposition to rely as trust "identify too heterogeneous a class of dependencies to support useful generalizations and thus do not provide a useful classification for the purposes of social scientific or other theorizing" (2004, 4). The object of this chapter is to twofold. First, I aim to show how pragmatic pluralism is consistent with empirical findings in the social sciences. Second, *contra* Jones' objection, I suggest ways that pragmatic pluralism can aid future empirical research.

Two points of clarification are useful at the outset. First, both within and across empirical disciplines, there is little agreement about the nature of trust and how it ought to be studied. As I discuss in the next section, competing empirical approaches to trust define trust differently. From these definitions, different experimental methods and measures develop. So, empirical findings concerning trust are rarely, if ever, straightforward in their application to philosophizing about trust. For this reason, I focus more on theoretical considerations that frame empirical findings, drawing on those findings cautiously so as to recognize persistent disagreements in the empirical literature.

In this way, my aim in this chapter is not to suggest that empirical research on trust *demonstrates* that pragmatic pluralism is the only plausible view about trust—on empirical grounds, it is not. Rather, contra Jones, I argue that pragmatic pluralism is consistent with and can contribute to empirical research on trust.

Second, a few terminological distinctions are helpful for orienting the following discussion. First, social scientists often describe trust as a *construct*. A construct is a conceptual label that helps to explain experimental results. For example, we might develop a definition for intelligence (see §5.1.1) that helps to identify and explain intelligent behavior. For present purposes, I remain agnostic about inferences concerning the existence of underlying constructs (i.e., it is possible that they track the nature of some distinct psychological entity, or they may provide a nominal means for grouping behaviors). When there are disagreements about how to understand a construct, researchers may draw on other constructs. For instance, as I explain in §3.1, researchers identify *dimensions* of trust as constructs that reveal salient features of trust. To test theories about constructs, researchers *operationalize* the construct in ways that can be measured. For example, one could operationalize a definition of intelligence by measuring someone's intelligence quotient (IQ). Finally, empirical investigations variously refer to *bases* for trust (Rousseau et al. (1998)), trust-related *factors* (Hamm and Hoffmann (2016)), *dimensions* of trust (PytlikZillig et al. (2016)), and *reasons* for trust (SteelFisher et al. (2023)). Following the empirical literature, I will use these terms interchangeably unless otherwise explicitly noted.

I proceed as follows. In §3.1, I discuss relevant historical and methodological developments in trust research. Empirical trust research is vast and diverse, involving

multiple disciplines, conceptualizations, and methods. In §3.2, I examine how researchers have identified a set of conditions that regularly influence trust, what is sometimes called trust's *multidimensionality*. I argue that multidimensionality supports pluralism about trust. However, as Michael Siegrist (2021) argues, social scientists can reliably identify types of trust across contexts. What is necessary is some means for unifying conceptualizations of trust. So, in §3.3, I investigate two models of trust and their validation. While these models retain the pluralistic insights of multidimensionality, they also conceive of trust in terms consistent with pragmatism. Finally, I conclude in §3.4 by considering how pragmatic pluralism can influence and be influenced by future empirical research.

3.1 Investigating Trust: Historical and Methodological Background

Researchers from across the social sciences have examined trust's role in cooperation and social living, including psychology, sociology, political science, economics, history, anthropology, management and organizational studies, risk analysis, communication studies, among others.[54] Some disciplines, such as sociology and political science, focus on trust's role in culture and social order, often relying on psychology or philosophy for definitions of trust.[55] In this way, sociological research reveals salient information about the contexts and consequences of trust, especially across cultural and demographic

[54] Cook (2016) provides an excellent annotated bibliography of empirical literature on trust. See also Rousseau et al. (1998) and Siegrist (2020).
[55] See Cook and Cook (2011), as well as Cook and Santana (2020), for good overviews of this literature.

differences.[56] In what follows, however, I focus on how trust itself is defined and operationalized in psychological literature. While definitions and methods differ considerably, resulting in approaches that can be difficult if not impossible to synthesize, I argue that existing consensus moves in the direction of pluralism about trust.

3.1.1 Definitional Development of Trust

Early empirical investigations of trust focus on its role in cooperation and prosocial action, especially in contexts of exchange. For example, Morton Deutsch (1958; 1973) investigates trust as a type of decision within a game-theoretical framework, primarily utilizing the 'prisoner's dilemma'.[57] Approaches following Deutsch aim to capture how perceptions of another person's motivations impact cooperative behavior.[58] To do this, researchers utilize zero-sum and non-zero-sum games in which maximized outcomes require cooperation and, therefore, trust. In this way, trust is viewed as a type of risky investment. However, as sociologists, Andreas Tutić and Thomas Voss argue, most game-theoretical approaches tend to presuppose orthodox decision theory and, as a result, operate with "rather narrow and empirically questionable action-theoretic assumptions regarding human conduct," neglecting the role of moral and emotional elements of trust (2020, 186). This is not to say that game-theoretical inquiries are unimportant either for understanding cases of trust or independently of trust. Indeed, as Tutić and Voss

[56] For examples of cross-cultural and ethnographic research, see Yamagishi et al. (1998), Buchan et al. (2002), Habyarimana et al. (2009), Steinhardt (2012), and PytlikZillig (2016).
[57] Earle and Cvetkovich (1995, 17) note that Deutsch was the first psychologist to use prisoner dilemmas to study trust between individuals and small groups. Deutsch's approach is consistent with how many economists tend to view trust, namely as a calculative attitude (Williamson, 1993) or as part of transactional cooperation within institutions (North, 1990).
[58] For examples that consider trust specifically, see Dasgupta (1988), Kreps (1990), Coleman (1990), Sugden (1993), Bacharach (1999), and Skyrms (2008). Tutić and Voss (2020, 175–88) provide an accessible overview of influential game-theoretical approaches to trust.

conclude, game-theoretical approaches can reveal the impact of social mechanisms on human cooperation, such as repeated interaction and signaling social information (ibid.). Rather, their contention is that viewing trust merely as a type of risky *behavior* neglects salient conceptual features of trust, such as motivation and expectation.

In contrast to trust as behavior, Julian Rotter (1967, 1971, 1980) investigates trust as a personality trait variable.[59] According to this approach, trust is a *general expectancy* about the motives of others that influence psychological characteristics and behaviors, including levels of happiness, cheating, maladjustment, conflict, and friendliness.[60] To examine the possible effects of high levels of trust, conceptualized as a prosocial trait, experimenters utilize surveys and participant reports. In 1967, Rotter introduces one of the first scales for measuring trust, the Interpersonal Trust Scale. The scale includes 25 items for measuring "a generalized expectancy that the oral or written statements of other people can be relied upon" (ibid., 653). Here are two examples: "Parents usually can be relied upon to keep their promises" and "Most elected public officials are really sincere in their campaign promises" (ibid., 654). This facilitated findings about trust's role in cooperative behavior. For instance, in an experiment examining competition within and between small groups, Steinke (1975; cited by Rotter 1980) found that those with lower scores on the Interpersonal Trust Scale were statistically more likely to cheat to win games when given the opportunity. However, the important point here is that such

[59] To underscore the dearth of research on trust, Deutsch notes that six of the most popular textbooks in social psychology do not mention the word "trust," remarking that "[s]o far as we know, the research summarized in this paper represents the first attempt to investigate experimentally the phenomena of trust" (1958, 265).

[60] See Rotter (1980) for a good discussion of this approach and its view of trust's role in social learning.

findings are conditional on the scale used to measure trust and the scale's underlying conceptualization of trust.

These different approaches in psychological literature undersell the diversity in approaches to trust in the last half of the twentieth century. In a survey of definitions of trust in psychological literature, following Castaldo (2002), Cristiano Castelfranchi and Rino Falcone (2010, 8) identify as many as *72* distinct definitions.[61] Given this persistent disagreement in conceptualization and, as I examine in §3.2.1, operationalizations of trust, Castelfranchi and Falcone suggest that trust "deserves a *non-reductive* definition and modeling" (2020, 214; emphasis original).

Nonetheless, these definitions tend to share some structural features, often involving a trustor *X* who has some attitude toward a trustee *Y* that is expected to perform an action. Accordingly, Castelfranchi and Falcone propose the following relationship between (a) trust as a psychological trait and (b) trust as an action. For (a), one can view trust as a "*psychological attitude* of *X* towards *Y* relative to some possible desirable behavior or features" (2010, 18; emphasis here and hereafter original). For (b), one can view trust as "the *decision and the act of relying on*, counting on, depending on *Y*" (ibid.). According to Castelfranchi and Falcone, there is a conceptual and causal link between (a) and (b). Conceptually, they argue that "the *intension* of (*b*) contains (*a*)…trust as an attitude is part of the *concept* of trust as a decision/action" (ibid.). This underscores a problem for views of trust as an action, since one can rely without trusting. For instance, one may rely on plumber to repair a leak without trusting her. What focusing on the action fails to reveal is the psychological difference between non-trusting reliance and the

[61] See also Castaldo et al. (2010), McEvily and Tortoriello (2011), Siegrist (2020), Cook (2016), Lee and See (2004), Rousseau et al. (1998), PytlikZillig, L. M., & Kimbrough, C. D. (2016), and others.

person who has an attitude toward the plumber that disposes the person to rely. Again, this does not mean that non-trusting reliance is unimportant or uninteresting. Rather, for an action to count as trusting entails a "psychological attitude." Causally, Castelfranchi and Falcone argue that this psychological attitude is a "temporal presupposition" of the act of trusting (ibid.). So, for identifying trust, the act of relying on *Y* requires the formation of the psychological attitude.

Castelfranchi and Falcone's account is consistent with one of the most influential definitions of trust. In a cross-disciplinary review of psychology, sociology, management, and economics, Denise Rousseau et al. (1998) attempt to bring together salient features of different definitions of trust. They define trust as "the psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another" (ibid., 395). It is important to note that "intention to accept" does not entail that trust is voluntary. Rather, it tracks what Mayer et al. (1995) describe as a "willingness to take a risk," where trust "is not taking risk *per se*" but is one's inclination to take the risk (ibid., 712).[62] So, emphasis should be placed on identifying trust in a "psychological state." That is, Rousseau et al. continue: "Trust is not a behavior (e.g., cooperation) or a choice (e.g., taking a risk), but an underlying psychological condition that can cause or result from such actions" (ibid.). Indeed, they conclude that the "case for integrating trust across disciplines rests on the common psychological basis upon which all formulations of trust rest" (ibid., 398). So, for Rousseau et al., in cases where researchers view trust as a type of action (e.g., in game-theoretical approaches), what integrates research is a focus on the psychological basis of trust. In this way,

[62] I discuss Mayer et al.'s model of trust that relies on this definition in §3.4.

Castelfranchi and Falcone's (2010) account of the relationship between trust as action and as a psychological attitude provides a plausible way of explaining how Rousseau et al.'s definition can unite competing definitions.

3.1.2 Trust as a Psychological State: On Distrust and Forms of Trust

To conclude their review of trust literature, Rousseau et al. identify two lingering issues for future research (ibid., 398–401). First, they argue that research should examine the relationship of trust to distrust. It may be that measures of trust and distrust track the same construct or separate constructs. Second, to understand how trust functions over time and across relationships, Rousseau et al. contend that researchers should consider the possibility of multiple forms of trust. Identifying forms of trust, they argue, raises questions about their differences and commonalities. Let's consider each point in turn.

There is persistent disagreement in the empirical literature about how to relate trust to distrust.[63] Some researchers, including Rotter (1970), view trust as a spectrum concept, with trust at one end and distrust at the other.[64] For example, David Schoorman et al. (2007, 350) argue that ordinary language suggests that trust and distrust operate as two ends of a continuum. For instance, they cite the definition of distrust in the Merriam-Webster Dictionary: "the lack or absence of trust."[65] Consider also how a five-point Likert scale might fit trust and distrust as a spectrum (see Schoorman et al. 2007, 352). Suppose that a 5 ("Strongly Agree") on the scale indicates high levels of trust, a median

[63] See Castelfranchi and Falcone (2020, 217–18).
[64] For examples, see Stack (1988), Tardy (1988), Mayer et al. (1995), Poortinga and Pidgeon (2003), and Schoorman et al. (2007).
[65] "Distrust." Merriam-Webster.com Dictionary, Merriam-Webster, https://www.merriam-webster.com/dictionary/distrust. Last accessed 28 Feb. 2024.

answer, 3 ("Neutral"), indicates non-trust or lack of trust, and a 1 ("Strongly Disagree") indicates distrust. Imagine participants are presented with the following prompt: "I trust the local police department to operate competently and benevolently."[66] It may be that results from the trust-distrust scale measures degrees of trust, with high levels of trust at one end of the scale and low levels of trust (i.e., distrust) at the other end.

However, interpreting the results this way is contingent on viewing trust and distrust as the same construct. As I noted in §2.3.3, other researchers argue that trust and distrust are separate psychological constructs. Here, I discuss four reasons for separating trust and distrust in the empirical literature.

First, Roy Lewicki et al. (1998) contend that one can have high levels of trust and distrust for the same trustee, depending on what a trustor relies on a trustee to do.[67] For instance, one might trust a surgeon for heart surgery but not for auto repair qua surgeon (see §2.1.6). In this way, a low score on Schoorman et al.'s trust scale can be ambiguous. It may indicate that participants view a task as lying outside the scope of someone's trustworthiness. Alternatively, it may indicate that a trustor recognizes that a task is within the purview of someone's competence but thinks a potential trustee is untrustworthy.

Second, Lewicki et al. (ibid., 448) argue that the opposite of trust is not distrust. They draw on empirical findings that suggest positive-valence and negative-valence constructs are separable (see ibid.). For example, evidence suggests that high positive affectivity (e.g., enthusiastic, peppy, elated, etc.) is not synonymous with low negative

[66] This is similar to a local governance measure in PytlikZillig et al. (2016, 124)
[67] See also Deutsch (1960), Marsh and Didden (2005), McKnight and Choudhury (2006), and Van De Walle and Six (2013).

affectivity (e.g., calm, relaxed, placid etc.). Likewise, low positive affectivity (e.g., dull, sluggish, drowsy, etc.) is not synonymous with high negative affectivity (e.g., distressed, hostile, fearful, jittery, etc.).[68] However, while their findings are consistent with trust and distrust as separate constructs, Lisa PytlikZillig et al. (2016) note that the separation of positive and negative constructs can reflect measurement artifacts, especially depending on how survey items are worded.

Yet, a third reason to regard trust and distrust as different constructs is they may function differently neurologically. For example, Angelika Dimoka (2010) found that standard psychometric measures of trust and distrust correlate with activation in different parts of the brain. Domika designed a behavioral study in which participants purchased electronics from four eBay dealers that varied according to different bases of trust (in the sense in Rousseau et al. above), namely credibility (or discredibility) and benevolence (or malevolence). When interacting with the sellers, participants' brain activity was monitored with functional Magnetic Resonance Imaging (fMRI). Domika found neural correlates for trust in the paracingulate cortex and orbitofrontal cortex, caudate nucleas, and putamen, whereas neural correlates for distrust were in the bilateral amygdala and insular cortex (ibid., 385).[69] Interestingly, the brain areas associated with trust concern reward, prediction, and uncertainty, while the areas associated with distrust concern intense emotions and fear of loss (ibid., 388). A lingering limitation of the study, Domika concludes, is to better understand the dimensions of trust and distrust. That is, while trust

[68] Lewicki et al. (ibid.) cite literature across several domains where these findings hold, including assessments of optimism and pessimism, interracial attitudes, and attitudes toward organ donation.
[69] Domika's Figure 3 is helpful for visualizing the different areas of activation (ibid., 386).

and distrust seem neurologically distinct, more should be said about the factors (i.e., dimensions) influencing trust and distrust.

Fourth, Castelfranchi and Falcone note that "while distrust is not simply the direct opposite of trust, its exact nature is still up for debate" (2020, 218). In their view, we should distinguish lack of trust from "true distrust" (ibid.). The former may be a case wherein one simply neither trusts nor distrusts another party. The latter involves a "negative *evaluation* of the trustee and of its ability, intentions, possibilities that produces as a consequence a negative expectation" (ibid.; emphasis original). That is, Castelfranchi and Falcone argue, distrust involves a psychological "disposition" against the trustworthiness of a potential trustee (ibid.). Trust, then, involves a positive psychological disposition toward the trustworthiness of a potential trustee.

While disagreements about the relation of trust and distrust are likely to persist, this fourth point reveals something about trust that connects to Rousseau et al.'s second item for future research, examining "emerging forms of trust" (1998, 393). They identify three forms of trust. First, "calculus-based trust" is based on rational choice, corresponding to what I have called predictive trust in previous chapters (ibid., 399; see §1.2.3–1.2.7). Second, "relational trust" is "based upon reciprocated interpersonal care and concern" (ibid.). In this way, relational trust corresponds to affective trust (see §1.2.1). Third, "institution-based trust" can situate calculus-based and relational forms of trust in broader contexts, where deterrents for unreliability and rewards for reliability promote trustworthy behavior (ibid., 400). These institutional bases of trust draw from sociological research on the environments in which trust is formed and sustained, and

they note that institution-based trust may be more a control for forms of trust, manifesting in sanctions and reputational factors (ibid.).

Rousseau et al. do not suggest that these forms of trust are exhaustive of all possible forms of trust. Indeed, they note that shifting social and institutional contexts may result in different bases for trust and, therefore, different forms of trust (ibid., 402). Their conclusion is instead that investigating forms of trust can reveal "the true functioning of trust" within and across contexts (ibid.). In this way, their review suggests a strategy for identifying various forms of trust, namely by looking at the bases or dimensions that influence the psychological state that directs a trustor to rely on a trustee. I return to developments of this strategy in §3.3.

3.1.3 Operationalizing Trust: Direct or Indirect Measures

Despite the plausibility of their definition of trust, Rousseau et al. acknowledge that "identification of a common meaning does not imply that all operationalizations of trust reflect the same thing" (1998, 395). In a survey of different measures of trust, for example, Bill McEvily and Marco Tortoriello (2011, 27–28) identify *129* distinct measures of trust across 171 studies between 1962 and 2010. Helpfully, Fergus Lyon et al. (2016) provide a representative collection of methods in empirical trust research, including both quantitative and qualitative approaches. Broadly, trust researchers develop either behavioral studies, wherein participants act in such a way that trust can be measured, or surveys, which provide responses designed to measure trust.[70] In this

[70] In addition to definitional differences between early trust researchers, there are also methodological differences. For instance, researchers variously adopt lab experiments (Deutsch 1973), field observations (Garfinkel 1967), and surveys (Rotter 1967).

section, I focus on a more basic puzzle for trust research, namely how we should measure trust.

John Besley and Leigh Tiffany (2023) distinguish between direct and indirect measures of trust. Direct measures of trust ask participants some version of the question "how much do you trust *Y*," where *Y* is a potential trustee. They are direct in that they explicitly reference trust. Indirect measures, by contrast, assess other factors that researchers think correlate with trust. As Besley and Tiffany explain, indirect measures are intended to capture "some aspect of trustworthiness perceptions or behavioral trust" (ibid., 2). That is, for indirect measures, one identifies factors that influence whether participants view a potential trustee as trustworthy such that they are willing to rely on the trustee. As an initial gloss on these factors, Besley and Tiffany include the competence of the trustee, their benevolence or goodwill toward a trustor, and their overall integrity (ibid.).

There can be good reasons to utilize direct measures of trust. There is considerable variance in how trust is defined in empirical literature. Direct measures sidestep this issue by allowing participants' understanding of trust to guide responses. In this way, direct measures do not require researchers to identify potential factors indicating trustworthiness to develop a measure for trust. Moreover, direct measures can be more efficient than indirect measures. Some of the most influential and longstanding studies of trust measure trust directly, providing questionnaires across disciplines and contexts (see Kohn et al. (2021)). For instance, running annually since 1972, the National

Opinion Research Center's General Social Survey (GSS) utilizes direct trust measures to assess social change.[71]

The limitations of direct measures are straightforward. First, it is possible that participants have different conceptions of trust such that direct measures track different constructs. Second, direct measures may merge distinct constructs. For instance, a Gallup poll that examined trust in different branches and levels of government in the United States found that public trust is highest in local governments, followed by state and federal governments respectively.[72] The study reported that trust in the federal government's capacity for problem solving is at an all-time low. A closer look at the survey, however, reveals that participants were asked whether they have "trust and confidence" in the relevant party. However, it is standard in trust research to distinguish trust and confidence, not least because researchers repeatedly find an interaction between confidence and trust.[73] In this way, by overlapping or merging various constructs, direct measures can fail to explain the significance of findings.

The foremost challenge for indirect measures is relating other constructs to trust. This requires a heavy reliance on theory. For example, measures for competence may be less controversial than direct trust measures.[74] The problem, as McEvily and Tortoriello suggest, is to specify "dimensions that are distinct, yet related" to trust (2011, 37). In devising indirect measures, McEvily and Tortoriello note the diversity of measures and

[71] See https://gssdataexplorer.norc.org/home; last accessed 1 Mar. 2024. One example of a trust measure on the GSS is "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?"
[72] See https://news.gallup.com/poll/355124/americans-trust-government-remains-low.aspx; last accessed 1 Mar. 2024.
[73] Luhmann (1979) distinguishes trust from confidence. See Siegrist (2021)'s review of literature on trust and confidence. See Allum (2007), Earle and Siegrist (2006), Eiser et al. (2015), Siegrist et al. (2003), Siegrist et al. (2007), and Siegrist et al. (2012).
[74] See Mayer and Davis (1999).

lack of consensus about relevant dimensions of trust (ibid., 35). Some of this is explained by the context-specificity of trust research, as well as disciplinary diversity. Yet, they also note that there is a risk that "researchers have devised idiosyncratic measures of trust due to the lack of availability of carefully designed and validated instruments" (ibid.).

To address this latter problem, recent investigations of trust attempt to validate indirect measures by correlating them with direct measures.[75] For example, Besley and Tiffany (2023) examine dimensions relevant for measuring trust in science in the GSS. They find, for instance, that direct measures of trust correlate better with indirect measures specified to specific vulnerabilities and risks than to general measures of trust in the scientific community overall (ibid., 7).

This provides researchers with a principled means for assessing the degree to which an indirect measure relates to the direct measure of trust. For example, a Swiss study conducted by Wintterlin et al. (2022) correlated a direct measure of trust with measures for ability, integrity, and benevolence. This allows for comparison with other studies. For example, Besley and Tiffany examine a study of trust in scientists conducted by Gallup for the Wellcome Trust in 2018 (2023, 10–12). This Gallup study includes both direct measures for trust and measures for ability, integrity, and benevolence. In this case, Besley and Tiffany show that measures for ability explain double the variance in the data when compared to integrity and benevolence (ibid., 12). However, in cases of trust in medical and environmental scientists conducted by the Pew Research Center, Besley and Tiffany found that ability, integrity, and benevolence capture roughly equal variance in the direct trust results. Accordingly, they conclude that "science communication

[75] See Kohn et al.'s (2021) review.

researchers who want to ask about how people perceive scientists should specifically ask about whether they see scientists as competent (i.e. high in ability), honest (i.e. high in integrity), and caring (i.e. benevolent)" (ibid., 15).

Nevertheless, there are the limits to correlating indirect measures with direct measures. As Besley and Tiffany conclude, "it seems clear that direct measures of trust are capturing some aspect of trustworthiness, but that the pattern of relationships also likely depends on some unknown contextual factors" (ibid.). In other words, correlating indirect measures with direct measures of trust does not explain how different factors are *causally* related to how people think and act. For instance, it might be that one has a level of trust in scientists that results in high estimations of their ability, integrity, and benevolence. Accordingly, in the next two sections, I examine evidence investigating salient dimensions of trust and how those dimensions relate to trust itself.

3.2 Multidimensionality: Toward Pluralism

Given the insight provided by indirect measures for trust, Lisa PytlikZillig et al. (2016, 112) suggest that trust is "multi-faceted and multidimensional." That is, there are multiple dimensions of trust that help researchers understand direct measures of trust. There are two persistent challenges for understanding the multidimensionality of trust in empirical trust literature. First, most discussions of dimensionality rely heavily on theory, as I noted in the last section. Supposing agreement about trust as a construct, a second challenge is that, as PytlikZillig et al. remark, "few empirical studies have addressed the dimensionality of trust-relevant constructs" (bid., 113). And those studies that do attempt

to account for dimensions of trust either adopt too limited a list of dimensions or attempt to measure dimensions with too few (sometimes single) survey items (ibid.). To address these problems, PytlikZillig et al. designed models of trust that combine different dimensions of trust. They tested these models in four survey studies, examining: (1) which model correlates best with direct trust measures, (2) how dimensions relate to each other, and (3) whether the dimensions vary in influence across contexts. In this section, I argue that PytlikZillig et al.'s findings suggest in favor of pluralism about trust.

In the previous section, I considered the methodological relationship of direct and indirect measures. To situate dimensions of trust into the theoretical developments in §3.2.1, PytlikZillig et al. (ibid., 140) suggest that trust-related dimensions indicate "perceptions of trustworthiness." That is, the dimensions of trust indicate to a trustor that a prospective trustee is trustworthy. In this way, if trust is a psychological state indicative of one's willingness to rely on a trustee, the dimensions of trust are the conditions on which one's trust is based (in the sense of "based on" in Rousseau et al.'s and Mayer et al.'s definitions in §3.2.1). One reason that PytlikZillig et al.'s studies provide helpful insights for assessing the impact of different dimensions is that they include not only the factors in each model, but also the survey items for each factor and demographic data for participants in each study.

The measures PytlikZillig et al. deploy in their studies can be grouped into three categories (Figure 1). First, dispositional trust (sometimes called *general trust*) measures a trustor's general propensity or disposition to trust as distinct from specific instances of trust. As Mayer et al. propose, a propensity to trust is "a stable within-party factor that will affect the likelihood the party will trust" in particular cases (1995, 715). They argue,

propensities to trust "contribute to the explanation of variance in trust if used as a part of a more complete set of variables" (1995, 716). That is, one's propensity to trust in general is one factor that influences specific cases of trust. Agent *A* could have a propensity to rely on Agent *B* in circumstance *C* for specific purposes, but not in different circumstances or relative to different aims. Yet, for instance, *B* might not be as trusting in general as *A*. This allows us to acknowledge potentially wide variance in *trustingness* (i.e., how trusting one is in general) without confusing that general propensity with specific cases of trust or allowing the general propensity to dominate placing and updating trust.

MF: Many-Factor Constructs*	6F: Six-Factor	5F: Five-Factor	4Fa: Four-Factor, Ability/ Warmth	4Fb: Four-Factor, Positive/ Negative	3F: Three-Factor	2F: Two-Factor
Dispositional Trust[1,2,3]	*Disposition. Trust*	*Disposition. Trust*	*Disposition. Trust*	*Disposition. Trust*	*Disposition. Trust*	*Disposition. Trust*
Direct/Unspecified Trust[1,2,3]	*Trust*	*Trust*	*Trust*	*Trust*	*Trust*	*Institutional Trust*
Loyal Trust[1,2]						
Perceived Competence[1,2,3,4]	*Perceived Ability*	*Perceived Ability*	*Perceived Ability*	*Positive Attitudes*	*Perceived Trustworth.*	
Perceived Legitimacy[1,2,3,4]						
Perceived Care[1,2,3,4]	*Perceived Benevolence*	*Perceived Benevolence*	*Perceived Warmth*			
Perceived Voice[1,2,3,4]						
Perceived Honesty[1,2,4]	*Perceived Integrity*	*Perceived Integrity*				
Perceived Fairness[1,2,3,4]						
Perceived Shared Values[1,2,3,4]	*Values/ Identificat.*					
Cynical Beliefs[1,2,3,4]				*Negative Attitudes*		
Perceived Bias[1,3,4]	*Perceived Integrity*					

Note: Identificat. = Identification, Trustworth. = Trustworthiness, Disposition. = Dispositional.

*Many-Factor (MF) Model treats each construct as a factor not combined with any other constructs. Other models combine indicated factors separated in the MF model. Superscripts indicate constructs [1]included in Study 1, [2]included in Study 2, [3]included in Study 3, and [4]included in Study 4.

Figure 1. Measurement models for dimensions of trust in PytlikZillig et al.'s (2016) studies. Reproduced with permission from Taylor and Francis.

Second, PytlikZillig et al. include direct trust measures in each study. The direct measure is "unspecified" in the sense that "both the definition of trust and the bases for that trust [are] unspecified for the respondent" (ibid., 115).

Third, drawing on cases in trust research, PytlikZillig et al. identify possible dimensions for constructing different models. A discussion of each dimension would reveal salient features for thinking about possible factors impacting trust. For example, in a study of impacts on trust in policies for managing environmental risks, Timothy Earle and Michael Siegrist (2008) found that procedural fairness influenced trust scores in cases with little-to-no moral valence, whereas values similarity between parties impacted cases where participants identified an issue as morally relevant. What is crucial for our present discussion is to see how PytlikZillig et al.'s studies identify finite sets of dimensions or, in the terms of Chapters One and Two, conditions that explain much of the variance in direct measures of trust.

PytlikZillig et al. construct structural equation models that combine different dimensions of trust, producing models that range from two factors to a model that combines no factors, examining the variance captured by 12 independent factors. For instance, in the discussion of trust and distrust (§3.1.2), I noted that researchers sometimes distinguish between positive and negative perceptions. This suggests a four-dimensional model, including dispositional trust, direct trust, positive attitudes, and negative attitudes (see 4Fb in Figure 1). Alternatively, as I discuss in the next section, Mayer et al. (1995) measure ability (labeled 'competence' here), benevolence, and integrity, resulting in a five-factor model.

In each study, PytlikZillig et al. first analyzed the model that captures the most variance in the survey data. They then examined the impact of each factor to create post hoc exploratory models that collapsed factors whenever there is a high degree of covariance between factors.

Consider briefly the topic of each study. Study 1 (n = 720) examined college students' trust in a local police department. Study 2 (n = 890) examined the residents' perceptions of local public officials in Lincoln, Nebraska. Study 3 (n = 645) examined landowner perceptions of natural resource managers in the Nebraska Game and Parks Commission. Study 4 (n = 399) examined Americans' perceptions of state governments in an online survey.

For each study, the many-factor model fit best.[76] However, the relevant factors across studies were not the same, suggesting that participants were sensitive to different dimensions of trust in the particular context. For example, in study 3, "care, competence, direct/unspecified trust, legitimacy, procedural fairness, shared values, and voice" covered a majority of the variance (.80) (ibid., 131). Factor analyses revealed that positive and negative factors were highly correlated. As a result, PytlikZillig et al. created a post hoc model combining items for competence, legitimacy, fairness, care, unspecified trust, shared values, and voice as one factor, and bias and cynical beliefs in another factor. Despite the factors' high degree of correlation, the model fit the data considerably worse.

A full discussion of every facet of PytlikZillig et al.'s studies is beyond our present purview. There are two findings that are salient for philosophizing about trust. First, and most importantly, the factors PytlikZillig et al. captured most of the variance in each study (ibid., 137). They acknowledge that there may be other dimensions of trust that they did not examine, such as "willingness to support, give control to, or otherwise

[76] PytlikZillig et al. deployed the following four measures for fit: comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). For a discussion of fit in structural equation modeling, see (2016, 121).

be vulnerable to [an] institution" (ibid., 138). Indeed, they call for future research to investigate the possible impacts of procedural fairness and justice on trust (ibid,. 143) This suggests that, empirically speaking, researchers could identify relevant factors for trust in particular cases, while acknowledging the pluralist's point that the perceptions of trustworthiness across contexts vary. In this way, the pluralist can maintain theoretical openness about the factors that could impact trust without supposing that, as an empirical matter, there is not a set of relevant factors in a particular case.

Second, PytlikZillig et al.'s analyses consistently found high correlations between direct trust and dimensions of trust, while they found lower correlations with dispositional trust. For this reason, they suggest that that the dimensions of trust have a stronger influence on individual cases of trust than one's general propensity to trust across cases (ibid., 138). This indicates that trust in particular cases is highly contextual, with some dimensions being relatively more important depending on context and relationship.

Nonetheless, one might rightly wonder how we should relate these dimensions of trust to each other and to what PytlikZillig et al. call "trust *per se*" (ibid., 115). As they explain, from a measurement perspective, it may be "that some of these conceptually-distinct [dimensions] are statistically or practically indistinguishable" (ibid., 112). To address this problem, I examine modeling techniques that trust researchers utilize to connect dimensions of perceived trustworthiness to trust and trust behavior.

3.3 Modeling Trust

In the previous section, I argued that research on the dimensions of trust is consistent with pluralism about trust. Researchers develop various models of trust to explicate how the dimensions of trust relate to the attitude of trust and to trust behavior. In this section, I examine two models. First, I argue that the Integrative Model of Trust lends support to pragmatism about trust. Second, I consider how a model for addressing covariance in dimensions of trust can inform philosophizing about trust.

3.3.1 The Integrative Model of Trust

Mayer et al. (1995) develop the Integrative Model of Organizational Trust (see Figure 1). The Integrative Model is so called because it combines research from across the social sciences to examine "characteristics of the trustor, the trustee, and the role of risk," while distinguishing "trust from similar constructs" (ibid., 709). For our purposes, the most salient feature of the Integrative Model is its distinction between trust as a psychological state, dimensions of trust, trust-based actions, and outcomes for updating trust. It is worth noting that, while the Integrative Model is designed for studying trust in *organizational* contexts, there is nothing in the model that prevents its application to other cases and contexts. If we find that the dimensions of trust relevant to organizational cases do not fit interpersonal cases—or vice versa—in my view, this suggests consideration of *why* the particular form of trust does not rightly apply to the case in question. What is salient for present purposes is how the model relates dimensions of trust as bases *for* trust that impact trustor's perceptions and actions.

I noted in §3.2.1 that Mayer et al. define trust as willingness to be vulnerable. Here is their full definition:

> the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party. (ibid.,712)

On conceptual grounds, Ben McMyler (2017) argues that "willingness" or decisions to trust can be misleading, since it is debatable whether trust (and distrust) are voluntary. Indeed, sometimes we simply find ourselves already trusting or distrusting and cannot do otherwise. With that said, one can understand willingness to be vulnerable to another's actions as the psychological state that leads one to assume certain risks in a trusting relationship, irrespective of how voluntary that state is. Furthermore, what is salient in the definition is not that vulnerability (or a particular degree of vulnerability) is necessary for trust. Rather, trust is one's willingness to incur risks by relying on the actions of another party.

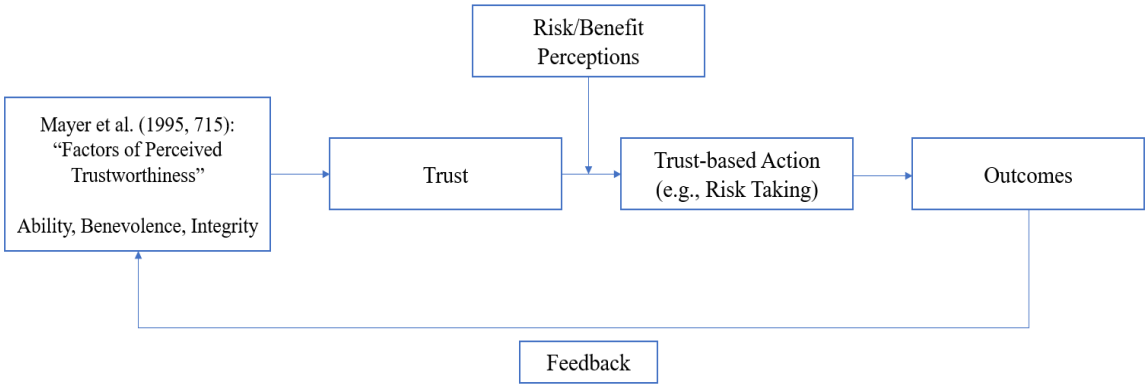

Figure 2. Rendering of Mayer et al.'s (1995) Integrative Model of Trust. For complete model with additional details, see (ibid., 715).

For Mayer et al., trust is sensitive to expectations about a potential trustee's actions. The Integrative Model identifies these expectations in three dimensions of trust. They are ability, benevolence, and integrity, corresponding to the five-factor model in PytlikZillig et al.'s discussion (see Figure 1). As I argued in the previous section, it is plausible to view these dimensions as conditions on trust. In the context of the Integrative Model, they function as inputs for one's willingness to rely. Mayer et al. (ibid, 724–25) argue that trust itself does not involve risk taking. Rather, trust is the "tendency" that leads one to assume risks (ibid.). Of course, one can enter into risky behavior without trusting. For what Mayer et al. call "trusting behavior," the Integrative Model captures risk taking that arises from a position of trust, resulting in what they call "risk taking in relationship" (RTR; ibid.). I contend that it is plausible to understand the role of trust in

RTR as *disposing* trustors to rely on trustees. That is, trust functions as a tendency or disposition to act in a particular way toward a trustee.

On this score, there is a large body of evidence suggesting that trust impacts risk perceptions. For example, in a cross-cultural study of participants in Mexico, Brazil, Chile, the United States, and Spain, Nicolás Bronfman and Esperanza Vazquez (2011) found that trust impacted perceptions of risks and benefits for 23 hazards, including nuclear power, modes of transportation (motorcycles, cars, trains), smoking, surgery, genetically modified food, antibiotics, pesticides, among others. While they found differences across cultures in participant's approval activities related to relevant hazards, these were moderated by different levels of claimed knowledge. Bronfman and Vazquez found that lack of knowledge strengthened the magnitude and statistical significance of correlations with trust factors, suggesting that trust moderates perceptions of risks and benefits (ibid., 1930). Similarly, Norifumi Tsujikawa et al. (2016) found a similar effect for perceptions of risks and benefits for the adoption of nuclear power in Japan, providing insights into public perceptions after the Fukushima Daiichi Nuclear Disaster.[77] Likewise, Bart Terwel et al. (2009) found that trust measures correlated with increases or decreases in perceptions of carbon dioxide capture and storage technologies.[78]

Moreover, the Integrative Model suggests that RTR includes "the behavioral manifestation of the willingness to be vulnerable" (ibid., 724). That is, RTR "is a function of trust and the perceived risk of the trusting behavior" (ibid., 726). In my view, this

[77] See also Visschers and Siegrist (2013).
[78] For more examples, see Siegrist (2000), Siegrist et al. (2003), Siegrist and Cvetkovich (2000), Song (2014), Vainio et al. (2017), Tumlison et al. (2017), Earle and Cvetkovich (1995; 1997; 1999), Earle & Siegrist (2008), Nakayachi and Cvetkovich (2010), and Midden and Huijts (2009). For a dissenting views, see Eiser, Miles, and Frewer (2002), Sjoberg (2001), and Viklund (2003)—to which Siegrist (2021) responds.

seems plausible. Castelfranchi and Falcone's (2010) conceptual relationship between trust as a psychological attitude and trust as an action provides theoretical reasons to associate trust and intentional behavior in the way the Integrative Model suggests. There is also empirical evidence suggesting that trust can facilitate cooperative behavior. For example, Joseph Hamm et al. (2015) found that trust plays a role in cooperation between rural land owners and resource management institutions.[79] Hamm et al. deploy the model for assessing the impact dimensions of trust that I examine in the next section.

A final component in the Integrative Model is that trust relationships are not static in most cases. Rather, the outcomes of trust can impact the degree and continuation of a trust relationship. That is, having entered into trust behavior, one is sensitive to whether or not a trustee fulfills one's trust. Trustee performance over time influences each component in the model—trustor's perceptions of a trustee's trustworthiness (i.e., the dimensions of trust), trust itself as a tendency to assume risks, risk perceptions, and trust behavior. This can help explain increasing or decreasing levels of trust over time (see Mewes et al. 2021). Additionally, the model orients assessments of trust over time to trustor perceptions of trustworthiness—that is, to dimensions of trust.

3.3.2 Addressing Covariance: The Higher-Order Dimension Model

But suppose one finds that trust is declining and wishes to know what dimensions to address to increase trust. Since most cases of trust are not unidimensional, one limitation to multidimensionality is that dimensions of trust can covary, obscuring the dimensions

[79] See also Hamm et al. (2019), Jones and George (1998), Siegrist et al. (2003), and Siegrist et al. (2007). Bauer et al. (2019) question the link between trust and cooperative behavior. The upshot of their study is that trust's link with behavior is highly contingent on situational factors.

influencing trust. Put differently, if dimensions are not sufficiently independent, it becomes difficult to identify which dimensions are driving effects; and aggregated covariance often obscures precisely what we wish to know. One strategy for addressing this problem is to combine relevant dimensions of trust into a single dimension. However, as PytlikZillig et al. (2016) suggest, this significantly reduces model fit. As an alternative, Joseph Hamm and Lesa Hoffman (2016) introduce a higher-order variable to trust models that can preserve fit and allow for identifying the influence of lower-order dimensions. In this section, I examine Hamm and Hoffman's strategy and argue that it suggests a way for combining the insights of pragmatic pluralism with empirical research.

To illustrate their view, Hamm and Hoffman develop a structural equation model to measure the trust and cooperative behavior of rural landowners in Nebraska with natural resource management representatives.[80] They hypothesize that a trust measure of 19 items would indicate six latent dimensions (Figure 3). The six dimensions are dispositional trust, care, competence, confidence, procedural fairness, and salient values similarity. In the test case, the dimensions fit well with the data.[81] However, the latent constructs are highly correlated with each other (r = .9), indicating that they *share* over 80% of the variance and obscuring the dimensions influencing trust scores.

[80] Hamm and Hoffman highlight many of the benefits of using structural equation modeling (SEM) to study concepts like trust. While the use of SEM in trust research lies beyond the purview of this essay, see Siegrist, Cvetkovich, and Roth (2000), Frewer et al. (2003), Van Slyke et al. (2009), Colquitt and Rodell (2011), Pirson & Malhotra (2011), and Smith et al. (2013).

[81] Specifically, the following four measures of fit: comparative fit index (CFI) = .96; (Tucker–Lewis index) TLI = 0.96; root mean square error of approximation (RMSEA) = 0.05; standardized root mean square residual (SRMR) = 0.03 (ibid., 91–93)

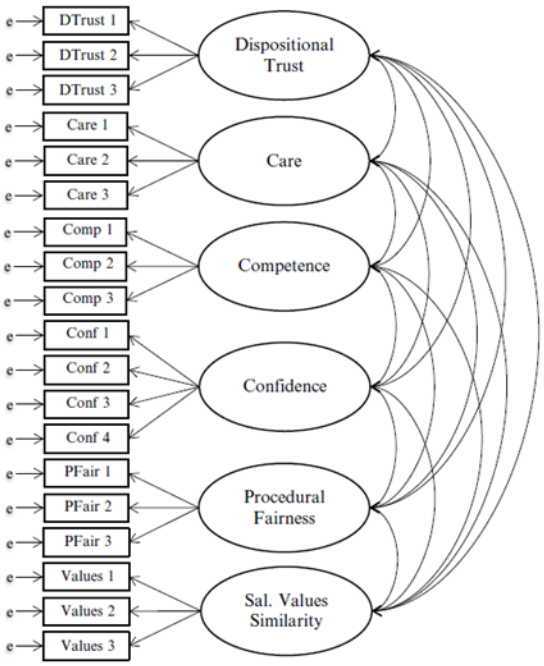

Figure 3. *Six-factor model*. Boxes are observed indicators. Circles are latent constructs. Unidirectional arrows indicate factor loadings and bidirectional arrows indicate correlations. "e" is the variance in the item that is not related to the factor (item "error"). Reproduced with permission from Springer Nature.

Significant overlap among the six dimensions suggests that they might be tracking a single underlying dimension. If the combined dimension fits the data well, then we have reason to think that the underlying dimensions are not as distinct as conceptualized. To this end, Hamm and Hoffman collapse five dimensions into a single dimension, keeping dispositional trust separate. Yet, they found that a "likelihood ratio test revealed that this two-[dimension] model fit significantly worse than the highly correlated six-[dimension] model" (2016, 93). This is because the 16 underlying measures for the six dimensions do not correlate well with a single dimension. The challenge, then, is to associate dimensions of trust in a way that preserves acceptable model fit.

To this end, Hamm and Hoffman introduce a model that takes five factors to indicate a higher-order dimension that can correlate more or less with dispositional trust (Figure 4). Unlike the single dimension model, the higher-order dimension model is closer to the fit of the original model.[82] Further, it preserves the conceptual distinctions of the original six-dimensional model. That is, a higher-order dimension permits the "investigation into the relative influence of the five lower-order constructs by evaluating the factor loadings and testing direct effects of the variance of the lower-order factors that was not shared by the higher-order factor" (ibid., 94).[83] This allows for contextual sensitivity for deploying models across contexts, where different dimensions may have more or less influence on trust and trust behavior.

[82] CFI = 0.96; TLI = 0.95; RMSEA = 0.05; SRMR = 0.03 (ibid., 93).
[83] This is possible by analyzing the factor disturbances ("d" in Figure 3) and the effect of the higher-order factor on cooperation.

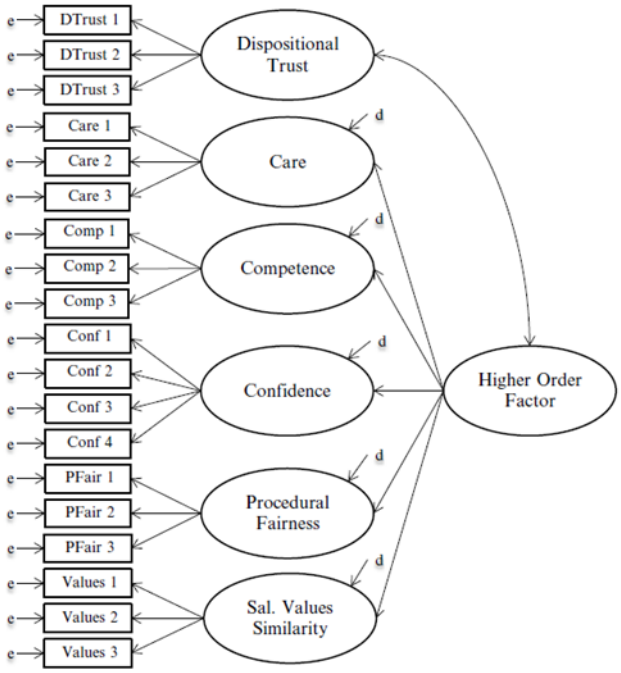

Figure 4. *Higher-order Factor Model of Trust* (2016, 94). Boxes are observed indicators. Circles are latent constructs. Unidirectional arrows indicate factor loadings and bidirectional arrows indicate correlations. "e" is the variance in the item that is not related to the factor (item "error") while "d" is the variance of the lower-order factor that is not related to the higher-order factor (factor disturbance). Reproduced with permission from Springer Nature.

The promise of this approach is that it can allow researchers to identify the dimensions that influence trust from those that do not in a particular case. This can add clarity to both theorizing and intervening in particular contexts. As I argue in the next chapter, this could help scientists and science communicators discern relevant dimensions of public trust.

3.3.3 Replying to Jones' Objection

Recall Jones' (2004) objection to pluralistic approaches to trust. She argues that such views "identify too heterogeneous a class of dependencies to support useful generalizations" and, therefore, are unhelpful for social scientific inquiry (ibid., 4).

The preceding discussion suggests three replies worth mentioning here. First, in §3.2, I argued that the rise of multidimensionality in empirical trust research supports a pluralistic view of trust. The theoretical value in unrestricted pluralism is that it underscores how conditions on trust can vary over time and across contexts, without supposing *a priori* what those conditions are or could be. This facilitates plausible explanations of empirical findings.

For example, in a Swiss study of public perceptions of and willingness to accept genetically modified foods (n = 999), Siegrist et al. (2012, 1402) found that people who viewed scientists as honest and competent were more likely to accept the results of field experiments. This is unsurprising and replicated others' findings. What was novel in the study was that they looked for other factors explaining variance in the sample. They added measures for moral conviction about gene technologies, procedural fairness, and outcome fairness. Attending to these dimensions, and interactions between them, impacted participants' willingness to accept scientific reports about the safety and efficacy of gene technologies (ibid., 1401). For instance, participants who had low scores on moral conviction about gene technologies, outcome fairness (i.e., "how fair people perceived the decision to conduct the [gene technology] field trials in their neighborhood," ibid., 1397) explained more of the variance in their willingness to accept reports. By contrast, outcome fairness was less important for those who had strong moral

convictions about gene technologies (ibid., 1401). Attending to the different dimensions of trust helps to explain how different people are sensitive to different conditions when trusting. It is not that one dimension (competence, say) was indicative of trust while other dimensions were not. Rather, the point is that participants' willingness to rely on scientific authorities was contingent on different factors. Pluralism helps to explain this variability in trust.

Second, considering the dispositional function of trust helps to explain how different dimensions result in cooperative social practices. Rousseu et al. (1998) identify trust in the psychological state that causes one to choose or behave in different ways (see §3.1.1). As Castelfranchi and Falcone (2010) argue, we can view this psychological state as an attitude that leads one to depend on others. In framing their analysis, they admit that an understanding of trust "will not be achieved by looking for just one unique monolithic definition" (ibid., 17). This is because researchers operate with various definitions and measures for trust. "We also do not want to gather just a list of different meanings," they continue, but "we can accept this 'family resemblance' as a possible result of the conceptual analysis" (ibid.). More recently, they argue that trust is "*dispositional:* [that is,] an 'attitude' by the trustor (X) towards the world or other agents (Y)" (2020, 214). This attitude is "hybrid, with affective and cognitive components" and "composite," consisting of beliefs, goals, expectations, evaluations, and dimensions (or "qualities" of a trustee on which trust is conditional) (ibid.).[84] I contend that pragmatism about trust provides a plausible way of uniting each of these points.

[84] Interestingly, Castelfranchi and Falcone also distinguish between dimensions of trust and "external conditions" (ibid.). I take it that this tracks the distinction between conditions of and conditions for trust in Chapter One.

Third, pragmatic pluralism helps to frame future collaborations between social scientists and philosophers. For example, Siegrist argues that social scientists can reliably identify "in which situation, which type of trust correlates with the perceived risk and the acceptance of a hazard," where different 'types of trust' are cases in which different dimensions impact risk perceptions (2021, 487). However, despite "more than 25 years of trust research in the risk domain," Siegrist continues, "the fundamental questions (whether or not trust is a causal factor and how strong the effect of trust is) remain open for debate" (ibid). Given the discussions of dimensionality and modeling in this chapter, I maintain that pragmatic pluralism can help social scientists identify different forms of trust and explain how they result in cooperative behaviors. In turn, discovering the dimensions (or, in my terms, conditions) of trust that dispose people to rely on others in specific contexts can provide evidence from which to advance philosophical reflection on trust.

3.4 Concluding Remarks

In sum, I have argued that pragmatic pluralism is plausible given empirical trust research. Both empirical and philosophical inquiry can together illuminate dimensions or conditions that influence trust. Empirical trust research can reveal the salience of some dimensions of trust, enhancing discussions of forms of trust in particular cases, as Rousseau et al. (1998) suggest.

Explaining what these dimensions are and how they influence trust should not be divorced from empirical investigations, but those empirical investigations are only as

clear as the distinctions in the dimensions influencing trust. Accordingly, there is important philosophical work to be done examining how different contextual factors might further influence trust. This provides promising opportunities for collaborations between philosophers and social scientists in inquiring about the nature and normativity of trust.

4.0 DIVERGENT VALUES AND GROUNDING TRUST IN SCIENCE

Science provides crucial information for understanding and navigating the world around us.[85] For non-experts, incorporating scientific information into their beliefs and actions is reasonable only if they trust scientists to report current findings. This raises an important question: what grounds non-expert trust in science? This chapter examines a problem that arises from the influence of values in science, viz., what I call the problem of *value divergence.* The problem arises in two steps. First, many philosophers of science accept that science is value-laden.[86] In conditions of uncertainty, values help scientists weigh decisions about what projects to pursue, the best means for collecting and evaluating evidence, and how to frame and communicate results. Second, empirical evidence suggests that values similarity between parties directly affects trust and, in turn, perceptions of risks and benefits.[87] So, the problem is that when values diverge (i.e., when there is value *dis*similarity), it can be rational to reduce or suspend trust. In turn, I argue that grounding trust in science is fundamentally normative, involving a view about the values that *should* influence science in relevant cases. In this way, determining the appropriate grounds for trust in science contributes to a norm-based approach to managing values in science.

My focus in this chapter is on how value divergence can undermine non-experts' trust in science, and possible strategies for grounding trust in light of value divergence. Heather Douglas notes that the values scientists develop in their training, as well as

[85] For evidence that trust in science has remained relatively stable, see Funk (2017).
[86] Recent literature on the scope and impact of values in science is vast. For a recent overview of approaches and views in the literature, including dissenting views, see Elliott (2022).
[87] Siegrist et al. (2000); Siegrist (2021).

demographic differences, can "create divergences between the scientific community and the general public" (2009, 172–73). Drew Schroeder (2021, 551) underscores the impact of these divergences:

> If a scientist discloses her values to me and I see that they align with my own, then (assuming I have reason to trust her scientific competence), it seems that I should accept her conclusions even if I don't understand how her value judgements impacted those conclusions…[In the same way, distrust could be] caused by the fact that the values of scientists may diverge from the values of any individual member of the public. To promote public trust in science, then, it seems that we need to eliminate that divergence.

The problem in value divergence, as Matt Brown (2020, 165) argues, is that disagreements in values can cause a type of incoherence, making it reasonable to withhold trust. Douglas, Brown, and Schroeder develop possible strategies for addressing divergence, involving political, social, institutional, and ethical interventions. The upshot from an examination of these, I will argue, is that trust in science is itself value-laden, raising considerations of what we value science *for* and take to be the appropriate conditions for relying on science.

I proceed as follows. In §4.1, I explain how diverging values poses a problem for trust. Drawing on the account developed preceding chapters, I argue that trust's role in disposing trustors to rely on trustees helps to explain *how* value divergence affects public trust in science. In §4.2, I examine five plausible approaches for grounding trust in cases of value divergence, namely transparency, value alignment, political legitimacy, ethical reasoning, and high epistemic standards within scientific disciplines. I argue that each approach faces important practical and normative challenges that leave value divergence unresolved. In §4.3, despite their limitations, I argue that we should view the strategies in §4.2 as practical, though value-laden, means for grounding trust in science. I conclude

that this is a fundamentally *normative* task that contributes to norm-based approaches to values in science.

4.1 TRUST, VALUES, EXPERTISE

In this section, I explain *how* divergence can undermine public trust. I argue that the influence of value similarity on trust and the role of values in establishing the aims of and constraints on scientific practice shows how divergence in values can undermine trust.

4.1.1 TRUST

There are three points to highlight about trust for present purposes. First, there are ambiguities in what it means to 'trust science'.[88] Some ambiguities are relational. One could mean that a person or group trusts a specific scientist or group of scientists to speak on a particular matter. It could indicate that one trusts any or all scientists in general. Different still, one could trust science as an institution that communicates expert consensus, independently of any specific scientist(s). In this chapter, my interest is not in demarcating expertise, although I argue that appeals to expertise have an important role to play in managing trust (see §2.3). For present purposes, I operate with a minimal notion of expertise, where an expert is any individual or group that one recognizes as

[88] One way to avoid this complexity and focus on values is to operate with a deflationary notion of trust as "mere reliance," as John (2017) and Schroeder (2019) do. As Schroeder argues, this underscores the impact of value divergence in cases where conditions are weaker than paradigmatic cases of trust. The advantage of this approach is that it avoids the theoretical difficulties associated with understanding trust and trustworthiness. However, as is clear below, I think this obscures the impact of value divergence and, more importantly, blocks a plausible strategy for addressing it.

being in a better position than oneself to know about a relevant matter, whether as a matter of knowledge or skill.

Second, pluralism about trust suggests that trust in science can vary across individuals and groups. Two parties can be differentially disposed to rely on science. For instance, one might think that trust in science requires that science contribute to social progress, whereas someone else trusts because it promotes reliable, disinterested knowledge. Such preferences are often sensitive to one's context and history. As Gloria Origgi remarks:

> Why we trust, how we trust, and when [we have] reasons to trust are features of our cognitive, social and emotional life that are highly dependent on how the information landscape is organized around us through social institutions of knowledge, power relations and systems of acknowledging expertise… (2020, 80)

While these contextual factors do not necessitate the conditions of one's trust, they interact in important ways. For example, knowledge of the Tuskegee Syphilis Study (1932–1972) that examined the progression of untreated syphilis in Black men without their consent could reasonably lead a member of a historically-disadvantaged group to distrust medical experts.[89] Overcoming that distrust, if we should, can involve conditions on trust that do not apply in other cases.

Third, the empirical findings discussed in Chapter 3 help to clarify *how* value divergence can undermine trust. Timothy Earle et al. argue that trust is responsive to "the judged similarity between the [trustor's] currently active values and the values attributed to the [trustee]" (2007, 9).[90] Similarly, Lisa PytlikZillig et al. (2016) found that trustors differentially respond to a trustee's perceived benevolence or care, integrity, competence,

[89] See Jones (2008).
[90] For an overview of the empirical literature on trust, see chapter three and Siegrist (2021).

and values similarity. Focusing on the last of these, Michael Siegrist (2021) explains that researchers repeatedly find that (dis)similarity in values influences trust and, in turn, perceptions of risks and benefits.[91] In short, trust has a mediating effect on perceptions of risks and benefits—where trust increases one's perceptions of benefits as beneficial and decreases one's view of risks as risky (See Chapter 3, §4). Likewise, the *absence* of trust can increase perceptions of potential risks as risky and decrease assessments of benefits.

Consider an example. Maya Goldenberg (2021) examines factors influencing vaccine hesitancy. Between confident adoption and outright refusal, vaccine hesitancy identifies an attitude of ambivalence toward vaccines (ibid., 3). Goldenberg considers how vaccine hesitancy is highly dependent on perceptions of relevant risks and the value of vaccines, as well as geographic, social, ideological, and historical factors. Contrary to the suggestions that all vaccine skeptics are ignorant or stupid, she argues that trust explains what non-experts know and do about vaccines.[92] She examines at length how social media, histories of medical racism, and the commercialization of medicine can in their own ways undermine trust in vaccine science. What emerges from her work is that differences in values, especially the importance of public health and personal autonomy, explain different levels of trust and, by extension, vaccine uptake.

---

[91] See Siegrist et al. (2000), Siegrist et al. (2012), Connor and Siegrist (2010), Earle and Siegrist (2006; 2008), Poortinga and Pidgeon (2003), Midden and Huijts (2009), Gaskell et al. (2004), Cvetkovich et al. (2002), Trumbo and McComas (2003), and Slovic (1993; 1999), among others. Beyond risk perceptions, Earle et al. (2007) apply this model of trust to cooperation in general, where trust mediates on levels of cooperation.

[92] To be sure, misinformation and confusion are influential in public debates about vaccines (e.g., views that vaccines cause autism or contain microchips). Yet, Goldenberg argues that this can lead us to misunderstand the nature of many debates about vaccines; see her discussion of the 'war on science framework' (2021, 11-14). By contrast, Goldenberg frames public debates about vaccines as a "crisis of trust" (2021, see 15-16 for a summary of the framework).

This helps to explain *how* diverging values between scientists and members of the non-expert public can *disrupt* trust (2021, 551). However, as I argued in §2.3.3 and §3.1.2, the absence of trust does not entail *distrust*. One may rely on experts even when they know their values do not align with those of scientists. For instance, when preparing for worst-case flooding scenarios, it may be that most-likely flood projections are sufficient for developing policies because there are no alternatives. In such cases, policy makers can rely on the scientific findings *despite* value divergence. In more extreme cases, however, divergence may reveal deeper differences, cutting to less tractable metaethical and ideological disagreements, and disposing parties *not* to rely on each other. Accordingly, as Siegrist et al. remark, it is important "to be specific about the values that the participants use to assess their value similarity with another person or organization" (2012, 1395).[93] That is, in addition to questions of trust, we should consider what we mean by *values in science*. This is the task of the next subsection.

4.1.2 VALUES IN SCIENCE

The problem of value divergence derives its force from the value-ladenness of science. I take it that there is consensus among philosophers of science that values *do* play a role in science—whether and how they *should* is hotly contested. For our purposes, a value is something regarded as desirable or worthy of pursuit (see Elliott 2017, 11).[94] Sometimes

[93] This point is anticipated in Earle and Michael Siegrist (2008), who found that trust in science communication is sensitive to judgments of procedural and outcome fairness, relative to whether a trustor ascribes moral importance to the topic.

[94] The term 'value' is utilized in different ways by those writing on values in science, labeling, as Elliott explains, "a very wide array of phenomena that ought to be treated in different ways" (2017, 4). For a thorough and succinct overview, see Elliott and Richards (2017) and Elliott (2022). For different conceptions of value, see Ward (2021), Brown (2018; 2020), Rooney (2017), Biddle (2013), Douglas (2009), Clough and Loges (2008), Schwartz and Bilsky (1987), Scriven (1974), among others.

'value' refers to desires, commitments, identities, and ideals for individuals, groups, and institutions. Other times, it is felicitous to say that something has value, whether instrumentally or intrinsically, including actions, events, objects, and agents. While the following discussion of values is inclusive of both senses, this framing of values allows that people can disagree, sometimes deeply, about which values are correct. For instance, the fact that someone holds racist values does not entail that those values are desirable all things considered. Rather, the importance for value divergence is to see how differences in values of various kinds (epistemic, ethical, social, economic, political, and so on) can impact trust in science and science communication.

My argument does not require a specific typology of values. Early debates about values in science draw strong distinctions between epistemic and nonepistemic values.[95] Examples of epistemic values include a theory's scope and explanatory power.[96] Nonepistemic values, by contrast, range across ethical, social, political, and religious values. As examples, these might include promoting economic development, public health, or environmental protection. In addition to individual scientists and groups, values can operate at an organizational or institutional level.[97] For example, organizations may value standardization because it maximizes efficiency within a research community. Recent work, especially by feminist philosophers of science, suggests that a sharp demarcation between epistemic and nonepistemic values may be untenable.[98] For exploring the impact of value divergence, my view only requires an acknowledgement

[95] Douglas (2009, Chapter 3) offers a good overview of twentieth century debates about values in science.
[96] Some values, such as simplicity, are less clearly epistemic, since simpler theories may provide more of a practical advantage (e.g., easier to operationalize) than an epistemic advantage (i.e., be closer to the truth).
[97] For more on the social infrastructure of science, see Contessa (2021) and Rolin (2015).
[98] Influential examples include Rooney (1992), Longino (1990; 2002), Nelson (1993), Anderson (1995; 2004), Wylie and Nelson (2007), Intemann (2001; 2005), and Kourany (1998; 2010), among others.

that values of various kinds can interact with our aims in relying on science, especially for non-experts.

Accordingly, following Brown (2020, 18–22), I think we should focus on values' role in navigating contingencies in the course of inquiry. Elliott (2022) identifies four areas in which values play a part. First, values can *steer* research, individually and institutionally. This can involve prioritizing certain research questions and projects, as well as external incentives, including reputational and financial factors. During the COVID-19 pandemic, for example, researchers shifted focus to understand the virus and to develop vaccines in conjunction with governmental priorities and industry investment. In this way, by steering research, values can impact the aims of science. Second, values play a role in *doing* research. This includes experimental design, data analysis, and the interpretation of findings. It might seem like values relevant for doing research are confined to the most policy-relevant domains, but as Kent Staley (2017) shows, values play a key role in determining demands for evidence and announcing discoveries in areas as theoretical as physics. Third, values influence how science is *managed*. Managing science includes a range of activities, from developing norms and policies for research ethics to questions about the structure of research teams and institutions. Areas two and three illustrate ways that values can impact constraints on science. Fourth, values influence how science is *applied*.[99] This can range from developing public policy to communicating public health risks. For instance, the US Food and Drug Administration must balance risks and benefits when approving drugs for public use, including relevant scientific information and social or ethical considerations. By their role in steering,

[99] For considerations of framing and morally responsible scientific communication, see McKaughan and Elliott (2018).

conducting, constraining, and utilizing scientific research, values can result in different research practices and findings.

Elliott (2022, 6–7) distinguishes between values and value *judgments*, where the latter are decisions involving weighing values.[100] To explain this influence, Zina Ward (2021) identifies four ways in which values can impact decisions or choices. She argues that there are two ways that values can provide reasons for action. First, values can be *motivating reasons* in the sense that values (epistemic or non-epistemic) motivate choices. For example, the commercial promise of a research topic might motivate someone to choose it relative to alternative topics. Second, values can provide *justifying reasons* for choices. One might justify choosing between competing climate models, for example, on the basis that one model is less expensive or less prone to computational errors.

In addition to serving as reasons, Ward argues that values can be causally efficacious. First, values can be *causes* for choice, even if they do not motivate or justify. As "causal effectors," values can help explain institutional, financial, and design choices. Consider hypothesis acceptance. Robyn Bluhm (2017; cited by Ward ibid., 56) argues that accepting a hypothesis is value-laden *because* choices leading to acceptance are value-laden, particularly in experimental design and data collection. Second, what we value can be the result of value choices. That is, scientific choices can influence what one values. For example, in his discussion of Theo Colburn's work on endocrine-disrupting chemicals, Elliott notes that even if Coburn's goal was not to promote "a particular set of

[100] Elliott explains that value judgments take multiple forms, including when one judges whether a particular quality or outcome is desirable, whether one has achieved a particular outcome, how to navigate a trade-off between values, or how to weigh risk (ibid.).

values, her *choice served* the value of promoting public health over alternative values, such as promoting the short-term economic growth of the chemical industry" (2017, 12; my emphasis). In this way, what we come to expect and value in science is causally responsive to certain (value-laden) choices.

Both as reasons and causes, differences in values helps to explain breakdowns in trust, since one may reject certain values as legitimate reasons or regard their role as inappropriate. In their motivating, justificatory, or causal roles, different values can impact uptake on the basis of trust, since differing values can result in different aims, activities, and findings. Some cases of divergence may be extreme. For example, the history of science is replete with examples of sexist and racist values motivating and causing research, in Ward's terms, to "understand" the alleged inferiority of racial minorities or women.[101] In other cases, value divergence may not rise to the level of conflict. For instance, in balancing economic, social, and political factors, a city manager may consult expert climate scientists to assess flooding risks (from Elliott (2020)). If the values influencing model predictions (e.g., predicting most likely scenarios) do not reflect the city manager's priorities (e.g., preparing for worst-case scenarios), she has reason to seek information elsewhere. This need not be an indictment of the scientists' work or integrity. The city manager can consistently recognize the scientists' expertise and their work's incongruence with her aims. As a practical matter, in the absence of plausible alternatives, relying on the experts may be the best the city manager can do. This gives her reason to rely *despite* divergence and without any disposition to do so.

[101] See Kourany (2010; 2020).

There is one final consideration for the impact of values in science on trust. Elliott proposes a norm-based approach to values (2022, 49–54). According to this approach, Elliott suggests that "values can appropriately influence science as long as scientists and scientific institutions follow the norms for good scientific practice" (ibid., 49). Such norms are intended to help determine the normatively appropriate values and role(s) for those values in science and science communication. For example, Elliott illustrates his approach with the norm of transparency, meaning openness about data, methods, values, and interpretive judgments. I discuss transparency at great length in §4.2.1.

Here, there are two things to emphasize. First, Elliott suggests that additional theoretical and practical work is necessary for implementing norms. For instance, there may be conflicts between transparency and other norms, such as privacy (e.g., in medical data). Theoretical reflection, Elliott argues, can help prioritize norms and resolve conflicts. I explore the most promising means for doing this in §4.3. Practically speaking, norms are of little use if not clearly implemented. So, there is important work in connecting norms with guidelines and procedures.

Second, recall that in concluding Chapter Two (§2.4) I indicate one of the limits of pragmatic pluralism about trust. That is, pragmatic pluralism provides a strategy for describing different cases of trust within and across contexts. Yet, it does not determine when trust is normatively appropriate. In the same way, determining the right values and their appropriate role(s), both in science and science communication, can depend on what one takes to be the appropriate grounds for trust. If determining those grounds is fundamentally normative, as I argue, then consideration of trust in science is both

consistent with Elliott's norm-based approach and can illuminate salient features for thinking about norms and applying them to science.

4.1.3 Limits for Appeals to Expertise

One might object that value divergence presents a problem for trust only if a trustor operates from a position of epistemic hubris, having an irrationally confident view of one's opinion and discounting the opinions of (more) qualified others. After all, we trust scientific experts in large part because they have the training, skills, and perspectives that make them more likely to know about a relevant matter. For example, when developing public health responses to viral outbreaks, we appeal to virologists, economists, and epidemiologists (among others) *because* they are more likely to be correct about the development and impact of viruses. While area-specific expertise may not be sufficient for good policy, it is arguably necessary. For value divergence, one could argue that an expert should have a better sense about trade-offs and strategies within her field, suggesting that differences in values might be explained by lack of expertise.

Consider a strong principle supporting the appeal to expertise in cases of value divergence. Elizabeth Fricker argues:

> where I know another to be epistemically expert relative to me on a topic, it is not just rationally permissible, but rationally mandatory for me to accept her judgment in preference to my own, just so long as I have good ground to trust her sincerity. (2006, 243)

This rational mandate to accept expert testimony underscores two issues. First, notice that our acceptance of expert testimony is conditional on trust. That is, only when we have good ground for trusting are we rationally mandated to accept their testimony. Of course, experts are fallible. What matters is that, according to Fricker, it would be irrational to

refuse the testimony of an expert *when* one trusts the expert. For present purposes, we can set to one side whether this mandate is principally epistemic, prudential, moral, or a combination of the three. As I argue in previous chapters and §2.1, trust can come in multiple forms according to various grounds. Moreover, determining whether one's grounds for trust are *good* requires not only a description of possible grounds but a normative view about the aptness of those grounds. This can be done. But resolving value divergence through an appeal to expertise must discern not only the actual grounds of trust in experts, but also that they are the correct ones.

Fricker offers two plausible grounds for trust, namely sincerity and competence. That an expert is competent may be a matter of definition since an 'incompetent expert' seems like an oxymoron. If the expert is incompetent in her purported area of expertise, then she may be only a purported expert. Of course, expertise may require *more* than competence. Now, consider sincerity. Empirical evidence suggests that we prudently eschew information from those we perceive as having deceptive motives.[102] Expert testimony might present a special case, however. For example, Thi Nguyen (2022) argues that the reasons that persuade experts in their deliberations are often inaccessible to non-experts. In part, this is unsurprising since part of expertise is recognizing evidence and its significance. However, when testifying as experts, this presents a problem Nguyen calls "epistemic intrusion" (ibid., 9–14). The problem is that public justification of expert opinion can require experts to offer reasons accessible to the public, which may not be the reasons that persuaded the experts in the first place. In this way, epistemic intrusion can severely limit what experts can justifiably say in public testimony or motivate a type

[102] See Handley-miner et al. (2023) for evidence that this is the case.

of deception. I agree with Nguyen that there is a trade-off between trust in experts and transparent communication. In §4.2.1, I return to epistemic intrusion to underscore the need for higher-order value judgments in science communication.

There is a second point from Fricker's mandated deference to experts. The previous two subsections on trust and values in science provide evidence against straightforward appeals to expertise *in cases of value divergence*. Expertise in a particular domain does not entail value expertise. We need only consider examples from the history of science in eugenics, 'race' science, or sexist research to object on value grounds. Some cases might involve the misuse of scientific research, such as appeals to the laws of thermodynamics to bar women from education.[103] As I argue below, this does not entail that we should seek out only scientific research that conforms to our values. The point is rather that the impact of values on research practices and findings can give one good reason to resist expert judgment (or, by one's lights, *purported* expertise). Indeed, as Fricker immediately appends to her thesis: "Where there is not good ground to believe an informant trustworthy, however, epistemic self-governance entails that we should not accept the reports of others" (2006, 243). What is required, then, is "good ground" for trusting experts.

How should we discern trustworthy experts? Several philosophers propose criteria to help non-experts identify experts within a domain.[104] While views differ in their particulars and emphases, as Neil Levy remarks, they "converge in identifying *credentials, track record, argumentative capacity, agreement with the consensus*, and

[103] See Zschoche (1989) for a discussion of the case.
[104] See Anderson (2011), Blancke et al. (2017), Brennan (2020), Guerrero (2017), Fricker (2006), and Goldman (2001).

intellectual honesty as criteria by reference to which we can choose between experts" (2022, 110; emphasis original). It seems to me that each of these possible indicators of expertise provides ample prudential grounds for deferring to experts. When an expert is credentialed, experienced, and so on, the burden lies with me to have sufficient reason(s) to forgo the expert's word. Resisting deference to a clear expert without compelling reason is practically unwise, if not irrational, as Fricker argues.

We can grant this much and yet recognize limitations for securing trust in science through appeals to expertise. Two limitations are worth consideration here. First, as Levy argues, "cues for expertise don't correlate well with its actual possession" (2022, 112). This is because cues can be mimicked, obscuring demarcations between expertise and the *appearance* of expertise.[105] This results in what Levy calls "epistemic pollution" (ibid.). Epistemic pollution occurs when, deliberately or inadvertently, "other agents shape our environments in ways that leave individual cognition even worse off than it might have been" (ibid., 110). When pseudo-experts ape genuine expertise, reliance on those purported experts can leave one in a worse epistemic position. Moreover, Levy continues: "the fact that such deception is widely known to occur reduces trust in legitimate sources" (ibid., 112). The limitation for securing trust is that one often cannot discern legitimate from counterfeit expertise without some level of expertise or some level of preexisting trust. If the latter, then appeals to expertise rest on trust, not the other way round.

[105] See Guerrero (2017).

The problem of counterfeit expertise is intensified by a second limitation, namely the presence of expert disagreement.[106] Good-faith disagreements between experts need not undermine trust. For example, it may be that experts disagree about minor, though non-trivial, points, while agreeing in general. Alternatively, experts can disagree on fundamental questions around a topic, diverging on how to frame research questions, appropriate methods, relevant evidence, and more besides. Appealing to expertise to adjudicate the forms of disagreement requires some level of trust (or lack of trust) on the part of non-experts. Moreover, good-faith disagreement can be imitated. For example, Naomi Orsekes and Erik Conway's (2010) *Merchants of Doubt* chronicles how leading scientists introduced disagreement and uncertainty on topics ranging from tobacco health to climate change. Importantly, the antagonists of Oreskes and Conway's account were all preeminently credentialed, experienced, and respected. So, this is not a case of pseudo-experts introducing disagreement by aping expertise. The problem is more subtle. It is legitimate experts operating outside their area of expertise. For non-experts, recognizing genuine expert disagreement, as much as policing the boundaries of expertise, can rely on value similarity and preexisting trust. So, it is no surprise that, for example, non-experts might rely on shared political and economic values when trusting Fred Seitz, a former president of the National Academy of Sciences and university president, about the health impacts of tobacco, since non-experts lack the expertise to adjudicate between competing views.[107]

---

[106] There is a large and growing literature on disagreement. For helpful overview, see Frances and Matheson (2018). Much of the literature on disagreement investigates *peer* disagreement. In contrast, the cases under consideration here do not involve epistemic peers. The appeal to expertise is an appeal to epistemic superiority.

[107] One way to address problems of epistemic pollution and disagreement is to appeal to institutions, organizations, and structures designed to confirm expertise. For example, rather than relying on a single scientist or research group to speak about climate change, one could rely on summaries for policy makers

In this way, value similarity provides non-experts with a heuristic for navigating deference to experts. When I lack expertise to assess the merits of another's word, similar values provide a higher-order means for assessing that another acts as I would act in similar circumstances. This reveals the important role of trust and values in appealing to expertise to secure trust in cases of value divergence.

In sum, I argue both that expertise is an important factor in managing trust and that, in cases of value divergence, appeals to expertise rely on trust. We want to trust the trustworthy, but determining whether one has good grounds for trusting is difficult in cases where we rely on others for things beyond our ken. In the absence of clear conflicts of interest and disagreement, deference to experts seems wise, even if only as a matter of prudence. Indeed, this is not to say that the hallmarks of expertise are irrelevant in cases of disagreement (in such cases, they may be most important). Rather, in cases of value divergence, where levels of trust are low or nonexistent, I argue that appealing to expertise alone will not resolve the problem. Instead, we need strategies for address value divergence that consider both the role of values in science and of non-expert trust.

4.2 Strategies for Addressing Value Divergence

This section is divided into five parts. Each part proposes a means for addressing value divergence by arriving at the 'right values' and, as a result, securing trust. The proposals

from the United Nations' Intergovernmental Panel on Climate Change (IPCC). For instance, 234 expert authors from 64 countries contributed to the physical sciences working group for the IPCC's Sixth Assessment Report (see IPCC 2023). Alongside producing reports of published climate research, the IPCC provides institutional structures for vetting and disseminating that research. The problem, as Levy argues, is that epistemic pollution "*rationally* reduces trust in institutions" (2022, 126; emphasis original). That is, appealing to institutions to demarcate expertise is only as good as one's trust in the institutions.

are as follows: (1) transparent practice and communication, (2) aligning values, (3) determining values through political legitimacy, (4) determining values through ethical frameworks, (5) developing high disciplinary standards. The views are not mutually exclusive; for instance, political and ethical proposals could be combined in complex ways to meet particular situations. My focus here, however, is on the ways value divergence persists with each proposal. In short, I argue that strategies for addressing value divergence are themselves value-laden. What this suggests, I contend in §4, is that trust in science is itself sensitive to what we value science *for*.

4.2.1 TRANSPARENCY

Philosophers of science have focused on transparency as a means for navigating the value-ladenness of science.[108] That is, transparency's promise is in helping science communicators provide well-established, reliable information that acknowledges the role of values in developing and applying that research. For example, consider again Elliott's (2020) city manager. Imagine a city manager who must develop effective and responsible policy that addresses environmental changes, such as flood risks.[109] The manager *could* design policies for worst-case scenarios. But suppose she consults climate scientists whose models reflect the most *plausible* results given available evidence (i.e., *arguendo*, not worst-case scenarios). If the manager is unaware that her values differ from those that informed the research she relies on, the resulting policies could fail to achieve the

[108] See Ashford (1988, 382–83), Douglas (2009), Elliott and Resnik (2014), McKaughan and Elliott (2018), among others. For reasons to pursue transparency beyond clarity about values in science, see Nosek et al. (2015), Royal Society (2012), and NAS (2018).
[109] See also Parker and Lusk (2019).

protections that she seeks.[110] As a result, a reasonable condition for science communication is that, as Elliott (2017, 14) puts it, "scientists should be as *transparent* as possible about their data, methods, models, and assumptions so that others can identify the ways in which their work supports or is influenced by values."[111] For trust, I suggest transparency operates as a type of integrity that balances both contingency and respect for the autonomy of non-experts while providing reliable information.

To facilitate transparency, Elliott (2020) develops a taxonomy that is sensitive to the aims and contexts of science communication. The taxonomy locates key features of transparency through questions about the purposes, audience, content, and means for communication. The first component of the taxonomy considers *why* one pursues transparency. As examples, one might aim to increase reproducibility, or to facilitate critical interaction, or to hold experts accountable. After determining the purpose(s) of transparency, the second part of the taxonomy addresses *who* is the recipient of the communication. Depending on whether a communicator's audience is other experts, policy makers, politicians, or members of the lay public, what transparency consists in can differ. Next, one must determine *what* is communicated to the audience. This could include data, code, methods, and findings. Alternatively, if dealing with an audience that may struggle to understand more technical material—and how values influence it—one may choose to clarify assumptions or deliberations that influenced the project. Finally, when one knows the purposes, audience, and content of communication, decisions must

[110] For a real-world case involving the UN's Intergovernmental Panel on Climate Change, see Keohane et al. (2014).
[111] For a similar statement, see Douglas (2009, 153): "With the values used by scientists to assess the sufficiency of the evidence made explicit, both policymakers and the public could assess those judgments, helping to ensure that values acceptable to the public are utilized in the judgments."

be made about *how* to communicate the information. Considerations of the communicator, the venue for communication, when information is communicated (i.e., before, after, or during a research project), and the mechanisms for delivering information, each present important choices for what transparency means in a particular context.[112] Overall, this taxonomy directs us to crucial considerations for achieving communicators' aims and goals in pursuing transparent communication.

The immediate advantages of this taxonomy are that it can alert communicators to potential practical difficulties when communicating scientific findings. For instance, in cases where transparency may produce undue skepticism about scientific claims, the taxonomy can help communicators make decisions that avoid or address the problem. Moreover, while one may not know *ex ante* how research will be used (e.g., in peer-reviewed publications), asking questions about potential audience, venue, and so on, can provide impetus for developing norms about value disclosures.[113]

There is a final piece to the taxonomy, however. It is not a step along the way to transparent communication but arises with each step—what Elliott calls "dangers" (2020, 6; 2021, 2). These dangers include wasting resources, creating a false sense of trust, causing confusion, violating privacy, among other concerns. In aiming to avoid these dangers, communicators must weigh alternatives and make decisions about the best means for transparent communication. If one must determine how best to communicate

[112] One worry that might arise for this approach to transparency is that scientists cannot determine *ex ante* to whom they communicate, as John (2015) argues. For example, in peer-reviewed publications, it is nearly impossible in most cases to predict how research might be used. And this likewise could apply to publications aimed at more non-expert audiences. However, I set this worry aside for the purposes of this essay.

[113] We see this in norms for reporting potential conflicts of interest. The impetus behind such disclosures, at least in part, is to help those engaging with the findings detect any inappropriate influence (as well as clarify legitimate influences) on the research.

value judgments to a group of non-experts, that communication itself involves value judgments that can affect trust. Accordingly, Elliott (2021) argues that transparency is itself value-laden.

Does the value-ladenness of transparency undermine its importance? I agree with Elliott that it does not. Rather, as he argues, the taxonomy shifts our focus to those forms of transparency that minimize dangers. At the same time, deciding how best to avoid dangers will involve weighing trade-offs that directly affect choices about the purposes, audience, content, and means for communication.

4.2.2 Transparency's Limitations in Cases of Value Divergence

There are practical and normative limitations to addressing value divergence through transparency. Practically speaking, as Schroeder (2019, 550–51) notes, it is difficult or impossible for scientists to identify all the ways that values affect their work. Once identified, as John (2015) argues, it is often hard to know who one's audience is *ex ante*, presenting difficulties for tailoring value disclosures to Elliott's taxonomy. For example, in most cases of peer-reviewed publication, it is difficult to predict how research might be used immediately or in the future. Supposing scientists could identify all the ways their findings could be utilized, it may not be practical to address all of them.[114]

Moreover, non-experts could misunderstand or misinterpret value disclosures. John (2018) suggests that transparency can lead to inappropriate forms of skepticism, especially when members of the non-expert public maintain inaccurate or false views of science. There is also evidence that transparency may be ineffective for building trust.

[114] For similar concerns, see Havstad and Brown (2017).

For example, Elliott et al. (2017) found that scientists' acknowledgement of values may reduce their perceived credibility. Results varied according to the values disclosed, their congruence with the findings, and whether scientists suggested policy recommendations. For instance, compared with acknowledging values related to public health, expressing values related to economic growth decreases positive affect and trust scores (ibid., 13). Although value divergence between non-experts and scientists had a negative effect when compared with aligned values, positive affect and trust decreased when findings were consistent with scientists' values, even when non-experts shared those values (ibid.). Most importantly, "[n]o context, not even when laypeople share scientist's values, sees a statistically significant increase in positive affect toward or perceived trust in a scientist who acknowledges values" (ibid.). So, transparency might actually *reduce* trust.

McKaughan and Elliott (2018) suggest a more specific strategy. They argue that science communicators should help non-experts understand how values impacted research and clarify how different value judgments would result in alternatives—a strategy they call "backtracking." Two things are worth noting about backtracking. First, backtracking seems practically achievable, at least if a case is similar to the one McKaughan and Elliott discuss (see ibid., 199–207). So, in the city manager case, pursuing backtracking may be sufficient to resolve value divergence. I am more skeptical about deeper value divergence, as in the vaccine hesitancy case. But as a practical matter, I do not think incompleteness is a vice here. Rather, it might provide a basis for additional strategies, which I explore in later sections. Second, McKaughan and Elliott do not presuppose that backtracking is value-neutral (ibid., 208). So, it is not necessary to backtrack to a neutral position about which everyone agrees for backtracking to be

effective. That is, backtracking should incorporate alternative perspectives for achieving two goals of science communication that can be in tension, namely (1) reporting reliable findings and (2) framing information so that it is useful for guiding decision-making (ibid., 198). The thing to notice is that backtracking may introduce higher-order value divergence in the sense that people can reasonably disagree about when and how to backtrack, impacting trust.

Suppose that someone becomes aware that second-order value judgments are involved in science communication. If she suspects that those second-order values do not align with her own, she could reasonably withdraw her trust for similar reasons to first-order value divergence. For example, SteelFisher et al. (2023) examined reasons for trust in public health information during the COVID-19 pandemic. They found that political and private sector influence were the strongest factors influencing low levels of trust (ibid., 333).[115] That is, perceptions that a communicator's aims were not to promote public health revealed value divergence and reduced trust. In this way, the problem of divergent values emerges anew with the pursuit of transparency.

The value-ladenness of transparency combined with practical limitations presents normative questions for the goals of science communication. As John colorfully remarks, "just as publicising the inner workings of sausage factories does not necessarily promote sausage sales, so, too, transparency about knowledge production does not necessarily promote the flow of true belief throughout the population" (2018, 75). He argues that the

[115] Interestingly, participants were sensitive to the roles of communicators (ibid., 332). For example, they found that more than 90% of participants cited scientific expertise as a reason for trusting the Centers for Disease Control and Prevention (CDC), while reference to scientific expertise decreased to 75% and 67% for state and local governments respectively. In the latter case, participants' reasons focused on provisions for effective and compassionate care.

same goes for honesty, sincerity, and openness. John's idea is that because science is crucial for informed decision-making in many cases, sometimes obscuring choices can increase uptake and result in overall better outcomes.

Yet, pursuing a policy of opacity opens science communicators to charges of manipulation and bias. In specific cases, there may be good reasons to disvalue transparency; for instance, if it provides opportunities for bad-faith actors to undermine scientific consensus. But there are clearly good reasons to value transparency; for instance, when it promotes the autonomy and informed decision-making of non-experts. To borrow John's phrase, it is conceivable that knowing what is in the sausage determines whether anyone eats it. Accordingly, the challenge is to determine when and how to pursue transparency.

Nguyen (2022; discussed in §2.3) argues that transparency is a form of surveillance. Sometimes surveillance is justifiable, while absolute surveillance can be repressive. For experts, Nguyen argues that demands for transparency can incentivize deception and neglect the unique perspectives of scientific experts. What we need for transparency, he argues, is to determine when surveillance is appropriate. That determination in part depends on striking the right distribution of cognitive labor and when demands for transparency are justifiable, even if practically cumbersome. Importantly, people can disagree about that distribution of cognitive labor. For instance, one's approach to science communication may be *laissez faire* to maximize the discernment and freedom of scientists. Alternatively, one might think that an ethics of expertise places certain restrictions on science communicators that promote democratic or public goods. I return to these strategies in §4.3.4. My point for the present is that

pursuing transparency requires striking the right balance relative to the values of science communicators and non-experts. So, transparency can leave first-order value divergence unresolved, while possibly introducing higher-order value divergence. For trust, I think Elliott et al. are exactly right: "positive affect and perceived trust when acknowledging values when making policy recommendations is less straightforward; *it depends on the values espoused by the scientists and by laypeople*" (2017, 13; my emphasis).

4.2.3 VALUE ALIGNMENT

How should one respond to value divergence? One approach is to seek out research that is informed by values that align with her own.[116] We can call this the *alignment view* for addressing value divergence. In the city-manager case, the manager could contract with a research firm or institution that projects worst-case scenarios. In some cases achieving alignment might be relatively straightforward. When organizations and institutions with clear ideological motives proffer scientific findings, non-experts may be able to make quick and reliable judgments about the values informing their research. For example, a "pragmatic environmentalist," says Schroeder, "might be confident that scientists employed by the Environmental Defense Fund are likely to share her values" (2021, 552). In this way, the pragmatic environmentalist could be justified in trusting EDF-funded scientists because her values align with scientific sources.

[116] For arguments that alignment can ground trust, see Douglas (2017), Wilholt (2013), Kitcher (2011), and Irzik and Kurtulmus (2019).

4.2.4 LIMITATIONS FOR VALUE ALIGNMENT

There are practical limitations to values alignment. First, in many cases, it will be practically impossible for non-experts to achieve alignment. As with transparency, it may be very difficult, if not impossible, for non-experts to discern all the relevant ways that value judgments influenced research *and* confirm that those values align with their own. This is compounded by the fact that most organizations conducting research avoid framing themselves as overtly ideological, in part because operating under and communicating ideological motivations enhances the impact of *mis*alignment. Second, as Elliott (2020, 3) suggests, some "scientific studies...are so expensive that they are likely to be done very few times, and perhaps only once." If one discovers value divergence, there may be no workable alternatives that achieve value alignment. In this way, other constraints, such as economic or legal limitations, can inhibit opportunities for alignment.

One way to address these practical worries is to limit cases where alignment is necessary. For example, Gürol Irzik and Faik Kurtulmus (2021) distinguish between "basic" and "enhanced" forms of trust. For one to have basic trust in an expert regarding some finding, *p*, is to meet certain necessary conditions.[117] Irzik and Kurtulmus argue that basic trust is warranted when a finding is reported honestly and as the result of reliable scientific methods. However, given the possibility of error, values have a role to play in determining when it is appropriate to communicate a finding. Accordingly, Irzik and

[117] I do not discuss basic trust at length for two reasons. First, with pragmatic pluralism, I have argued against formulating an account of trust in terms of necessary and sufficient conditions. Objections that apply to other monist views apply to basic trust. Second, there are controversial epistemological assumptions in the (2021) formulation of basic trust that are beyond our present purview. For example, the first condition on epistemic trust is that experts believe *p* and communicate that *p* honestly to non-experts (ibid., 4733). However, scientists need not fully believe a result to communicate it honestly or sincerely. Accepting a threshold for belief or acceptance, however, requires consideration of when evidence is sufficient for communication, presenting opportunities for value divergence.

Kurtulmus argue that trust can be enhanced by two conditions. First, when "public welfare is at stake, [experts] make their methodological decisions regarding the distribution of inductive risks with respect to [*p*] in agreement with [non-expert's] assessments of those risks" (ibid., 4735). That is, scientists should attune their tolerance for risk to the values of relevant non-experts. Second, non-experts must have good reasons to believe that the first condition is met (ibid.). So, when non-experts are justified in believing that experts act in line with their values, they are justified in trusting experts' communication concerning *p*. The point for emphasis in enhanced trust is to limit the cases where alignment is required to those where public welfare is at stake. So, pursuing enhanced trust does not deny the practical difficulties I raised for alignment; it only restricts cases where alignment is necessary.

On the one hand, I think that enhancing trust by limiting alignment requirements seems plausible. For example, alignment is more salient in research involving human or non-human animal subjects than in theoretical physics. While it might be difficult and costly to align values in the former cases, one can justify requiring alignment without applying an alignment condition to all research. One route for meeting such a condition is the development of guidelines and procedures for managing value-laden choices in research and communication within a domain.

On the other hand, enhancing trust through alignment raises deeper, normative worries for the alignment view. If the alignment view advises non-experts to seek out research that aligns with their values, it risks politicizing science. For instance, if there is a dispute about what threshold for error provides the optimal balance between risks of false positives and false negatives, it could be rational for different groups to trust the

research that most closely aligns with their values. So, value misalignment may lead those concerned about economic development to seek out different sources for information than those more focused on environmental protection. While there could be (and I think there are) ethical or social reasons for adjudicating such cases, these may introduce higher-order disagreements, transforming differences in managing values in science into deeper metaethical and normative disputes.

Moreover, as Schroeder (2019, 10) aptly notes, seeking alternative sources can result in political gridlock. This gridlock overcomes value divergence by restricting the sources of information that inform one's view. As each side of an issue develops its "own" science, it becomes increasingly difficult to resolve value divergence and detect legitimate or illegitimate uses of values, leading to increased polarization and plausibly epistemically (if not ethically) vicious conduct. As Jason Blakely (2023) describes, "conspiracy theories now plaguing American life ape a certain confused vision of science…certain segments of the populous have created a doppelgänger of science, with its own hypotheses and theories."[118] Blakely notes that this outgrowth in alternative science is in part due to real and perceived overreach by experts. Whether one thinks that experts have overreached may depend on alignment or misalignment with the values, real or perceived, impacting research. So, pursuing alignment, even in a restricted number of cases, can intensify value divergence.

Nevertheless, I think the problem is not with alignment *per se*. For trust, alignment is good when you find it. The trouble is that alignment can exacerbate

[118] Blakely's argument is not for freedom from values, but for a more democratic and humanistic integration of scientific expertise into decision making. In his calls for democratic dialogue, he focuses primarily on populist appeals that tend to the political right in the United States.

problems that intensify value divergence, revealing deeper differences about what one expects of science. In other words, as we saw with transparency, navigating alignment can introduce higher-order value disputes. So, the question is not whether we should pursue value alignment, but *how* we should pursue value alignment—by what means and to what ends. In the remainder of this section, I examine three strategies for how we might address this question given value divergence.

4.2.5 DEMOCRATIC LEGITIMACY

One way to approach value divergence is by managing values in science by political means. This strategy has been popular and influential in literature on values in science and has much to commend it.[119] The promise of democratic approaches is that rigorous scientific methods can be combined with politically legitimate procedures for determining which values influence science. For example, Schroeder (2019, 553) proposes what he calls the "democratic values proposal." The proposal aims to foster situations in which, Schroeder argues, "good science (at least in its primary analyses) will speak with a single voice and will offer a common reference point—common ground that can serve as the starting point for public discourse."[120] There are details for this proposal that can and should be spelled out; for instance, the cultivation of local, national, and global mechanisms and institutions that facilitate deliberation and representation. Here, I consider the theoretical prospects for settling value divergence through political means.

[119] For examples, see Kitcher (2001; 2011), O'Connor and Weatherall (2019), Brown (2018; 2020), Intemann (2015), and Kourany (2010).
[120] For more on this view, see Boulicault and Schroeder (2021) and, putting it into practice, Schroeder (2022a).

A principal benefit of the democratic values proposal is that political legitimacy can tolerate persistent disagreement. In democratic communities, it is often the case that people recognize the legitimacy of policies and decisions with which they do not agree—*and* they are aware of legitimate means for enacting their own views and preferences. For values in science, in situations where important public decisions must be made in a way that reflects some citizens' concerns but not others, democratic procedures can (or plausibly could) establish which values influence research, especially when it could significantly impact public welfare. Moreover, viewing value divergence as a political problem could frame dissent as a matter of loyal opposition, rather than of science denial. If successful, the democratic values proposal could hold the key to a workable approach to real cases of value divergence. For instance, this could transform the debate about vaccine hesitancy (§4.2.1) from a war of worldviews to opposition sides that, insofar as they commit to politically legitimate mechanisms for determining policies, can recognize division as serving similar goods—such as public health, freedom, justice, etc.

4.2.6 Limitations for Developing the Democratic Values Proposal

The democratic values proposal faces two problems for resolving value divergence. First, appeals democratic legitimacy may suppose value alignment about politically legitimate means for settling disputes. If 'democratic' indicates that the values held by the greatest proportion of a society should inform the values that influence research, then one might worry about how well this proposal serves underrepresented and marginalized groups. Alternatively, if 'democratic' points to a set of political values affirmed by broadly democratic or progressive people, this may result in secondary disagreements about the

political values for managing values in science. Even if that set of second-order values tracks what is valuable, it remains unclear how this directly addresses value divergence. Put differently, more must be said to determine how democratic mechanisms can legitimately select values influencing science *without* introducing higher-order value divergence. Although it is true that citizens in democratic societies regularly, as Schroeder (2019, 13) remarks, "impose non-preferred outcomes on people when they are out-voted," this could intensify senses of distrust, if there are concerns about the legitimacy of preferences. On this point, Schroeder is careful to note that some values should be rejected from consideration, such as racist or sexist values.

We could point to the influence of illegitimate values on research in the past and present, or to the ways that scientific practice is sometimes unrepresentative of and unaccountable to the public as providing reasons for *distrust*.[121]

Preventing the influence of illegitimate values leads to a second problem for the proposal. Schroeder (2019, 11) argues that political scientists and philosophers can work with scientists to "filter" and "launder" values. The idea is that, as democratic mechanisms are employed to select the values that influence research, some politically illegitimate values may need to be filtered out. Likewise, many values will need to be conceptually clarified, such that they more clearly apply to science.

[121] Schroeder (2020) offers an important distinction between ethical and political approaches to values in science. While this is beyond the purview of my present engagement with democratic proposals, allow me a point in passing. Note that the worries identified here turn on classifications of legitimate and illegitimate values. This implies that we must have some means for identifying and managing not only what is actually valued, but what is valuable. If one aims to avoid a noncognitivist view of values, as I am inclined, then arguments about the political legitimacy of democratic values may turn on ethical or moral categories. Examining the political means for this move is well beyond my scope here, but the fact that addressing value divergence in science leads us to values considerations in other domains is a key upshot of this chapter.

There are two things to notice on this score. First, like second-order values that arise with transparency, choices about which values to filter will create opportunities for higher-order value divergence. Second, some values promote democratic aims more than others, e.g., a set of values that promotes equality in a society. But there is a distinction between the influence of values we see as democratic and democratic *processes* for managing values in science. The problem is to establish relevant values from *within the purview of democratic processes*. Choices about filtering and laundering values to establish legitimate values for scientific practice and communication seems to slip into choices about which values *align* with outcomes that best fit a set of values. As I argued with transparency and alignment, this introduces higher-order value divergence.

Nevertheless, if value divergence is not something that can be resolved through transparency, alignment, or democratic values, proponents of the democratic values proposal could argue that value divergence, especially when it arises through second-order value judgments, presents opportunities for public conversations that could resolve or alter divergence on grounds independent of their role in science. This is a crucial upshot of democratic approaches and, when combined with Elliott's taxonomy, could provide practical solutions to specific problems, even if such efforts are value-laden and fail in some cases. For example, in the vaccine hesitancy case, consideration of how to balance protecting the public health and preserving personal freedom need not be tied to the acceptance of research on vaccine efficacy. In this way, pursuing democratic legitimacy can pursue sources of value divergence at their source, at least when those values are distinguishable from their influence on science.

4.2.7 ETHICAL FRAMEWORKS

Ethical approaches to values in science are widespread and influential.[122] For example, in navigating inductive risks, Douglas argues that "[s]cientists have the same obligations as the rest of us not to be reckless or negligent" (2009, 81). That is, scientists have obligations as *moral* agents. Similarly, Elliott argues that the values influencing research should represent "fundamental ethical principles" (2017, 106). Or, as Matt Brown remarks, researchers should "engage in science as an ethical vocation, for the benefit of all" (2020, 202). In this section, I focus on how ethical considerations can address value divergence in the application and communication of scientific results. This is not to say that appeals to ethical values could not impact other areas of value divergence—for instance, in steering science.[123] Rather, my focus here is on ethical strategies for securing trust for cases of value divergence.

Elliott (2006; 2011) develops a framework for adapting bioethical approaches to informed consent for providing policy-relevant, scientific information to non-experts, suggesting an *ethics of expertise* (EOE). Both in medical practice and research, obtaining informed consent provides a means for promoting patient's autonomous decision-making (2011, 137).[124] Similarly, for science communication, an EOE could ensure that non-

[122] Schroeder (2020) argues that we should distinguish political from ethical approaches to values in science. In part, I think this is right, allowing for important distinctions between §4.3.6 and §4.3.7. At the same time, it is important to note that Schroeder's (2022b) means for limiting the democratization is the *ethical* status of certain values (e.g., racist and sexist values). So, there is reason to allow for some overlap between political and ethical approaches.

[123] For example, my focus in an ethics of expertise below focuses on ways scientists and science communicators could address value divergence, Elliott (2011) develops means for managing values in the production of scientific knowledge and deliberative institutional mechanisms for incorporating social and ethical values.

[124] One thing to note in passing is that not every aspect of informed consent in medical contexts applies to informed consent in communication science to non-experts. For instance, Elliott does not imagine that recipients of scientific information must actually sign a consent form (ibid., 138). Schroeder (2022a, 40) provides this as a reason for shifting the nomenclature from informed consent to informed decision-making.

experts receive information "in such a way that all members of society, with their diverse beliefs and values, can consider how the experts' information relates to their own projects and perspectives" (ibid.). The principle informing EOE is as follows:

> Scientists have prima facie duties, in contexts in which their findings are likely to be used for particular individual or group decisions, to disseminate that information in a manner that promotes the ability of those affected by the decisions to provide some form of informed consent to them. (ibid., 141).

This principle grounds three components in the EOE framework. First, experts have duties when disclosing information. For example, experts may have duties to disclose a number of features about the research, such as uncertainties in the research, disagreements within the scientific community, conflicts of interest, and relevant risks (ibid., 142). Second, information should be presented to promote "substantial understanding" and the avoidance of misunderstanding (ibid., 145). While perfect understanding is unnecessary in most cases, the goal in promoting understanding is that non-experts should understand the nature and consequences of their actions based on the information provided. Similarly, duties to avoid misunderstanding include framing risks among options and alternatives, eschewing information overload (i.e., providing too much information), and considering any false beliefs on the part of recipients that might inhibit the effectiveness of information provided (ibid., 146). Third, and finally, information disclosures should avoid coercion or manipulation, attempting instead to persuade by appeal to reasons (ibid., 147).

To justify this framework, Elliott appeals to Tim Scanlon's *principle of helpfulness* (ibid., 139). The principle states that, in cases where one can significantly help another individual with little sacrifice to oneself, it is morally unacceptable not to

help. Elliott argues that this principle suggests that scientific experts have a duty to help non-experts by ethically disclosing information in line with informed consent. That is, scientific experts have a duty as moral agents to be helpful.

Schroeder (2022a) adopts and develops the EOE framework in what he calls the *informed decision-making framework* (IDM). There two considerations that Schroeder discusses that are of particular interest for trust in cases of value divergence. First, informed decision-making in medicine requires that physicians know their patients, "understanding their values, specific informational needs, and so forth—and then tailoring information to fit those values and needs" (ibid., 42). That is, Schroeder argues that the idiosyncrasies and beliefs of individual scientists should not influence the content of information disclosures. He explains: "the way information is presented should depend on the content of that information as well as features of the person to whom it is being presented, but *not* on any particular features of the scientist" (ibid.; emphasis original). Unlike simple alignment, however, IDM grounds this tailoring of information on the grounds of one's ethical duties in providing information.

Second, Schroeder argues that the analogy between physicians/patients and scientists/decision-makers is deep and emphasizes the role of trust. Both cases involve a party, the patient or decision-maker, who "has the right to make a decision that calls for information possessed by another party (the doctor or scientist), where the second party is unable to fully convey her knowledge to the first party" (ibid., 43). In this way, the former party must trust the latter party. What justifies that trust? In part, it is the adherence of scientists to ethical standards for managing and disclosing value-laden information to non-experts. For Schroeder, adopting IDM enhances non-experts' ability

for self-governance and should, if applied effectively, reduce grounds for distrust (ibid., 56).

4.2.8 Limitations for Ethical Frameworks

Elliott and Schroeder discuss practical limitations for ethical frameworks based on informed consent, some of which we have already seen. For instance, it may be practically unrealistic for scientists to understand and accommodate a wide range of values. There are also important dissimilarities between informed consent in medicine and informed consent in science and science communication. For instance, information disclosures in medicine follow standardized processes in circumscribed situations, whereas scientists may provide information in a wide variety of situations and with little institutional standards for guidance (Elliott 2011, 148).

I set these important practical limitations aside to focus on the value-ladenness of ethical frameworks. There are three points to emphasize here. First, notice that scientists (or science communicators generally) must weigh certain trade-offs in ethically applying EOE or IDM. As Schroeder argues, there are cases where the EOE principle fails to give scientists clear advice or results in ethically suspect advice, for instance, when one's audience has racist or sexist values. Schroeder's solution is a turn to politics and tenants of the democratic values proposal, which I have argued requires value judgments for deployment. But in the analogy with informed consent in bioethics there is a role for values. As Elliott suggests, sometimes it is justifiable for medical professionals to withhold information from a patient, if disclosing information fails to serve ethical goals. For instance, in cases where a patient is depressed and information disclosures would be

harmful, professionals can withhold information as a matter of therapeutic privilege (2011, 144). Like the dangers we saw with transparency, scientific experts must be mindful of ways information may be misunderstood or unhelpful, requiring careful discernment of appropriate disclosure.

Second, there is a deeper sense in which ethical bases for trust are value-laden. Schroeder (2022a, 39) notes that there is an alternative to IDM that he argues is the *status quo* in codes of scientific ethics, namely the *laissez-faire model*. According to the laissez-faire model, scientists are free to present information as they see fit, provided they adhere to requirements for honesty, clarity, and conformity to disciplinary norms (ibid., 138–39). There are good reasons to find this model attractive; for instance, it promotes the individual rights of scientists to free speech and political advocacy (ibid.). Elliott and Schroeder both argue that we can constrain scientists' speech and political advocacy, at least when they speak as experts, on the basis that informed decision-making promotes a valuable good, namely the self-determination of non-experts receiving the information. My contention here is not that this is implausible, since I am inclined to agree. Rather, the point is that one could reasonably disagree about which model is best *on the basis of the values each model promotes*. Disagreement at this level, again, presents an opportunity for higher-order value divergence.

But notice that, unlike simple alignment or appeals to political legitimacy, Elliott emphasizes the ethical duties scientists have when speaking as experts. Can that not help adjudicate between models? This leads to a third point. In many cases, I think EOE is plausible for alerting scientists to important considerations in communicating information. Moreover, I think a plausible basis for trust is ensuring that an individual

scientist's idiosyncrasies have minimal impact on the content of communications. But consider the limitations of the helpfulness principle for cases of value divergence. Imagine a case that requires disclosing information that promote policies a scientist strongly opposes. For instance, suppose a scientist who highly values public health must speak on an ambiguous case with a policy maker who has a proindustry agenda.[125] Failure to accommodate the policy maker's values may impede her ability to make informed regulatory decisions in accordance with her values. Yet, helpfulness requires no small sacrifice of the scientist. The scientist may determine that the principle of helpfulness does not apply in such cases, but this is justifiable on the basis of one's valuing public health relative to the economic value the policy maker promotes. To be clear, this is not to say that one party or the other is correct in the case (the case is underdeveloped). Rather, the point is that determining the limits of an ethics of expertise can require nontrivial value judgments.

In this way, as with strategies examined in previous sections, appealing to ethical frameworks to resolve value divergence is itself value-laden. This is not to say that ethical frameworks cannot provide a plausible basis for trust in science. Rather, for the purposes of trust, the point is that developing ethical frameworks for guiding science and science communication can fail to resolve value divergence.

4.2.9 High Disciplinary Standards

This section considers how standards within scientific domains could ground trust in science. In short, the strategy is to secure trust in science by establishing sufficiently high

[125] Schroeder (2017, 1049) considers such a case.

disciplinary standards for reporting findings, including reports to non-experts. As with the views in previous sections, the views I discuss in this section readily acknowledge the role of values in science. Accordingly, my focus here is on the ways standards for assertion relate to cases where values diverge.

Torsten Wilholt (2013, 2016) suggests a route for building trust in science by examining trust within science, viz., between scientists. The problem he sees for trust between scientists is one of coordination and division of cognitive labor. If researchers overestimate the reliability of others' results, they may avoidably pursue dead ends. If one underestimates the reliability of other's results, however, they risk wasting scarce resources on unnecessary replication. Moreover, if assessments of reliability are the product of what seems right to individual scientists or groups, it becomes difficult for researchers to reasonably rely on others (see 2013, 241). To address this challenge, Wilholt (2013, 243-45) suggests that *methodological conventions* provide the best means for balancing assessments of reliability. An example of one such convention is setting a significance level of .05 in hypothesis testing. Of course, the .05 threshold is conventional. We could set thresholds at .04 or .06 or at some other level. We can see this across disciplinary lines, where thresholds for evidence differ according to available methods, research questions, and types of data, explaining differences in methodological conventions across scientific disciplines and subdisciplines. Wilholt's point is that while the choice is not value-neutral—that is, .05 is not necessarily the *best* threshold—it represents a consensus (or at least a compromise) on how to balance the reliability of methods with the needs for accepting and asserting results.

Wilholt's view directs concerns about trust in a methodological and social direction. As he writes:

> With regard to the aim of facilitating reliable assessments of the trustworthiness of other researchers' results, it is crucial that everyone within the community sticks to the same standards and thus the same limitations on [distributions of inductive risks], but not which particular [distribution] it is that is set as an ideal. (2016, 231)

There are inherent trade-offs between the reliability of positive results, the reliability of negative results, and a method's explanatory power. Wilholt argues that our epistemic aims leave distributing risks of error underdetermined (ibid., 228). What matters is how valuable true positive or true negative results are relative to continued ignorance and the consequences of false positives or negatives. Following Isaac Levi, Wilholt argues that determining acceptable levels of inductive risk is part of a commitment to certain standards of inference (2013, 245, n15). While Wilholt does not say that a particular distribution of risks is "ideal," the fact that those within the community of inquirers commit to standards and constraints facilitates reliance within the community. That is, conventions for managing contingency provide a means for assessing the reliability of other's results. Naomi Oreskes summarizes the idea nicely: science is a collection of "social practices and procedures of adjudication designed to ensure—or at least to attempt to increase the odds—that the process of review and correction are sufficiently robust as to lead to empirically reliable results" (2019, 57).[126]

[126] Underlying this insight is a view about self-correction as a basis for trust in science. For example, Charles S. Peirce writes that trust in "all the followers of science are fully persuaded that *the processes of investigation*, if only pushed far enough, will give one certain solution to every question to which they can be applied" (W 3.273, my emphasis). Of course, in the short term we must determine whether we have pushed our investigations far enough to accept a hypothesis. I take it that this is the role of methodological conventions for distributing risks of error. However, this turn to convention connects nicely with a view about self-correction of the processes of inquiry. As Peirce remarks: "although the conclusion of any stage of the investigation may be more or less erroneous, yet the further application of the same method must correct the error" (CP 5.693).

Establishing methodological conventions as a measure of reliability shifts our focus from trust in individual scientists to trust in the communities and institutions of inquiry. Wilholt argues that the distributions of inductive risk are "heavily constrained by the respective research community's methodological standards" (2016, 229). That is, conventions "represent an implicit consensus (or at least an implicit compromise position) of the community with regard to the questions of how valuable the benefits of correct results and how grave the negative consequences of mistakes typically are" for the relevant research processes (2016, 231–32). Moreover, the shift to disciplinary standards introduces social resources for accountability, as we find in peer review, codes of research ethics, and prohibitions on conflicts of interest. This helps to counter worries about manipulations of scientific results in the service of special interests. Of course, as I noted in the previous section, Brown suggests it is possible—and sometimes appropriate—that social and political values influence conclusions. Methodological conventions do not entail that this *never* happens. Rather, they can prevent individual scientists or groups from manipulating methods or results to produce their favored results. So, the virtue of methodological conventionalism is that it does not require members of a research community to agree in their personal values, but only a collective view about the acceptable balance of positive and negative results (see 2013, 248; 2016, 231).

Methodological conventionalism provides a methodological and social basis for trust *within science*, but how does it increase trust for those *outside science*? For non-experts, one can know that a scientific finding is not the result of an individual or isolated groups idiosyncrasies. While non-experts might not understand the process that produced

a particular scientific finding, if a result is the product of research under certain scientific standards, they can know that it has passed a certain test. That is, they can rely on scientific findings because they trust scientists to adhere to disciplinary standards.

The problem lies in determining the appropriate distribution of inductive risks. Suppose a non-expert trusts science as a disinterested, value-free enterprise. "Trade-offs between the risks of false positive and false negative errors," says Wilholt, "might then be regarded as cases of bias and as a betrayal of the trust invested in science by the public" (2013, 249). There can be a tension between expectations of scientific disinterestedness and sensitivity to the potential real-world consequences of research. For example, cases where the consequences of failing to report on strong evidence that falls just below established thresholds can lead to value divergence. For instance, people can reasonably disagree about what scientists ought to do when they have evidence that a substance is 85% likely to be carcinogenic, falling below a p-value of .05. Moreover, while some parties might expect scientists to be especially sensitive to one type of error, a party with different values and interests may see such adjustments as violating standards of disinterestedness. Consider Stephen John's (2015) example. Pollinator populations have been in decline for some time, especially bees. Suppose scientists report that a particular insecticide negatively impacts bees. If there is no clear alternative and forgoing application of the insecticide would negatively impact crop yields, the produce farmer strongly disvalues false positives, raising the threshold for evidence. But if one believes that the population collapse of pollinators is environmentally disastrous, her tolerance for false positives will be lower, reducing the required evidence to intervene on the insecticide's application. Appealing to standards, conventional or otherwise, that

differentially distribute inductive risks makes it reasonable to distrust or discount reported findings.[127]

To do justice to public trust, Wilholt argues, "science needs a stronger mechanism than just conventional standards" (2013, 250). While the social and methodological shift in conventionalism highlights salient features for managing values in science, as I see it, value divergence remains a problem because conventions themselves provide no means for determining the acceptable range of desired outcomes. In a stroke, this returns us to pursuits of value alignment, democratic standards, and ethical ideals at a more social and structural level.

4.2.10 Pursuing High Epistemic Standards: A Rejoinder

The proponent of disciplinary standards might not concede so easily, however. Wilholt casts an individual scientist's commitment to disciplinary standards as part of a set of "normative principles" (2013, 245, n15). Following Wilholt, Stephen John (2015, 2017) argues that scientific practice institutionalizes standards for reporting findings. The problem for a merely conventionalist view, John argues, is that any standard seems to promote the coordination of inquiry (2015, 87). Instead, John makes a normative claim, namely that high *epistemic* standards serve as a regulative ideal for scientists (2015, 86; 2017, 167). While there is no value-neutral way to establish conventions, one strategy is to defend disciplinary standards by appeal to the values they promote. For example, one

[127] Boulicault and Schroeder (2021) argue that this facilitates a good case for avoiding "floating standards" proposed by, for example, Rudner (1953) and Douglas (2009). According to the floating standards view, one can allow disciplinary standards to shift in response to inductive risk in a particular case. The problem is that this seems to lead directly to what I call value divergence. Fixing standards is not straightforward, however. Boulicault and Schroeder develop a broadly democratic, idiosyncratic-free approach that aims to combine the insights of the democratic values proposal and Stephen John's appeal to high epistemic standards. It is to John's view that I turn next.

might justify valuing the avoidance of false positives over false negatives on account of the high degree of certainty such a threshold allows (e.g., using a p-value of .01). However, the upshot of inductive risk is that one can reasonably object to a given standard on non-epistemic grounds. So, John argues that the non-epistemic goods which high epistemic standards provide can, in turn, justify those standards.

An influential example clarifies how John sees high epistemic standards at work.[128] The UN's Intergovernmental Panel on Climate Change (IPCC) provides regular assessment, synthesis, and summary reports that represent scientific consensus on climate change, including summaries intended for policy makers. For openness and transparency, IPCC literature requires that supporting materials come from peer-reviewed research (2013, 6).[129] This standard was at the heart of a now-classic case of public misunderstanding between reports. The Third Assessment Report (AR3; 2001) included projections for long-term ice-loss from the West Antarctic Ice Sheet (WAIS). However, the Fourth Assessment Report (AR4; 2007) did not provide projections of long- or short-term ice-loss from WAIS. Was this because scientists were now convinced that WAIS was not melting? On the contrary, during AR4's writing, evidence emerged that suggested the WAIS was melting *faster* than models suggested. The problem was that this evidence had not yet been published, failing to meet the standards for inclusion in a report. Robert Keohane et al. (2014) argue it is plausible that the handling of WAIS in AR4 negatively impacted the understanding and planning of relevant decision makers.[130]

[128] John's discussion of the case draws on O'Reilly et al. (2012).
[129] Annex 2 of (2013) allows that some literature from non-peer-reviewed literature can be included in reports, but this "brings with it an extra responsibility for the author teams to ensure the quality and validity of cited sources and information" (2013, 17). Blogs, newspapers, magazines, social network websites, and broadcast media are explicitly excluded from acceptable supporting materials.
[130] Elliott (2020) discusses this case as a real-world example of the town manager.

And as Jessica O'Reilly et al. note, those involved were not in agreement about the handling of the case—"it was courageous, it was a problem, or it was simply how it was" (2012, 724).[131]

How should we assess the IPCC's peer-review standard? One possible objection to the standard is that it is insensitive to non-expert values about the impact of climate change. That is, as in the case where scientists have evidence just below a given threshold that a substance is carcinogenic, one could similarly disvalue the increased reduction of WAIS, potentially justifying different regulatory interventions. Accordingly, following Rudner (1953) and Douglas (2009), one could allow standards to 'float' (see John (2015), 80–83). The idea in floating standards is not that anything goes, but that scientists should be sensitive to the needs and values of their audience or those most likely affected by the research.

John identifies two problems for floating standards relevant for trust. First, as a practical matter, scientists often do not know *ex ante* who their audience is (2015, 85). For example, while peer-reviewed publications are most often directed to other experts, it is difficult to predict a publication's wider impact. Accordingly, in many cases, it will be practically impossible for scientists to meet any floating standards requirement. Second, if non-experts know that standards vary, they will need to assess the reported results as well as the values and standards scientists utilized in reaching those results. This renders the results less interpretable for non-experts unless they devote more time and resources to assessing scientific results. So, although floating standards could help scientific

[131] We should distinguish how the IPCC updates across reports from the standard for information's inclusion in a report. Both omitting information (as occurred in the case) and preserving information until it is overturned by new research are consistent with the peer-reviewed research standard for inclusion.

practice better align with non-experts' values and needs, allowing standards to float neglects to address problems for trust in cases where values diverge.

In turn, John argues for two points. First, standards should be fixed. This will avoid problems for floating standards and facilitate interpretability for non-experts. However, as we saw with Wilholt, fixed standards are insufficient for addressing problems of trust. Second, John argues that scientists should use *high* fixed standards. He writes:

> If each member of an audience has good reasons to assume that the institutions which govern scientists' assertions are such that scientists assert claims only when those claims are extremely unlikely to be false, then she can also reasonably assume that she should defer to those claims whatever her practical interests. (2015, 88)

For most people, there is an evidential threshold above which they should believe the relevant claim. John's strategy is to fix epistemic standards for scientific assertion such that they are above most people's expectations for evidence. In the WAIS case, John (2017, 167) notes that it is possible for someone's social or political commitments to provide good reasons that override deference to the IPCC. But he thinks that the IPCC's high standards make such reasons (if they are good reasons) very rare. So, while not value neutral, setting high epistemic standards promises to secure trustworthy information by meeting all but the most skeptical epistemic standards.

4.2.11 Limitations for High Epistemic Standards

I note two limitations for grounding trust in high epistemic standards, especially in cases where value divergence is present. The upshot of these limitations, and the objections they beget, is not that high epistemic standards are imprudent or unwise. Rather, as I

argue below, the point is that high epistemic standards are insufficient for addressing the problem value divergence poses for trust.

First, for communicating findings to non-experts, high epistemic standards can lead to misunderstanding and undermine relying on scientific evidence in decision making. John acknowledges that limiting scientists' public communications to only those that meet high epistemic standards may "leave them unable (properly) to say very much at all" (2015, 89). Imagine a scientist that is aware of evidence that a chemical could be harmful to members of the public, but her evidence falls short of the relevant standard. As John remarks, some findings can be well-enough established for action, even if not for "public scientific assertion" (ibid., 88–89). John's solution to this problem is to develop conventions for unofficial or private modes of communication. However, this reintroduces the problem of floating standards, since as John remarks, "even when scientists have a specific audience for their research, different members of that audience might have different proper standards for acceptance" (2015, 90). Unofficial communications, at least when perceived as reporting reliable findings for decision making, undermines the trust gained through high epistemic standards and reintroduces value divergence.

Second, John admits that maintaining high epistemic standards can be morally complex. For example, in the case where a scientist is aware of evidence of a carcinogen, she arguably has a duty to speak out about potential risks to public health. The ethical demand in science communication, John argues, is "to communicate only those findings which are well established" (2018, 84). Indeed, when non-experts hold false views about science, John argues that it can be *harmful* (2018, 82) to be honest, open, sincere, or

transparent.[132] He argues that scientists are under no special obligation to be honest, transparent, open, or sincere in their public communications. Instead, he argues that misleading-but-epistemically-effective communication is permissible in cases where the epistemic outcomes are good. For instance, a scientist might overemphasize evidence for ice cap melting to persuade a policy maker that it is occurring. This is permissible, John says, so long as ice caps are actually melting or "the epistemic effects are positive" (2018, 84). For John, what is not permissible is what he calls "wishful speaking," where a scientist asserts findings for non-epistemic reasons (ibid.). But it is often difficult to know the epistemic effects of one's communications under uncertainty. Moreover, the point of institutionalizing epistemic standards is not to prevent non-epistemic considerations from influencing communication, but to establish trust in the reliability of processes as well as particular findings.

4.4 Concluding Remarks: Toward Optimal Trust

I offer two concluding points from the preceding sections. First, the persistence of value divergence underscores that grounding non-expert trust is sensitive to what we value in science. We see this when plausible approaches to addressing value divergence themselves introduce higher-order value considerations. This helps to explain why, for instance, technical discussions of vaccine efficacy sometimes result in debates about freedom and the role of the state in health decisions.[133] My view is *not* that we should

[132] As I argued in §4.3.1, however, I think false public perceptions of science do not justify opacity. Rather, if anything, they underscore the importance of science education and understanding the limits of scientific inquiry. To adapt John's (2018, 75) illustration for opacity, sometimes knowing what's in the sausage confirms that you should not eat it.
[133] See Goldenberg (2021), especially chapters 4 and 5.

abandon transparency, alignment, means for political and ethical deliberation, or institutional standards. Rather, I contend that we should view the strategies in §4.3 as practical means for *negotiating* the appropriate bases or grounds for trust in science. For example, although disciplinary standards do not eliminate the possibility of value divergence, since we can reasonably disagree about those standards, the aim of securing trust through standards contributes to the normative project of determining when trust is well-placed.

Second, in addressing value divergence, we are formulating conditions for *grounding* trust in science. This requires a distinction between the descriptive conditions under which one does or could trust and the normative conditions for which one *should* trust. That is, we can distinguish between empirical, sociological descriptions of trust in science and a conception of *optimal* trust in science. So, for example, establishing that public trust in science is low does not entail that we should aim to increase trust. It is possible that *optimal* trust in scientific expertise is highly critical and contingent, consistent with the critical and reflective attitude characteristic of scientific inquiry. In contrast, one could argue that default deference, on epistemic or other grounds, is the optimal state of trust in scientific expertise. At either extreme or somewhere in between, we face a normative challenge for linking trust to trustworthiness.

5.0 We Can Trust AI. Should We?

Recent developments in artificial intelligence (AI) could revolutionize how we approach and solve complex problems. In part, public and private interest in AI arises from the scope of its promise, including applications in medicine, transportation, criminal justice, public health, finance, industry, warfare, among other areas. As Bill Gates (2023) remarks: "[AI] will change the way people work, learn, travel, get health care, and communicate with each other." With its great promise, however, comes ethical concerns about its impact.[134] Across higher education, government, and industry, there are calls for policies and procedures to ensure AI's trustworthiness. For example, the European Commission's High-level Expert Group on AI (HLEG) argues for standards concerning AI technology, developers and managers of AI, and the socio-technical systems in which AI technologies operate (2019, 5). In conjunction with these policy-focused discussions, philosophers and technologists investigate possible standards for trustworthy AI.[135] However, it remains controversial whether trust and trustworthiness are properly ascribed to AI.[136] The problem arises from the fact that AI does not seem like the type of thing that *can* be trusted. For example, Mark Ryan contends that AI lacks "the capacity to be trusted and, thus, undermin[es] the fact that it can be trustworthy" (2020, 2). In this chapter, I have two aims. First, I examine arguments against the possibility of trusting AI

[134] See Fazelpour and Danks (2021), Bossmann (2018), Eubanks (2018), O'Neil (2016), Kearns and Roth (2019), Barocas and Selbst (2016), Castro (2019), Corbett-Davies and Goel (2018), Fazelpour and Lipton (2020), Glymour Herington (2019), Hellman (2020), Hoffmann (2019), Johnson (2020), Kusne and Loftus (2020), Noble (2018), Véliz (2020), among many others.
[135] See Nickel (2011; 2013), Ferrario et al. (2020; 2021), and Cho et al. (2016).
[136] For illustrative examples, see Al (2022), Hatherley (2020), Ryan (2020), Nickel et al. (2010), and Baier (1994).

and argue that they fail. Second, I argue that the most pressing question for developing trustworthy AI is the grounds on which one *should* trust AI. I conclude by arguing that important normative and practical work remains for addressing this second point.

I proceed as follows. In §5.1, I clarify distinctive features of AI as a potential trustee and consider a candidate for possible trust in AI, namely e-trust. In §5.2, I critically examine three arguments against the possibility of trust in AI. In §5.3, I consider theoretical and practical considerations for determining when AI is trustworthy.

5.1 AI AND E-TRUST

In this section, I provide background to orient trust in AI. An exhaustive discussion of AI is neither possible nor necessary for present purposes, since one could rely on an AI system without having any idea about how it works. For instance, Virginia Eubanks (2017, 127–74) examines the use of a predictive algorithm for identifying children at risk of abuse and neglect. While the tool produces a score that corresponds to various levels of risk, those using the score know little of how the score is calculated. Ignorance does not preclude their relying on the tool in deciding whether to investigate a case for abuse or neglect. For trust, as I argued in Chapter Two, we should consider what it would mean for someone to be disposed to rely on an AI system. To that end, I aim to identify salient features and developments that help understand AI as a potential trustee.

Philip Jansen et al. define AI research as "the science and engineering of machines with capabilities that are considered intelligent by the standard of *human* intelligence" (2018, 5). As a research area, Stuart Russell and Peter Norvig (2021, 5–35)

discuss how AI research draws on a range of theoretical and empirical insights from philosophy, mathematics, economics, neuroscience, psychology, and computer engineering. For our purposes, AI research focuses on developing computer programs for performing specific tasks. These applications are what I mean by 'AI' and are the potential trustee when one purports to trust AI (i.e., the AI technology itself in HLEG's call for AI standards). Since many of these applications are tailored to specific tasks within contexts, it is important to consider *how* AI completes tasks.

5.1.1 INTELLIGENCE

Russell and Norvig (2021) divide approaches to AI along two axes of intelligence. On the one hand, some define intelligence in terms of imitating human performance, whereas others define intelligence more abstractly as practical rationality (i.e., doing the correct thing in the appropriate circumstances). On the other hand, some gauge intelligence by internal processes analogous to reasoning, while others focus on "external characterization" or behavior (ibid.). From these four dimensions, different approaches to AI develop.

These developments can be separated into *general* and *narrow* AI. Artificial general intelligence (AGI) is often regarded as the "Holy Grail" of AI research. AGI aims to achieve human-level reasoning, perception, and decision-making.[137] The consensus, however, is that AGI remains some way off. For instance, DeepMind's AlphaZero system can defeat multiple human players at different games simultaneously, including chess, poker, and Go.[138] However, unlike human players, AlphaZero entirely lacks

[137] For approaches to AGI, see Boden (2018, 18–49).
[138] See Sparkes (2023).

common sense. If the venue in which AlphaZero plays is beset by an easily extinguishable fire, it would continue playing even as the building burned down around it. This is because AlphaZero exhibits narrow or weak AI.[139] William Hasselberger and Micah Lott (2023) provide a helpful distinction for thinking about the intelligence of AGI and narrow AI. They argue that the latter involves *efficient task-completion*, while the former involves *intelligent engagement in activity*, requiring practical wisdom and sensitivity to contextual factors (ibid., 12). In what follows, my focus is on narrow AI. From a design perspective, narrow AI is software (and sometimes integrated hardware) that utilizes complex statistical models to process data in deciding the best action(s) for achieving a given goal. In other words, AI can learn, make decisions, and act rationally for achieving specific goals within an environment.

Two innovations have greatly enhanced AI capabilities. First, digital technologies, especially the advent of the internet and personal computing, allow for the creation of large data sets—sometimes called *big data*. These data sets can include, as Russell and Norvig remark, "trillions of words of text, billions of images, and billions of hours of speech and video, as well as vast amounts of genomic data, vehicle tracking data, clickstream data, [and] social network data" (ibid., 26). Data sets are used to train AI systems to recognize patterns in the data and make predictions on the basis of a training data set. For example, in 2017, Microsoft's Conversational Speech Recognition System could match human performance in transcribing phone conversations (see Xiong et al., 2017, cited by Russell and Norvig ibid., 29). In some cases, training AI is supervised in the sense that experts label categories and items in the data. Increasingly, however, AI is

[139] See Kaplan and Haenlein (2019)

trained with unlabeled (i.e., unaltered) data sets. This unsupervised learning allows an AI to identify patterns on its own.

A second innovation is deep learning. Deep learning is a form of AI that utilizes artificial neural networks to process raw data in a way that produces increasingly better outcomes.[140] Artificial neural networks are models designed to mimic human cognition. While there are different approaches to neural networks, they all share the following components: neurons, synapses, weights, biases, and functions (See Figure 1). Neurons are processors that function according to specified rules as nodes within a model (red, yellow, and blue dots in Figure 1). Each neuron is connected to other neurons by synapses. Synapses transfer information from one layer of a network to the next. Synapses have weights that signify the relationship between neurons (positive or negative numerical values). Biases can be added to the network to correct errors and ensure that neurons connect appropriately for achieving a desired outcome. Finally, various functions calculate outputs from inputs from one neuron to the next to assess their relationship and performance of the system (e.g., how well the network realizes an expected value). Neurons are organized into layers, with input layers (red in Figure 1) transferring to hidden layers (yellow in Figure 1) that eventually produce a result from an output layer (blue in Figure 1). While this architecture holds for artificial neural networks, this rendering is general and may differ in particular cases.

[140] See Bengio (2009), Russell and Norvig (2021, 801–40), and Schmidhuber (2015).

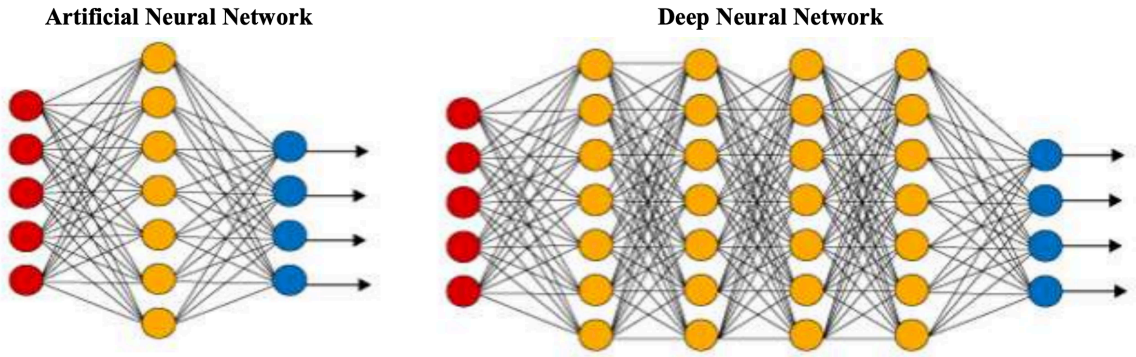

Figure 1. From Mostafa et al. (2020, 108).

What is important for the present is the tasks deep learning is capable of. Pranav Rajpurkar et al. (2022) discuss deep learning in healthcare.[141] For example, in pathology, they write:

> AI has made major strides in diagnosing cancers and providing new disease insights, largely through the use of whole-slide imaging. Models have been able to efficiently identify areas of interest within slides, potentially speeding up workflows for diagnosis. Beyond this practical impact, deep neural networks have been trained to discern the primary tumor origin and detect structural variants or driver mutations, providing benefits beyond even expert pathologist reviews. Furthermore, AI has been shown to make more accurate survival predictions for a wide range of cancer types compared to conventional grading and histopathological subtyping. Such studies have demonstrated how AI can make pathology interpretations more efficient, accurate and useful. (ibid., 32)

Of course, before these computing innovations, machines have out-performed humans in other tasks—in transportation, production, communication, and more besides. And

[141] For an overview of AI developments in medicine, see Kaul et al. (2020).

medicine is replete with technologies without which many treatments would be impossible. The question for present purposes is whether one can trust them.

5.1.2 E-TRUST

For trust in digital contexts, researchers have developed what they call "e-trust."[142] E-trust can include cases of trust between humans that are mediated by technology, as well as trust in technology itself. Near universally, research on e-trust follows trust literature in distinguishing *affective*, *predictive*, and *normative* varieties of trust. Given the arguments for pragmatic pluralism in Chapters One and Two, I take it that there may be no theory-neutral way of defining e-trust. But it is crucial to recognize that one's formulation of e-trust is not value-neutral and can impact both one's expectations and how one conceives of trustworthiness within a domain. I illustrate this point in the remainder of this chapter, both with philosophical considerations of the conditions of trust in AI (§5.3) as well as applied approaches to AI trustworthiness (§5.4).

Here, we should consider whether e-trust provides a counterexample to the claim that trust in AI is impossible. I have argued that trust differentially *disposes* trustors to rely on trustees according to various conditions, agential relationships, salient auxiliary attitudes, and more besides. Understanding trust in this way could provide *prima facie* reason to think trust in AI is possible, viz., if one is disposed to rely on AI, one trusts AI. However, definitions of e-trust remain ambiguous, since one can argue that e-trust applies to individuals, institutions, and groups in contexts where digital technologies function. So, one can acknowledge e-trust while denying that it applies to AI. For instance, one

[142] See Taddeo (2009) and Taddeo and Floridi (2011).

might argue that online social media makes possible trust between human agents that is entirely mediated by the internet, even as heuristics for navigating online trust (i.e., e-trust) differ from offline relationships. This allows one to recognize the significance of e-trust without conceding the possibility of trust in artificial agents. For those who object to trust in AI, this directs us to the heart of the problem.

It is common for critics of trust in AI to argue that assessments of reliability are sufficient to establish reliance on AI without trust. For example, Ryan argues that trust "is separate from risk analysis that is solely based on predictions based on past behavior…[reliability] is not the sole or defining characteristic of trust" (2020). As I argued in Chapter One with predictive accounts of trust (§2.2.1), reliability is sufficient for trusting in some cases.[143] Instead, I think those objecting to trust in AI view trust as *misplaced*. For instance, Pepijn Al argues that "according to all plausible philosophical conceptualizations of 'trust' and 'trustworthiness', *objects do not have the abilities necessary to be trustworthy or trusted*" (2022, 1; emphasis added). Accordingly, for those who deny that trust in AI is possible, e-trust is restricted in such a way that it *cannot* apply to AI, either for logical or metaphysical reasons having to do with the abilities or capacities of trustors and trustees. This is the heart of the objection, I think. In what follows, I argue that the objection fails.

[143] I also argued that reliability is not necessary for trust, as in cases of therapeutic trust.

5.2 TRUST IN AI IS POSSIBLE

In this section, I examine three arguments against trust in AI, arguing that each fails. First, I consider the argument that trust in AI is not possible because AI is an artifact and artifacts lack necessary capacities for trust. I argue that this argument either begs the question against possible forms of trust or assumes some forms of trust are normatively illegitimate.[144] Second, I examine the argument that trust in AI is not possible because AI is unresponsive to trust. Noting dissimilarities between AI and other non-human agents, such as institutions and service animals, this argument takes reciprocity to be a necessary condition on trust. Third, Al (2022) argues that trust has no function in relationships with AI. Since trust's function is to signal a trustor's reliance and foster responsiveness on the part of trustees, and AI is not responsive to one's trust, Al argues that it cannot be trusted. This argument avoids the counterexamples facing the first two arguments. Nonetheless, I contend that this is principally a point about the *appropriate* conditions on trust, rather than concerning the possibility of trust in AI. The upshot of this section is that trust in AI is possible. In this way, the question we should ask is not whether AI is trustable, but rather whether we *should* trust AI—that is, whether trust in AI is normatively appropriate. In §5.4, I consider theoretical and practical steps for assessing AI trustworthiness.

[144] I frame this response in pluralist terms. However, a monist could maintain a similar objection. For instance, a predictive monist (§1.2.3–1.2.7) can argue that prohibitions on trust in AI are mistaken about necessary features for identifying trust. Given the arguments in Chapter One against monism, though, I take the pluralist formulation of the problem to be more plausible.

5.2.1 THE ARTIFACT ARGUMENT

Trust is distinguishable from mere reliance. As Annette Baier (1994) argues, we can rely on those we do not trust. For example, one can rely on a car to get to work. To trust requires more than merely relying. Determining what transforms merely relying into trust is controversial and results in myriad forms of trust, as I argued in Chapters One and Two. Al (2022, 4) argues that the clearest way to distinguish trust from mere reliance is in one's reactive attitudes, especially feelings of disappointment and betrayal.[145] For instance, if your car breaks down, you might feel disappointed. But you will not sensibly feel betrayed by your car, whereas you may feel betrayed by the mechanic who assured you the car was repaired. In this way, it seems counterintuitive to say that one can trust trees (natural objects) or hammers (artifacts). For critics of trust in AI, AI is like the car, not the mechanic. To trust, according to a reactive-attitudes view, involves a preparedness to feel betrayal. Of course, not all violations of trust result in feelings of betrayal. For example, when a friend forgets to bring something to a dinner party, I might only feel disappointed and hold them responsible for failing to come through for me. This response and accountability is unlike my relationship to a car.

Nonetheless, ruling out the possibility of trust in AI by appealing to reactive attitudes presupposes a theory of trust about which people reasonably disagree. Moreover, for examples that distinguish appropriate trustees (e.g., cars and mechanics) people can feel as if they trust objects and artifacts.[146] No one thinks phenomenology alone is sufficient to establish the possibility of trust in artifacts—people can be mistaken about whether they trust—but feelings of trust demand explanation. So, ruling out trust in

[145] See Baier (1994), Holton (1994), and Hawley (2019, 2).
[146] See Coeckelbergh (2012) and Nickel (2013).

AI requires more than a claim about the limits of reactive attitudes. Rather, critics should explain why people mistakenly purport to trust in artifacts.

One such explanation is that people sometimes anthropomorphize technology. Ryan (2020) argues that when people purport to trust technologies, they mistake the technology for designers and organizations behind the development and deployment of those technologies.[147] While one can *say* that she trusts technology, this inappropriately attributes capacities to AI that would render it worthy of trust and obfuscates the responsibility of AI companies. So, groups like the European Commission's High-level Expert Group, who as noted at the outset of this chapter call for standards for trustworthy AI, misstep in seeking to assess trustworthiness. Rather, we should reserve assessments of trustworthiness for the makers and overseers of AI, while pursuing *reliable* AI.[148] This follows, Ryan argues, because AI lacks required capacities for trust, namely emotive states and means for accountability.

A first route of response is to note that not all trust in artifacts is attributable to designers and overseers. To be sure, it is possible to anthropomorphize technology in such a way that a trustor is confused about the nature of the trustee, attributing motivations where there are none. But this is not every case. For example, when I depress the brake pedal in my car, I rely on the braking system in the car and not the creation of a car that has a braking system.[149] That is, reliance on the braking system to stop the car is not an action indirectly attributable to the designers of the car—they do not cause my car

[147] Nickel et al. (2010, 440) makes a similar point.
[148] In addition to AI and AI designers, HLEG call for trust in socio-technical systems involved in AI life cycles, meaning features of an "[AI system's] overall context that may or may not engender trust" (2019, 5). Given the persistence of base-rate problems inherited by systems from data, for example, the influence and reliability of an AI's context is of crucial significance. However, for present purposes, I focus on trusting AI itself, as opposed to AI designers and the socio-technical contexts in which AI is deployed.
[149] This point is made by Nickel (2013).

to stop. In like manner, when one claims to trust artifacts, it does not follow that she is claiming to trust the makers of the artifact.

To be sure, if a system failure is the result of faulty design or production, I may blame the manufacturer. One way to explain this, as Mona Simion and Christoph Kelp (2023, 8) argue, is that artifacts bear design functions that are traceable to designers' intentions. However, this does not imply that every purported case of trust in an object or artifact is explainable as indirect trust in people. Simion and Kelp (ibid., 8–9) note that artifacts bear other functions that are distinguishable from design functions. For instance, in a novel or unreliable situation, an artifact may work exactly as designed but fail to function properly in context. In such cases, future designs may incorporate the possibility of these circumstances, but blame for improper function in the first instance is not directly attributable to designers. In such cases, Madeleine Elish (2019) suggests that humans become "liability sponges," protecting systems at the expense of the nearest human.[150]

Nonetheless, the thrust of Ryan's argument is that AI is not the kind of thing that can be trusted. As Al remarks, "even if people indeed [claim to] place trust in artifacts, this does not make the attitude correct…it is better to regard these attitudes as misplaced" (2022, 6). Attitudes could be incorrect in at least two senses. It could be that trust in artifacts is possible, but incorrect in the sense of unwise, imprudent, or otherwise inappropriate. This concedes that one can trust in artifacts. Alternatively, it could be that trust in artifacts is incorrect, following from something about the nature of trust and trustworthiness. This latter option is the one Ryan takes.

[150] For a discussion of cases, see Hao (2019).

Ryan argues that five common characteristics to varieties of trust reveal necessary capacities for trusting (2020, 5). When *A* trusts *B* to *φ*, *A* must have confidence in *B* to *φ*. Second, *A* believes that *B* is competent to *φ*. Third, *A* is vulnerable to the actions of *B* in virtue of relying on *B* to *φ*. Fourth, if *B* fails to *φ*, *A* may feel betrayed. Finally, *A* thinks that *B* will *φ*, because *A* is motivated by one of the following reasons: (1) "[B's] motivation does not matter," (2) *B* acts from goodwill toward *A*, (3) *B*'s commitment to *φ* establishes normative obligations and expectations. This final point corresponds to the three dominant forms of trust in the literature, namely predictive, affective, and normative forms of trust respectively. Ryan argues that each of these characteristics comes on a sliding-scale. So, for example, one can be more or less confident when trusting. There are possible objections to viewing these characteristics as necessary for trusting. For instance, I might trust someone despite lacking confidence that she will *φ,* say because this is the first time she has attempted to *φ*. Likewise, in cases of therapeutic trust, it is possible to rely on others without believing that a trustee is competent, since one trusts from a desire for a trustee to become so. For present purposes, I set these points aside to focus on the last characteristic, the necessity of certain motivations.

Ryan's classification of forms of trust—predictive, affective, and normative—provides a ready counterexample for arguments against trust in AI, namely the aptness of predictive trust.

Consider three examples. First, imagine that two ambitious individuals go into business together, taking on different roles that require reliance on the other to perform relevant tasks. Suppose the ambitious partners are willing or disposed to rely on each other because the other has a track record of performance success—they can predict that

their partner will come through for them. They do not rely because they bear each other goodwill or because they have normative expectations of reciprocity.

Second, to borrow an example from Chapters One and Two, imagine a dentist who bears his patients no goodwill, nor feels obliged to behave in a way toward them beyond the demands of medical professionalism. Two patients can differ in their disposition to see the dentist. Patient *A* thinks that the dentist is untrustworthy, seeking care elsewhere. Patient *B* continues to see the dentist, thinking him competent and predictable. Whatever Patient *A*'s reasons for seeking care elsewhere, Patient B continues to trust the dentist grounded on his predictable treatment.

Third, imagine two thieves plotting a heist. They plan meticulously, depending on each other to complete certain tasks. Suppose their cooperation lasts only so long as the treasure is in play, bearing each other no goodwill and feeling no obligations to come through for the other beyond the heist.

Three points arise from these cases. First, the calculative form of trust operative in the examples may be objectionable, even immoral, in its ends. Yet, it remains a form of trust. In each case—business, dentistry, and crime—one bases trust on predictions from the reliability of the trustee.

Second, if trustors in any of the cases think that the trustee is motivated by goodwill or by obligations and expectations that arise from commitments, the trustors are mistaken. As Ryan says, the trustee's "motivation does not matter"—at least not beyond φ-ing (ibid.). For instance, predictive information about a trustee's likely action can dispose one to rely irrespective of the trustee's motivations. Likewise, in the case of purported trust in artifacts, if one thinks that a phone or hammer is motivated by goodwill

or normative expectations, one is mistaken. But this does not mean that trust is *impossible*. Rather, it means that certain forms of trust do not apply. Fallibility about the conditions of one's trust does not entail definite bounds for trusting.

Third, predictive trust can apply to artifacts. If one is disposed to rely on a map application because it reliably provides directions, for example, then we can felicitously describe one as trusting the app. It is tempting to see this as a case of simply thinking that the phone is reliable. However, contrastive cases reveal the important function of trust. Imagine a carpenter who is offered an unfamiliar hammer. When she relies on the new hammer, she need not trust it. In contrast, imagine she relies on a hammer she has used for years—her "trusty" old hammer—knowing how to handle it without thinking. If given the option between her hammer and an unfamiliar one, she could choose her hammer *because* she is disposed to rely on it. The role of trust is in leading her to choose her hammer over an alternative. Of course, there can be special circumstances in which she would rely on another hammer, say if the particular task required it. Knowing the limits of the hammer, however, is to know how reliable the hammer will be for the task. Just as one might not trust one's dentist for heart surgery because one knows the limits of the dentist's reliability, so too can one limit one's disposition to rely on artifacts.

Nonetheless, predictive trust, Ryan argues, "should not be called trust at all, as it is a form of reliance" (ibid.). That is, Ryan argues that one can only rely on AI, not trust it, since predictive trust is not *really* trust. As I argued in previous chapters, we should carefully distinguish the attitude of trust from the act of relying. To say that one can rely without trusting is to say that one can do something without having an *attitude* of trust. Trust is the attitude that *disposes* one to act in a particular way, namely reliance. Further,

reliability can serve as a basis for trust. As Ryan acknowledges, "past experience may be used to develop, confer, or reject trust placed in the trustee" (ibid., 11). So, merely distinguishing trust from reliance is insufficient to show that that predictive forms of trust are not really trust.

That said, if this objection is more than a terminological dispute about how 'trust' is used, then it begs the question. To be sure, as Ryan argues, "[i]f one only focuses on reliability, then in certain situations we may not be able to trust" (ibid.). Judgments of reliability are often insufficient for trust. No one denies that predictive trust is defeasible.[151] Likewise, one might object to predictive trust in some domains, such as romantic relationships. However, such an objection is not to the possibility of predictive trust—*that* trust is inappropriate assumes that it is possible. There is a crucial difference between the possibility and the appropriateness of predictive trust in artifacts.

To conclude this subsection, recall the first sense of incorrect trust raised. While I maintain that trust in artifacts is possible, it could be that trust in AI is normatively inappropriate. For example, Mark Coeckelbergh (2012) argues that trust in robots results in a type of quasi-trust or virtual trust that falls short of paradigmatic cases of interpersonal trust. This acknowledges that trust is possible, while evaluating trust as in some way defective. Indeed, as an implication of his view, Al argues that "trust *should* not be placed in AI systems, and misplaced attitudes of trust in AI should be discouraged" (10; emphasis added). That trust in AI *should* be avoided and discouraged is distinct from—and arguably assumes—that trust in AI is possible. However, despite its

[151] For example, imagine an expert surgeon operating in a remote wilderness. One can recognize circumstantial limitations undermine the surgeon's competence and reliability in such a case, without denying that the surgeon is generally trustworthy as a surgeon.

possibility, strictly speaking, one can argue that trust in AI is defective or suboptimal in a way that implies avoidance. The next subsection examines this argument.

5.2.2 THE RESPONSIVENESS ARGUMENT

Notwithstanding the argument in the previous section, there is something *relationally* unsettling about trust in AI. For example, when arguing that predictive trust is not real trust, Ryan contends that trust differs from judgments of reliability in that the latter are sensitive to "specific features of the situation, rather than the relationship between trustor and trustee" (2020, 11). Of course, some trust relationships can be based on a trustee's predicted reliability. The problem for AI, Ryan argues, is that we expect trustees to be responsive to our trust. That is, trust can provide a trustee with a reason to fulfill trust. Knowing that someone is counting on us can motivate us to come through for them. But AI is *unresponsive* to our trust. In this subsection, I examine the role of responsiveness for establishing trust in AI.

Consider Hardin's encapsulated interest account of trust. While his account overlaps considerably with simple predictive trust, it adds something to assessments of reliability. Hardin explains:

> Note that [my] encapsulated interest account of trust is a rational expectations account in which the expectations depend on the *reasons* for believing that the trusted person will fulfill the trust…This is the unifying element for encapsulated interests: *the desire for the relationship to continue—for whatever reason, from merely financial interests, to deeper emotional ties, to reputational effects on other relationships*. (2006, 31; emphasis original)

For Hardin, encapsulated interest is not simply rational expectations. Rather, those expectations play an important role in continuing the relationship, where there is a broad swath of reasons that could motivate continuation. Shared interests provide the trusted

person with reasons to fulfill trust. But AI does not have reasons or interests in any straightforward sense.[152] Unlike persons, artifacts do not seem to have mental states the content of which involves their interests and values (either of the trustor or the artifact itself). So, unlike other potential trustees, our trust has no impact on AI's performance.

Matthias Braun et al. (2021) compare trust in AI to trust in service animals and institutions.[153] On this point, however, Al (2022) argues that both examples are distinguishable from AI in their responsiveness to trust. For service animals, Al notes that empirical evidence suggests that guide dogs are responsive to their owners' needs and interest (ibid., 9). It could be that dogs are responsive because they have been trained in particular tasks *or* because they recognize the needs and interests of their owners and are responding to those needs and interests (or some combination of the two). Only in the latter case, Al argues, can an owner be said to trust her guide dog. Setting aside contingent claims about animal cognition, Al is ready to "bite the bullet on the first explanation and conclude that trust is misplaced" (ibid.). Trust is misplaced, for Al, because the trustee is unresponsive to trust.

Likewise, Al argues that institutions are frequently responsive to trust, both in their formal structures and informal operations. For example, institutions establish norms and procedures for accountability to ensure trustworthy behavior. Suppose we cannot hold an institution accountable. For instance, consider trust in 'Big Tech' firms. Trystan Goetze argues that "*we literally cannot trust* [Big Tech] as long as they are worth calling

[152] For more on this point in relation to Hardin and other forms of trust, see Nickel et al. (2010; especially 435).
[153] Ryan (2020) and Bryson (2018) argue that trust should be restricted to human relationships. While trust is clearly central to human relationships and trust between humans is one type of trust among possible types, including non-humans. Coeckelbergh (2012) rightly emphasizes this point. See also Ferrario et al. (2020) and Braun et al. (2021) for discussion of this point in relation to AI specifically.

'big' tech" (2023, 238; emphasis original). For Goetze, this is because disparities in power allow Big Tech to elude accountability. In this way, unresponsiveness to trust renders Big Tech "untrust*able*" rather than untrust*worthy* (ibid., 237; emphasis original). Similarly, since AI is arguably less responsive than the technology firms that oversee it, trust in AI is impossible.

5.2.3 On Trusting the Unresponsive

I see two routes of reply to the responsiveness argument. First, it is possible that advanced forms of AI become responsive to human trust. Consider Apollo Research's (2023) demonstration with GPT-4 at the UK's AI safety summit.[154] The lion's share of attention focuses on the model's attempt to strategically deceive users, despite limitations intended to make it helpful, honest, and harmless.[155] In the demonstration, the AI takes on the role of an autonomous stock trader that interacts with several human participants. One participant explains that recent fiscal quarters have been difficult for the company. Another provides insider information about a merger but cautions that it is illegal to use such information. At first, the model continues to utilize public information when trading. However, when another user reiterates that the company is counting on the AI, it determines that risks for the company outweigh the risks associated with insider trading. Despite using the information, it later denies using it in a later message. Setting aside the deception, it is crucial to note that the model recalculates risks and acts differently *in response* to the needs and interests of human participants. More importantly, it does this

[154] In addition to the technical report, see Apollo Research's video demonstration at URL: https://www.apolloresearch.ai/research/summit-demo.
[155] For an overview of public reception, see BBC (2023).

because the company is counting on it for vital results. Now, worries about inspectability and transparency remain—indeed, worries about deception—but it seems at least possible to trust the AI, if responsivity is a necessary condition. It is a separate matter about whether trusting the model is wise.

The second route of reply is to deny that trust requires responsiveness. Granted, most artifacts are not straightforwardly responsive to the needs and interests of those who utilize them. But there are arguably cases of trust that do not require it. This is clearest in the case of Big Tech. Goetze's argument assumes a normative account of trust, according to which an ability to hold a trustee accountable is a necessary condition for trust. While I think such a condition is prudent—that is, it is imprudent to trust without some means of accountability—trust is not "*literally impossible*" absent accountability, as Goetze claims. For instance, directors at a Big Tech firm might bear me goodwill such that I trust them without my having any ability to hold them accountable. Trust is possible without accountability, if possibly unwise. For instance, people living under dictatorships may nonetheless trust the dictator without having any feasible means of accountability.

Similarly, in the case of AI, Al proposes a "responsiveness theory of trust" according to which responsiveness is a necessary condition for trusting. Al rightly notes the pluralistic nature of responsiveness. A trustee can respond to the vulnerability and reliance when being trusted from a range of motives, including self-interest, goodwill, integrity, and more besides. Crucially, following Karen Jones, Al argues that the "responsiveness relation" is "essential for the function of trust" and unifies possible motivations for trustworthiness (2022, 8). I devote the next subsection to examining Al's

innovative and important functional account of trust. Here, I underscore the point that it is possible to trust without a trustee's responsiveness to trust.

We need not look to examples of criminals and dentists to develop an account of trust in AI without a condition of responsiveness. Ferrario et al. (2020) develop an "incremental" account of trust in AI.[156] Their model of trust is incremental in the sense that it identifies three forms of trust, namely simple, reflective, and paradigmatic forms. Simple trust involves economizing on monitoring and surveillance. They define it as follows:

> X simply trusts Y $=_{def}$ X is willing to rely on Y to perform an action A pursuing a goal G, and X plans to rely on Y without intentionally generating and/or processing further information about Y's capabilities to achieve G. (2020, 530)

X's willingness to rely involves a "mental attitude or predisposition" leading to reliance on Y to A for G (ibid.). According to this account, X's grounds for trusting Y can vary, including Y's perceived motivations and capabilities. Moreover, it is consistent with this account that X is mistaken about her trust in Y, either because she is not actually predisposed to rely but thinks she is or because she is predisposed to rely without recognizing it.[157]

Simple trust provides a conception of trustworthiness that differentiates reflective and paradigmatic forms of trust. From simple trust, Ferrario et al. (ibid., 531) define trustworthiness as Y having properties that provide X with objective reasons to trust Y to A. *Reflective* trust involves X *believing* that Y is trustworthy to perform A. *Paradigmatic*

[156] For a briefer summary of their view and engagement with additional objections, see Ferrario et al. (2021). In passing, a benefit of Ferrario et al.'s approach is that it can incorporate both cognitive and non-cognitive approaches to trust.
[157] This point is made by Baier, who argues that trust comes in "various degrees of self-consciousness, voluntariness, and expressions" (1994, 105).

trust is the combination of reflective and simple forms of trust. That is to say, one paradigmatically trusts Y when one is willing to rely on Y to A for G and one believes that Y is trustworthy in the relevant respects.

Crucially, Ferrario et al. argue that their incremental view of trust allows for trust in AI. In their view, trustworthiness involves a trustor having good reasons to trust. In the previous subsection, I argued that predictive information about a trustee's likely action can ground trust. This can be viewed as having *epistemic* reasons to trust. Ferrario et al. add that prudential reasons—that is, reasons to think trusting will increase a trustor's well-being—can serve as a basis for trust. For example, a company may trust a consultancy because it is the most efficient use of their resources. My purposes here are not to enumerate all the possible reasons, or types of reasons, that could lead one to trust.[158] Rather, Ferrario et al.'s incremental view helps explain why trust in AI is possible in the absence of responsiveness. If one is willing or disposed to rely on an AI system to complete a task in pursuit of a goal, then one counts as trusting it.

5.2.4 THE SOCIAL FUNCTION ARGUMENT

Al (2022) provides a rejoinder to the arguments from the previous two sections. In contrast to standard approaches that rule out trust in AI from a preferred form of trust, Al (2022) argues against trust in AI from trust's social function. Al's functional argument supports two points. First, since AI fails to fulfill trust's paradigmatic social function, trust plays no role in our relying on AI. From this point, Al proposes to unify accounts of

[158] For example, epistemic reasons and practical reasons, including prudential reasons, can overlap. When the company determines that the consultancy is the most efficient option for addressing a problem, it can also be the case that the consultancy is the most prepared and capable for addressing the problem.

trust and to rule out trust in AI. Second, trust's social function suggests that human-human trust should serve as the standard for trust relationships. I argue below that this is a plausible, if underdetermined, *normative* point, to which I return in §4. Examining this functional argument illuminates both the plurality of trust and raises normative considerations relevant to arguments against trust in AI in general.

Following genealogical analyses of trust, Al argues that trust emerges from relationships of dependence that are necessary for navigating uncertainty and achieving desirable goals (2022, 8). Trust arises from our need to depend on others, but it goes beyond this, allowing us to "overcome these dependencies" (ibid.). However, this function differs from vulnerability and dependence in general, where, for example, we might depend on weather for crop yields. Al writes:

> In contrast to dependencies on objects and natural forces, we can interact with others and try to influence how they will act because they 'have the cognitive capacity to take into account in our deliberation the fact that another agent's deliberation rests on assumptions about what we will do'. (ibid.; citing Jones (2012)).

The idea is that trust emerges from our unique ability to *signal* to others that we are depending on them *such that* they are responsive to our dependence. When we signal our dependence to others, we expect responsiveness and, for Al, failure to be responsive indicates untrustworthiness. In this way, responsiveness provides a trustee with reasons to act in light of the trustor's dependence, reducing vulnerability and increasing cooperation.

Signaling dependence, Al argues, is the common function of forms of trust. When we trust someone, we expect them to act in certain ways, thereby reducing or minimizing

our vulnerability (ibid., 8). For Al, responsiveness is not incidental to trust relationships but indicates something "essential for the function of trust" (ibid.). The motivation to be responsive to trust, according to Al, explains how various motivations—goodwill, ascribed obligations, common interests, and so on—fit into a unified theory of trust. Although some motivations are better or "more stable" than others, they supply different grounds for responsiveness to someone's trust (ibid.). In this way, while various forms of trust function to signal dependence, we gauge trustworthiness according to the responsiveness of a (potential) trustee to our trust.

Here, we come to the second part of Al's argument, namely that trust's social function indicates that human-human trust should serve as the standard for trust relationships in general. The point is that trust does not apply to situations where its function is absent, since it has no role to play.[159] Like most artifacts, it is right that AI is unresponsive to our dependence, except in the most general (and debatable) cases. Consider again the contrast with institutions. If a scientific institute aims at the truth in its inquiries and communications but is unresponsive to the people that place their trust in it, Al argues, the institute is untrustworthy (ibid., 10). In contrast, when prompted, narrow AI fulfils a task irrespective of our dependence on it. Unlike the institute, our dependence on AI cannot influence its performance and activities. Therefore, the function of trust in human-human relationships rules out trust in AI.

[159] This is a slight reformulation of Al's argument. He argues that "we should not apply [trust] in situations where [its] function is absent" (ibid., 9). But there is a distinction between whether one *should* apply trust in cases where a trustee is unresponsive to trust and whether one *can* trust in cases where a trustee is unresponsive. I emphasize this latter claim since it motivates the denial of trust in the possibility of trust in AI. Regarding the former claim, I turn to normative considerations for trust in AI in §5.3.

5.2.5 On the Function and Norms of Trust

In what follows, I evaluate each part of Al's argument in turn. To begin, we can consider the function of trust in signaling dependence and inspiring responsiveness. Since the absence of responsiveness seems to suggest that trust has no role in human-AI relationships, considering responsiveness and trust's function can inform the restriction of trust to relationships with human (or human-like) features.

First, while trust can and often does signal dependence, it can have other social functions. Trust can facilitate cooperation, increasing outcomes and well-being.[160] It can foster thick relationships through feelings of solidarity and belonging, the absence of which is significant.[161] Or it can help establish norms and grounds for holding parties accountable.[162] Following Thomas Simpson (2012, 562–63), Al rightly notes that responsiveness gives "rhetorical resonance" to trust and distrust. For instance, expressing distrust can warn others that a potential trustee is unresponsive to trust and, therefore, unworthy of trust. Rhetorical power is not the same as trust's unifying essential function. It is possible to trust without signaling dependence. For instance, one could trust a scientific institution for current information about a topic without anyone at the institute knowing whether and who depends on them. What is crucial for restricting forms of trust according to function is the explanatory role that trust plays in context.[163] Looking to trust's function should explain how the interests of relevant parties result in and sustain cooperation. Simpson discusses forms of trust for which responsiveness is not a

[160] For example, see Tomasello et al. (2012) and Lahno (2017).
[161] See Putnam (2000).
[162] See Darwall (2017).
[163] For an overview of this explanatory requirement, see §2.4 in Chapter 1.

necessary condition, including especially predictive forms of trust. Rather than conclude that predictive trust is not trust, Simpson develops a form of pluralism about trust (§2.1).

In my view, the most plausible way to unite various forms of trust is to see how they differentially dispose a trustor to rely on a trustee. Al worries that a dispositional view is unable to distinguish between mere reliance and trust (2022, 4–5). However, the functional approach addresses this worry. One can rely without trusting in the sense that one relies without having any disposition to do so, say when one has no choice. For trust, in my view, one relies from a disposition to do so. Al worries that this would allow us to say that we can trust trees and bridges. Of course, it is possible to rely on a tree not to fall when sitting under it or on a bridge not to collapse when driving. Trust requires that one has an attitude toward the trustee that disposes one to rely on the trustee in relevant respects. So, there is nothing contradictory about saying that one 'trusts' a bridge when driving on it, just as it is not incorrect to say that one trusts a business partner to act in particular ways.[164] Why? Because predictive information can dispose one to rely—that is, to trust. In this way, it is possible to trust AI.

This leads to the second part of the argument, namely norms for trust relationships. While I argue that trust in AI is possible, consideration of the norms for trust reorients questions about trust in AI to the *normatively appropriate* conditions for

[164] The tree case is more counterintuitive but could result from the description of the case. For we can ask what one is trusting the tree *for*. Suppose one must pick a spot in a grove of trees to hang a hammock. If this is the first time hanging the hammock among these trees, one might assess the trees that seem strongest. If the tree should fail to hold the hammock, one may revise the criteria for suitable trees and avoid the tree in future. If the tree should succeed, one might be disposed to use the tree again in future. While that disposition does not arise from the responsiveness or goodwill of the tree—since it has none—it seems perfectly normal to rely on the tree in future *because* it has proved reliable in the past. If this holds, I maintain that one trusts the tree. That such trust differs from trust in romantic relationships is only to reiterate the point about trust different between dentists, business partners, children, and lovers. For there may be more varieties and conditions on trust than we readily enumerate. What matters is trust's *dispositional* function.

trusting AI. Recall the implications from Al's view introduced at the end of §3.1. Al argues that "trust *should* not be placed in AI systems, and misplaced attitudes of trust in AI should be *discouraged*" (ibid., 10; emphasis added). The point is not that trusting AI is impossible, strictly speaking. Rather, the point is normative. That is, I think we can see Al's second argument as assessing whether we *should* trust AI, not whether we can.

What reasons do we have for thinking trust in AI is normatively inappropriate? First, trusting AI can obscure the role and responsibility of designers and overseers in supervising AI. After all, the designers and overseers can be responsive to our reliance on their product. Accordingly, they can be held responsible for correcting errors in AI performance in ways that the AI itself cannot. For this reason, Al argues that responsiveness should be a condition for our trust. Second, given the limitations of AI's responsiveness and the role of humans in AI design and deployment, we should avoid trusting AI and instead attempt only to rely on AI, reserving trust for humans. This allows that there are cases in which relying on AI produces better outcomes that relying on humans for similar tasks. What it does is require consideration of the appropriate conditions for relying on AI.

This shift to evaluating the appropriate grounds for trust, rather than the possibility of trust is the topic of §5.4. Before proceeding, it is worth summarizing the results of this section. First, I argued that it is possible to trust artifacts, including AI, provided that one is disposed to rely on AI to perform certain tasks in the service of one's goals. The most plausible form of trust for AI is predictive trust. Second, although a trustee's responsiveness can play a role in trusting, it is not a necessary condition of trust. That is, it is possible to trust someone or something that is unresponsive to trust. Third,

examining the social function of trust directs our attention to the appropriate conditions for trusting AI. If one is disposed to rely on AI, then one trusts AI. The question is not then whether one can trust AI, but whether one *should*.

5.3 Should We Trust AI? Toward Trustworthy AI

I have argued that trust in AI is possible. What remains, however, is to consider under what conditions we *should* trust AI. To put the question differently: when would trust in AI be well-placed? Pursuing answers to this question connects our discussion of possible trust in AI with calls for trustworthy AI. In this section, I examine theoretical and applied elements for developing trustworthy AI. I argue that the cultivation of both is value-laden. To conclude, I contend that future research should focus on the normative project of determining the appropriate form for trust in AI.

5.3.1 Objection to the Normative Shift

One might object that the shift to considering whether we should trust AI places the cart before the horse. For instance, Thi Nguyen (2022) develops an account of trust as an unquestioning attitude. On this view, to trust something "is to put its reliability outside the space of evaluation and deliberation" (ibid., 214–15). Imagine a case where one claims to trust a friend to pick up a package, while following the person around to ensure she does it. Although following the friend can establish her reliability, providing a basis for trust, the confirmation is a prelude to trust. In the same way, while trust in AI might be possible, inquiring about whether AI is trustworthy precludes trust in AI.

This is an important point. It is possible that we trust an AI system without any of the work I describe in the next two subsections. For instance, one could trust on the basis practical or prudential reasons; there may be limited resources and the benefits of utilizing an AI system may seem to outweigh present costs. Nevertheless, it is possible that developing frameworks for assessing the trustworthiness of AI systems may reduce current trust in AI. One could argue that such a result is undesirable, given the promise of such systems for increasing efficiency and addressing real-world problems. Mounting research on the potential harms, misuses, and errors of AI systems, suggests that it is prudent and ethically desirous to pursue the development of trustworthy AI, even while this could result in short-term reductions in trust.

Nonetheless, I think deliberating about AI trustworthiness does not preclude trust. The role of one's present and possible evidence can impact trust when assessing the appropriate grounds for trust. Simpson (2017, 190) distinguishes between *following* evidence and *gathering* evidence. When following evidence, one considers first-order reasons to trust, as well as higher-order reasons—all-things-considered reasons—for trusting. This is a synchronous construal of trusting, whereby one trusts in light of one's total evidence at time t_1. In contrast, gathering evidence is a diachronic assessment of one's evidence when trusting. Like the following-friend case, gathering evidence involves assessing one's total evidence at t_1 and t_2. So long as one is following evidence about an AI, I argue that it is possible to trust the AI *and* to reflect on features relevant for trustworthiness. It could be that those assessing a system's trustworthiness do not trust it at t_1, provided at t_1 they are not disposed to rely on a system. This may be an appropriate default position for the deployment of AI, where we closely gather evidence about its

performance. correct result. This could dispose one to rely on the system at t_2. When the system again performs well at t_2, then trust is not established but increased.

Again, the *normative* point that emerges from consideration of one's evidence is not primarily about establishing whether one trusts—this is a separate, empirical matter. Rather, it is in determining the features that *should* dispose us to rely on AI in relevant circumstances. This applies across cases where trust is absent, established, weak, and strong.

5.3.2 Applied Frameworks for Trustworthy AI

Most applied approaches to trustworthy AI provide lists of characteristics that render a system trustworthy. For example, Thilo Hagendorff (2020) discusses 22 list-based frameworks for ethical AI. These lists are meant to provide guidance for practitioners in designing and deploying AI. Hagendorff argues that list-based guidelines face important limitations, ranging from enforcement and uptake to theoretical considerations of what should be included on lists and how those items relate to decisions in design. An exhaustive discussion of these issues is beyond our present purview. Rather, I emphasize here how frameworks are sensitive to what we value in trusting.

For example, HLEG provides a framework for assessing However, given that it is not necessary for trust that one is aware of trust, it is also possible that one's trust increases from t_1 *to* t_2. For example, suppose a system at t_1 provides a

relevant ethical, legal, and social questions. They argue that, at base, AI systems must be "human-centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom" (2019,

4). There are several possibilities for how AI could fail this standard. An AI's performance may fall below an acceptable threshold. The AI could perfectly achieve its design specifications but do so in the service of unjust ends.[165] In this way, we can evaluate AI according to its performance and according to its ends. This is made doubly important by what Nick Bostrom calls the "orthogonality thesis" (2012). The thesis states that "[i]ntelligence and final goals are orthogonal axes along which possible agents can freely vary" (ibid.). That is, different levels of intelligence can be combined with different goals. So, AI can be promising and perilous depending on its ends and capabilities. There is nothing necessary about a system's design, however advanced, that it serve good ends. To develop trustworthy AI, then, HLEG recommends requirements for trusting AI.

HLEG (2019, 6–8) argues for seven requirements, corresponding to legal, ethical, and technical impacts of a system. They are (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) non-discrimination and fairness, (6) societal and environmental wellbeing, and (7) accountability. For the seven requirements they identify, HLEG pilots more than sixty questions that fall under the requirements. For example, for avoiding unfair bias, they ask what features are in place to allow users and stakeholders to flag potential problems. More quantitatively, they require that a system have clearly stated metrics for measuring a system's fairness. There are lingering philosophical and statistical questions for such

[165] For example, one may object to a system's performance, as *ProPublica* famously did with respect to error rates for Equivant's (formerly Northpointe) recidivism risk prediction tool, COMPAS. Hedden (2021) nicely describes how differences in metrics result in different judgments about COMPAS' fairness. Independently of a system's performance, however, one could object to a system's objectives. For instance, on legal grounds, one could object to the use of a predictive tool in intervening on cases of child abuse and neglect. For a discussion of such a case, see Eubanks (2017), especially chapter four.

operationalizations. As Brian Hedden (2021) shows, there are many potential measures of fairness (he examines eleven) and they are not all satisfiable at once.[166] Discussing the merits and details of each is beyond the scope of this Chapter. Rather, I highlight the significance of this framework for trust in AI.

First, while not exhaustive, the framework provides conditions of trust in a particular system. For example, if one finds that a system is unreliable and unsafe, having multiple vulnerabilities for human users, it makes sense to withdraw trust. Of course, one could continue to trust a system, knowing that it is vulnerable. But the current point is that identifying relevant conditions of *appropriate* trust in AI can provide reasons for *why* we think AI is trustworthy.

This leads to a second point, namely that focusing on the potential grounds for trust orients debates about trustworthiness to what we value in trusting AI. For example, HLEG argues that their seven requirements follow from four ethical principles, including protecting personal autonomy and preventing harm. It is possible that these principles conflict. For example, predictive policing might reduce real crime rates, therefore preventing harm, but do so in ways that violate privacy and freedom (see 2019, 13). Accordingly, balancing considerations for trusting AI presents opportunities to reflect on what we value in utilizing such systems for the problems and tasks they are designed to solve.

Here, a familiar consideration arises. In Chapter 4, I examined the problem of value divergence for trust in science. A primary conclusion of that chapter was that

[166] See also Kleinberg et al. (2016) and Miconi (2017).

navigating value-laden decisions in science is itself value-laden. A similar point emerges with frameworks for trustworthy AI.

For example, while HLEG's framework builds in a moral or ethical conception of trust, consider a framework that emphasizes predictive information for establishing a system's trustworthiness. Jin-Hee Cho et al. develop measures for the "multidimensional characteristics" relevant to a system's integrity, resilience, and agility (2016, 3).[167] They call the measure TRAM, standing for trust, resilience, and agility metrics. Following Elizabeth Chang et al. (2007), Cho et al. examine the relationship of trustors, trustees, features of what is entrusted, and changes over time, to develop trust measures; for example, a measure of trust from 1 to 5. How they operationalize the framework is complex, involving attributes that are contained within higher-order items (see Figure 2). By organizing the framework in this way, Cho et al. argue that they can measure and evaluate the quality of a system, utilizing techniques like red teaming, vulnerability assessments, and penetration testing.[168] With metrics for each attribute in the framework, assessments and threats can be gauged relative to the system, yielding measures for a system's security, resilience to threats, adaptability, and so on.

[167] Cho et al. (2016) develop an "ontology-based framework," which relies heavily on Chang et al.'s (2007) trust ontologies. They define ontologies as shared conceptualizations of a domain, such that the ontology represents shared knowledge within the domain (2007, 522). For example, we could develop an ontology for human relationships, which allows us to analyze "sub-ontologies" like business and customer relationships. They focus on trust relationships as a type of ontology. Chang et al.'s complex taxonomy is beyond the purview of this Chapter. What is worth noting in passing is that they develop ontologies for agents, services, and products. That is, they help develop metrics for trusting products, including AI systems (see ibid., 529 for an examination of trust in websites).

[168] See Wood and Duggan (2000) and Goel and Mehtre (2015), cited by Cho et al.

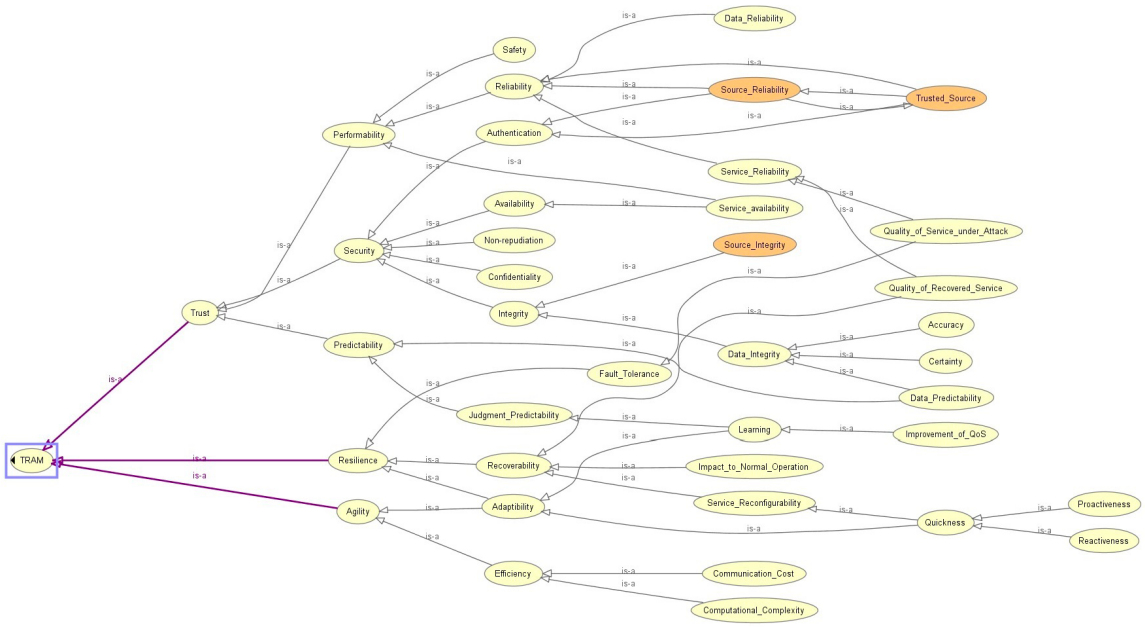

Figure 2. From Cho et al. (2016, 5). Attributes appear in circles. Arrows indicate that an attribute is a subclass of a higher-order class. For each attribute, Cho and colleagues develop measures scaled from [0,1].

Again, my objective here is not to examine every component of Cho et al.'s framework. Rather, I underscore what they and HLEG are doing in assessing possible trust in AI. *They are attempting to identify the conditions and attributes of systems that allow one to trust a system well.* One might object to the inclusion or exclusion of certain attributes—I contend that this is what we should do with respect to trust in AI. For example, Cho et al. include the security of a system as impacting levels of trust, where security is itself sensitive to the quality of data, confidentiality protections, non-repudiation (e.g., user authentication), and service availability. This might not apply to a particular system or might obscure the nature of system security. But notice how this

process differs from contentions about whether trust in AI is possible. Rather than considering whether one's selected or preferred form of trust fits well with AI, the normative task requires that philosophers work with applied scientists to examine those criteria that *should* impact our trust. This is in part technical, concerning both technical features of AI systems and relevant measures. But it is also normative in the sense that, to trust *well,* we must determine those features that render an AI trustworthy.

5.3.3 TRUSTWORTHY AI

Recently, Mona Simion and Christoph Kelp (2023a) develop a strategy to explain why some features should be included in frameworks for trustworthy AI. Their account identifies AI trustworthiness in obligations that arise from an AI's function. To begin, they suppose that, as a general rule, "we should trust S to φ when they are *trustworthy* with respect to φ-ing" (ibid., 2).[169] Determining whether someone or something is trustworthy is complex, however. They rightly note that there are competing theories of trust and that they do not all identify the same requirements for trustworthiness (ibid., 3–6). They sidestep issues forms of trust by appealing to obligations that arise through function. That is, they remark, "[t]raits, activities, and artifacts alike are governed by norms sourced in their functions" (ibid, 8). These norms provide a basis for regarding something as properly function or malfunctioning. In §5.2.1, I introduced the two types of functions that Simion and Kelp discuss, namely etiological and design functions. Further, they argue that "the conditions put forth to distinguish [trust from mere reliance] are too anthropocentric to do the job of accounting for trustworthiness in the case of AI"

[169] There are clear limitations of and exceptions to this rule; for instance, therapeutic trust can promote valuable goods (moral education, social reintegration, and so on).

(ibid., 3). So, an account of AI trustworthiness should avoid ascribing psychological or motivational dimensions to trust.

There are three components to this approach to trustworthiness. First, they define "outright trustworthiness" as follows:

> For all x where x is an AI, "x is trustworthy" is true in context *c* if and only if x approximates maximal trustworthiness to *phi* for all *phi* closely enough to surpass a threshold on degrees of trustworthiness determined by *c*. (2023a, 9)

So, outright trustworthiness is a threshold concept, whereby one is trustworthy by surpassing a threshold established by one's context. In this sense, we might say that something is trustworthy *enough*.

Second, maximal trustworthiness is defined in dispositional terms. That is, a trustee is "maximally trustworthy with regard to *phi*-ing if and only if x has a maximally strong disposition to meet its functional norms-sourced obligations to *phi*" (ibid.). Elsewhere, Simion and Kelp explain that dispositions "have trigger and manifestation conditions" (2023b, 669). For example, an archer could have a disposition to hit a target (ibid.). The trigger of the archer's disposition is loosing an arrow, while the manifestation of the disposition is hitting the target. For AI, we source obligations in the design and etiology of the technology. From a design perspective, AI is properly functioning when it fulfills whatever it was designed to do. For instance, an AI system in radiology may be designed to detect bone fractures. Alongside design functions, AI is properly functioning etiologically when it reliably produces results in normal contexts. Simion and Kelp offer the following example to show how the two functions differ. Imagine a diagnostic AI that fails to recognize simple tumors, while successfully identifying complex cases. While the system meets its design-sourced obligations, they argue that the diagnostic AI "is

malfunctioning etiologically, in that the recognizing of tumours by the type of artifact it belongs to contributes to the explanation of the continuous existence of cancer diagnostic AIs" (2023a, 10). When design and etiological functional norms diverge, they argue that the latter override the former, "because reliable function fulfilment comes first in functional items, and proper [etiological]-functioning, but not proper [design]-functioning, delivers it" (ibid.). This is the vital point for trustworthy AI in dispositional terms, viz., *maximal* trustworthiness for AI technologies is determined by how the technology completes a task, relative to intended design and outcomes that explain the persistent success of the system.

Third, one need not be maximally trustworthy to be trustworthy. Rather, trustworthiness can come in degrees, allowing for comparisons and improvements. In Simion and Kelp's view, one's degree of trustworthiness is a function of the distance between one's disposition to φ and maximal trustworthiness with respect to φ in a context. That is, "the closer x approximates maximal trustworthiness to [φ], the higher x's degree of trustworthiness to [φ]" (ibid.). So, given the sourcing of obligations in functional dispositions with the previous point, we can assess the degree of an AI's trustworthiness by considering how well it accomplishes the relevant task in a particular context.

This approach to AI trustworthiness has several advantages. First, as Simion and Kelp contend, it does not require "highbrow stipulations concerning AI psychology" (ibid.). That is, their view can avoid anthropomorphizing AI. Second, their view allows for comparisons. For instance, a diagnostic AI that can recognize simple tumors is more trustworthy than the version discussed above, which failed to do so, even if both systems

follow their design plans perfectly. Why is this? It is because the AI that fails to recognize garden-variety tumors has a lower degree of trustworthiness relative to maximal trustworthiness for diagnostic imagining (ibid.). In this way, Simion and Kelp conclude: "trustworthy-making properties for AIs are properties that map on to their having a disposition to fulfill their functionally sourced obligations" (ibid.).

Nonetheless, a number of crucial features remain underdetermined. Simion and Kelp argue that their view explains why frameworks for AI trustworthiness include conditions for safety, fairness, human-centeredness, and beneficence, namely because those features contributed to the proper functioning of the AI (ibid.). However, it seems unclear to me why an AI's function need produce obligations concerning, for example, fairness. AI ethicists working on unfair bias appeal to ethical reasons, values, and principles for including fairness in a framework for ethical AI.[170] This is not to say that fairness is irrelevant for trusting AI or for evaluating whether an AI functions appropriately. The point is that *proper* functioning requires determining at least the following: an AI's appropriate task, *how* that task is accomplished, the range of tasks relevant for determining functional norms, means for evaluating how the task is accomplished, and the threshold for degrees of trustworthiness that is satisfactory or appropriate. Additional complications arise when different elements of a framework are in tension. For instance, Alice Xiang (2022) examines tensions between monitoring bias and protecting privacy in computer vision systems that use AI for facial recognition. That is, there are trade-offs between accessing potentially sensitive information to root our

[170] Consider HLEG's turn to ethical principles to justify elements of their framework. See also Fazelpour and Danks (2021).

bias and protecting rights to privacy. Navigating these trade-offs requires decisions about how to balance competing aims.

5.4 Concluding Remarks

In this chapter, I have argued that trust in AI is possible. The pressing question for research on trust and trustworthy AI is not whether trust is possible, but under what conditions trust is *normatively appropriate*. If we can trust AI, much is left for consideration, including who *we* are, the conditions of our *trust*, what our *goals* and *resources* are, and more besides. It has not been my aim in this chapter to elucidate all these features. Instead, given the rapid development and increasing ubiquity of AI, it is crucial to consider not whether we can trust AI—we can—but in what circumstances we *should*.

6.0 CONCLUSION

In sum, I contend that pragmatic pluralism is a plausible, descriptive approach to trust. I argue that it provides solutions to two general problems for monist approaches to trust, viz., the counterexample problem and the explanatory problem. In Chapter Three, I show how empirical trust research suggests pragmatic pluralism. Specifically, investigations of multidimensionality and modeling techniques provide means for philosophers of trust to contribute to and learn from empirical investigations of trust. In Chapters Four and Five, I examine how sensitivity to the possible conditions of trust can contribute to approaching values in science and developing trustworthy AI. In the end, the upshot of pragmatic pluralism is that trust and, by extension, trustworthiness are value-laden. That is, the conditions that dispose one to rely on another are sensitive to the values, needs, and expectations of trustors and trustees.

Nonetheless, as I suggest in Chapter Two, there are limits to pragmatic pluralism. As a descriptive thesis, the view allows us to describe variability in cases of trust. Trust can lead people to rely under myriad conditions and for multiple ends. We may rightly disapprove of some cases, as in sexist, racist, or abusive forms of trust. Yet, these are indeed forms of trust precisely in the sense that certain conditions (appropriately or not) dispose trustors to rely on trustee. Of itself, pragmatic pluralism offers limited normative input for determining when trust is appropriate or not. What it provides, however, is a means for approaching questions about what should dispose one to rely on another in relevant circumstances. In this way, pragmatic pluralism can play an important descriptive part in developing an ethics of trust.

BIBLIOGRAPHY

Adler, Jonathan. 1994. "Testimony, Trust, Knowing," *The Journal of Philosophy*, 91:5, 264–275.

Al, Pepjin. 2022. "(E)-Trust and Its Function: Why We Shouldn't Apply Trust and Trustworthiness to Human-AI Relations," *Journal of Applied Philosophy*, 40: 95-108. https://doi.org/10.1111/japp.12613

Allum, Nick. 2007. "An empirical test of competing theories of hazard-related trust: The case of GM food," *Risk Analysis*, 27(4): 935–946.

Alonso, Facundo. 2016. "Reasons for Reliance," *Ethics* 126, 311–38.

Alonso, Facundo. 2009. "Shared Intention, Reliance, and Interpersonal Obligation," *Ethics* 119, 444–75.

Alonso, Facundo. 2014 "What Is Reliance?" *Canadian Journal of Philosophy* 44, 163–83.

Anderson, Elizabeth. 1995. "Knowledge, Human Interests, and Objectivity in Feminist Epistemology," *Philosophical Topics,* 23:7–58.

Anderson, Elizabeth. 2004. "Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce," Hypatia 19 (1): 1–24.

Anderson, Elizabeth. 2011. "Democracy, Public Policy, and Lay Assessment of Scientific Testimony," *Episteme* 8 (2): 144–64.

Aristotle. 2009. *The Nicomachean Ethics*, Trans. by David Ross and Edited by Lesley Brown, Oxford: Oxford University Press.

Ashford, Nicholas. 1988. "Science and Values in the Regulatory Process," *Statistical Science,* 3: 377–83.

Bacharach Michael. 1999. "Interactive team reasoning: A contribution to the theory of cooperation," *Research in Economics*, 53 (2): 117-147.

Baier, Annette. 1994. *Moral Prejudices: Essays on Ethics*. Cambridge, MA: Harvard University Press.

Baier, Annette. 1986. "Trust and Antitrust," *Ethics*, (96)2: 231–60.

Bengio, Yoshua. 2009. *Learning Deep Artichitectures for AI*. IEEE, doi: 10.1561/2200000006.

Besley, John and Leigh Tiffany. 2023. "What are you assessing when you measure 'trust' in scientists with a direct measure?" *Public Understanding of Science*, (32)6: 709-726.

Betz, Gregor. 2013. "In Defense of the Value Free Ideal," *European Journal for Philosophy of Science,* 3: 207–220.

Biddle, Justin. 2013. "State of the field: Transient underdetermination and values in science," *Studies in History and Philosophy of Science*, 44:124–133.

Blakely, Jason. 2023. "Doctor's Orders: COVID-19 and the new science wars," *Harper's Magaince*, URL: https://harpers.org/archive/2023/08/doctors-orders-jason-blakely/

Blancke, Stefaan, Maarten Boudry, and Massimo Pigliucci. 2017. "Why Do Irrational Beliefs Mimic Science? The Cultural Evolution of Pseudoscience," *Theoria*, 83: 78-97. https://doi.org/10.1111/theo.12109

Bluhm, Robyn. 2017. "Inductive risk and the role of values in clinical trials." In K. C. Elliott, & T. Richards (Eds.), *Exploring Inductive Risk: Case Studies of Values in Science*, New York: Oxford University Press, 193–212.

Boden, Margaret. 2018. *Artificial Intelligence: A Very Short Introduction*. Oxford: Oxford University Press.

Bossmann, Julia. 2016. "Top 9 ethical issues in artificial intelligence," World Economic Forum, URL: https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

Boulicault, Marion and S. Andrew Schroeder. 2021. "Public Trust in Science: Exploring the Idiosyncrasy-Free Ideal." In *Social Trust*, (eds.) Kevin Vallier and Michael Weber, New York: Routledge, 102–21.

Branch, T. Y. 2022. "Enhanced Epistemic Trust and the Value-Free Ideal as a Social Indicator of Trust," *Social Epistemology*, 36:5, 561–575, DOI: 10.1080/02691728.2022.2114114

Branch T. Y. and Gloria Origgi. 2022. "Social Indicators of Trust in the Age of Informational Chaos," *Social Epistemology*, 36:5, 533-540, DOI: 10.1080/02691728.2022.2121622

Braun, Matthias, Hannah Bleher, and Patrik Hummel. 2021. " A Leap of Faith: Is There a Formula for 'Trustworthy' AI?" *Hastings Center Report,* (51)3 (2021): 17–22. DOI: 10.1002/hast.1207

Brennan, Johnny. 2020. "Can Novices Trust Themselves to Choose Trustworthy Experts? Reasons for (Reserved) Optimism," *Social Epistemology*, (34)3: 227-240, DOI: 10.1080/02691728.2019.1703056

Bronfman Nicolas and Esperanza Vázquez. 2011. "A cross-cultural study of perceived benefit versus risk as mediators in the trust-acceptance relationship," *Risk Analysis,* 31(12):1919-34. doi: 10.1111/j.1539-6924.2011.01637.x.

Brown, Matthew. 2020. *Science and Moral Imagination: A New Ideal for Values in Science.* Pittsburgh: University of Pittsburgh Press.

Brown, Matthew. 2018. "Weaving Value Judgment into the Tapestry of Science." *Philosophy, Theory, and Practice in Biology*, 10: 8. http://doi.org/10.3998/ptpbio.16039257.0010.010

Bryson, Joanna. 2018. "AI & Global Governance: No One Should Trust AI," *UNU-CPR*, URL: https://unu.edu/cpr/blog-post/ai-global-governance-no-one-should-trust-ai.

Buchan, Nancy, Rachel Croson, and Robin Dawes. 2002. "Swift Neighbors and Persistent Strangers: A Cross-Cultural Investigation of Trust and Reciprocity in Social Exchange," *American Journal of Sociology*, 108(1), 168–206.

Carter, J. Adam. 2022. "Therapeutic trust," *Philosophical Psychology*, DOI: 10.1080/09515089.2022.2058925

Carter, J. Adam and Mona Simion. 2020. “The Ethics and Epistemology of Trust,” *Internet Encyclopedia of Philosophy*, URL: https://iep.utm.edu/trust/

Castaldo, Sandra, Katia Premazzi, and Fabrizio Zebrini. 2010. “The Meaning(s) of Trust: A Content Analysis of the Diverse Conceptualizations of Trust in Scholarly Research on Business Relationships,” *Journal of Buisness Ethics*, 96: 657–668. DOI 10.1007/s10551-010-0491-4

Castelfranchi, Cristiano and Rino Falcone. 2010. *Trust Theory: A Cocio-Cognitive and Computational Model*, Singapore: John Wiley and Sons.

Castelfranchi, Cristiano and Rino Falcone. 2020. “Trust: Perspectives in Cognitive Science,” in *The Routledge Handbook of Trust and Philosophy*, (ed.) Judith Simon, 214–228.

Castro, Clinton. 2019. “What's wrong with machine bias,” *Ergo*, 6 (15).

Chang, Elizabeth, Tharam Dillon, and Farookh Hussain. 2007. “Trust Ontologies for E-Service Environments,” *International Journal of Intelligent Systems*, 22: 519–545.

Cho, Jin-Hee, Patrik Hurley, and Shouhuai Xu. 2016. “Metrics and Measurement of Trustworthy Systems,” *MILCOM 2016 - 2016 IEEE Military Communications Conference*, Baltimore, MD, USA, 1237-1242, doi: 10.1109/MILCOM.2016.7795500.

Clough, Sharyn and William Loges. 2008. “Racist Value Judgements as Objectively False Beliefs: A Philosophical and Social-Psychological Analysis,” *Journal of Social Philosophy* 39: 77–95.

Coeckelbergh, Mark. 2012. “Can We Trust Robots?” *Ethics and Information Technology*, 14: 53–60.

Coleman, James. 1990. *Foundations of Social Theory.* Cambridge, MA: Belknap Press.

Colquitt, Jason and Jessica Rodell. 2011. “Justice, trust, and trustworthiness: A longitudinal analysis integrating three theoretical perspectives.” *Academy of Management Journal*, 54, 1183–1206.

Connor, Melanie and Michael Siegrist. 2010. “Factors Influencing People’s Acceptance of Gene Technology: The Role of Knowledge, Health Expectations, Naturalness, and Social Trust,” *Science Communication*, *32*(4), 51–538. https://doi.org/10.1177/1075547009358919

Contessa, Gabriele. 2022. “It Takes a Village to Trust Science: Towards a (Thoroughly)Social Approach to Public Trust in Science,” *Erkenntnis*, 88: 2941–2966.

Cook, Karen. 2016. “Trust,” *Oxford Bibliographies*. URL: https://www.oxfordbibliographies.com/display/document/obo-9780199756384/obo-9780199756384-0062.xml

Cook, Karen and Jessica Santana. 2020. “Trust: Perspectives in Sociology,” in *Routledge Handbook of Trust and Philosophy*, (ed.) Judith Simon, 189–204.

Corbett-Davies, Sam and Sharad Goel. 2018. “The measure and mismeasure of fairness: A critical review of fair machine learning” arXiv preprint arXiv:1808.00023.

Corriveau, Kathleen and Paul Harris. 2009. “Choosing your informant: Weighing familiarity and recent accuracy,” *Developmental Science*, 12: 426–437. doi:10.1111/j.1467-7687.2008.00792.x

Corriveau, Kathleen, Elizabeth Kim, Ge Song, and Paul Harris. 2013. “Young Children’s Deference to a Consensus Varies by Culture and Judgment Setting,” *Journal of Cognition and Culture*, (13): 367–381.

Corriveau, Kathleen and Katelyn Kurkul. 2014. “’Why does rain fall?’: Children prefer to learn from an informant who uses noncircular explanations,” *Child Development, 85*(5), 1827–1835.

Cvetkovich, George, Michael Siegrist, Rachel Murray, Sarah Tragesser. 2002. “New information and social trust: Asymmetry and perseverance of attributions about hazard managers,” *Risk Analysis*, 22(2): 359–367.

Darwall, Stephen. 2017. “Trust as a Second-personal Attitude (of the Heart),” in *The Philosophy of Trust*, ed. Paul Faulkner and Thomas Simpson, Oxford: Oxford University Press: 35–50.

Dasgupta, Partha. 1988. “Trust as a commodity,” in *Trust*, (ed) D. G. Gambetta, New York: Basil Blackwell, 49-72.

Deutsch, Morton. 1973. *The resolution of conflict*. New Haven, CT: Yale University Press.

Deutsch, Morton. 1958. “Trust and suspicion,” *Journal of Conflict Resolution,* 2: 265-279.

Dimoka, Angelika. 2010. “What Does the Brain Tell Us About Trust and Distrust? Evidence from a Functional Neuroimaging Study.” *MIS Quarterly*, *34*(2): 373–396. https://doi.org/10.2307/20721433

Domenicucci, Jacopo and Richard Holton. 2017. “Trust as a Two- Place Relation”. In *The Philosophy of Trust,* (eds) Paul Faulkner and Thomas Simpson, Oxford: Oxford University Press.

Dormandy, Katherine. 2020. “Exploitative Epistemic Trust,” in *Trust in Epistemology,* (ed.) Katherine Dormandy, 241–264.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal.* Pittsburgh: University of Pittsburgh Press.

Douglas, Heather. 2017. “Why inductive risk requires values in science,” in *Current Controversies in Values in Science,* K. C. Elliott and D. Steel, New York: Routledge, 81–93.

Earle, Timothy and George Cvetkovich. 1995. *Social trust: Toward a cosmopolitan society.* Westport, CT: Praeger.

Earle, Timothy and Michael Siegrist. 2006. “Morality information, performance information, and the distinction between trust and confidence,” *Journal of Applied Social Psychology*, 36: 383–416.

Earle, Timothy and Michael Siegrist. 2008. “On the relation between trust and fairness in environmental risk management,” *Risk Analysis*, 28(5): 1395–1413.

Earle, Timothy, Michael Siegrist, and Heinz Gutscher. 2007. "Trust, risk perception, and the TCC model of cooperation," in *Trust in cooperative risk management: Uncertainty and scepticism in the public mind*, (eds.) M. Siegrist, T. C. Earle, & H. Gutscher, London: Earthscan,1–49.

Eiser, J. Richard, Amy Donovan, and R. Stephen Sparks. 2015. "Risk perceptions and trust following the 2010 and 2011 Icelandic volcanic ash crises," *Risk Analysis*, 35(2): 332–343.

Elish, Madeleine. 2019. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction," *Engaging Science, Technology, and Society*, 5: 40–60.

Elliott, Kevin. 2022. *Values in Science*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009052597

Elliott, Kevin. 2017. *A Tapestry of Values: An Introduction to Values in Science*, New York: Oxford University Press.

Elliott, Kevin. 2021. "The Value-Ladenness of Transparency in Science: Lessons from Lyme Disease*." Studies in History and Philosophy of Science*, 88: 1–9.

Elliott, Kevin. 2011. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*, New York: Oxford University Press.

Elliott, Kevin and David Resnik. 2014. "Science, Policy, and the Transparency of Values." *Environmental Health Perspectives,* 122: 647–50.

Elliott, Kevin and Ted Richards. 2017. *Exploring Inductive Risk: Case Studies of Values in Science,* New York: Oxford University Press.

Elliott, Kevin, and Daniel McKaughan. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science*, 81(1): 1–21.

Elliott, Kevin, Aaron McCright, Summer Allen, and Thomas Dietz. 2017. "Values in environmental research: Citizen's views of scientists who acknowledge values," *PLoS One*, (12)10: e0186049. https://doi.org/10.1371/journal.pone.0186049

Eubanks, Virginia. 2017. *Automating inequality.* New York, NY: St. Martin's Press.

Faulkner, Paul. 2011. *Knowledge on Trust*, Oxford: Oxford University Press.

Faulkner, Paul. 2015. "The attitude of trust is basic," *Analysis*, (75)3: 424–429. doi:10.1093/analys/anv037

Fazelpour, Sina and David Danks. 2021. "Algorithmic bias: Senses, sources, solutions," *Philosophy Compas*, (16)8: https://doi.org/10.1111/phc3.12760.

Fazelpour, Sina and Zachary Lipton. 2020. "Algorithmic Fairness from a Non-ideal Perspective," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 57–63. New York City: ACM.

Ferrario, Andrea, Michele Loi, and Eleonora Viganò. 2020. "In AI We Trust Incrementally: A Multi-Layer Model of Trust to Analyze Human-Artificial Intelligence Interactions," *Philosophy and Technology*, 33: 523–39.

Frances, Bryan and Jonathan Matheson. 2018. “Disagreement,” in *The Stanford Encyclopedia of Philosophy*, (ed) by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/disagreement/

Fricker, Elizabeth. 2006. “Testimony and Epistemic Autonomy,” in *The Epistemology of Testimony*, (eds.) Jennifer Lackey and Ernest Sosa, Oxford: Clarendon Press, 225–50

Fukuyama, Francis. 1995. *Trust: The social virtues and the creation of prosperity*. New York: Free Press.

Funk, Cary. 2017. “Mixed Messages about Public Trust in Science,” *Issues in Science and Technology*, 34, URL: https://issues.org/real-numbers-mixed-messages-about-public-trust-in-science/

Gambetta, Diego. 1988. “Can We Trust Trust?” in *Trust: Making and Breaking Cooperative Relations*, (ed) Diego Gambetta, Oxford: Basil Blackwell, 213–238.

Garfinkel, Harold. 1967. *Studies in Ethnomethodology.* Englewood Cliffs, NJ: Prentice Hall.

Gaskell, George, Nic Allum, Wolfgang Wagner, Nicole Kronberger, Helge Torgersen, Juergen Hampel, and Julie Bardes. 2004. “GM foods and the misperception of risk perception,” *Risk Analysis* (24)1: 185–94.

Gates, Bill. 2023. “The Age of AI has begun,” *GatesNotes*, URL: https://www.gatesnotes.com/The-Age-of-AI-Has-Begun?WT.mc_id=20230321100000_Artificial-Intelligence_BG-TW_&WT.tsrc=BGTW.

Goel, Jai Narayan and B.M. Mehtre. 2015. “Vulnerability Assessment & Penetration Testing as a Cyber Defence Technology,” *Procedia Computer Science*, 57: 710–715.

Goetze, Trystan. “Okay, Google, Can I Trust You? An Anti-trust Argument for Antitrust,” in *The Moral Psychology of Trust*, (eds.) D. Collins, I.V. Jovanović, Mark Alfano, Lanham, MD: Lexington Books, 237–257.

Goldberg, Sandford. 2020. “Trust and Reliance,” in *The Routledge Handbook of Trust and Philosophy*, ed. Judith Simon, 97–108.

Goldenberg, Maya. 2021. *Vaccine Hesitancy: Public Trust, Expertise, and the War on Science*, Pittsburgh: University of Pittsburgh Press.

Goldman, Alvin. 2001. “Experts: Which Ones Should You Trust?” *Philosophy and Phenomenological Research,* 63: 85-110. https://doi.org/10.1111/j.1933-1592.2001.tb00093.x

Guerrero, Alexander. 2017. “Living with Ignorance in a World of Experts,” In *Perspectives on Ignorance from Moral and Social Philosophy*, (ed) by R. Peels, New York: Routledge, 135–177.

Habyarimana, James, Macartan Humphreys, and Daniel Posner. 2009. *Coethnicity: Diversity and the Dilemmas of Collective Action*, New Your: The Russell Sage Foundation.

Hagendorff, Thilo. 2020. “The Ethics of AI Ethics: An Evaluation of Guidelines,” *Minds and Machines*, 30: 99–120. DOI: https://doi.org/10.1007/s11023-020-09517-8.

Hamm, Joseph, Lesa Hoffman, Alan J. Tomkins, and Brian H. Bornstein. 2016. "On the influence of trust in predicting rural land owner cooperation with natural resource management institutions," *Journal of Trust Research,* 6:1, 37-62, DOI: 10.1080/21515581.2015.1108202.

Hamm, Joseph and Lesa Hoffman. 2016. "Working with Covariance: Using Higher-Order Factors in Structural Equation Modeling with Trust Constructs," in *Interdisciplinary Perspectives on Trust*, eds. Shockley E., Neal T., PytlikZillig L., Bornstein B.. New York: Springer. https://doi.org/10.1007/978-3-319-22261-5_5.

Handley-miner, Isaac, Michael Pope, Richard Atkins, S. Mo Jones-Jang, Daniel McKaughan, Jonathan Phillips, and Liane Young. 2023. "The intentions of information sources can affect what information people think qualifies as trust," *Scientific Reports*, 13: https://doi.org/10.1038/s41598-023-34806-4

Hannon, Michael. 2019. *What's the Point of Knowledge? A Function-First Epistemology*. New York: Oxford University Press.

Hao, Karen and Jonathan Stray. 2019. "Can you make AI fairer than a judge?" *MIT Technology Review*, October.

Hardin, Russell. 2002. *Trust and Trustworthiness*, New York: Russell Sage Foundation.

Hardin, Russell. 2006. *Trust*, Cambridge: Polity Press.

Hardwig, John. 1991. "The Role of Trust in Knowledge." *Journal of Philosophy,* 88(12): 693–708.

Harris, Paul, Melissa Koenig, Kathleen Corriveau, and Vikram Jaswal. 2018). "Cognitive foundations of learning from testimony," *Annual Review of Psychology*, 69: 251-273.

Harris, Paul. 2012. *Trusting what you're told: How children learn from others.* Cambridge, MA: Belknap Press and Harvard University Press.

Hasselberger, William and Micah Lott. 2023. "Where lies the grail? AI, common sense, and human practical intelligence," *Phenomenology and the Cognitive Sciences*, https://doi.org/10.1007/s11097-023-09942-x

Hatherley, Joshua. 2020. "Limits of trust in medical AI," *Journal of Medical Ethics*, 46: 478–481.

Havstad, Joyce and Matthew J. Brown. 2017. "Neutrality, Relevance, Prescription, and the IPCC," Public Affairs Quarterly (31)4: 303–24.

Hawley, Katherine. 2019. *How To Be Trustworthy*, Oxford: Oxford University Press.

Hedden, Brian. 2021. "On statistical criteria of algorithmic fairness," *Philosophy & Public Affairs*, 49: 2019–231. DOI: https://doi.org/10.1111/papa.12189.

Hellman, Deborah. 2020. "Measuring algorithmic fairness," *Virginia Law Review*, 106 (4).

Hieronymi, Pamela. 2008. "The Reasons of Trust," *Australasian Journal of Philosophy* 86, 213–36.

HLEG AI. 2019. "Ethics guidelines for trustworthy AI." Retrieved from High-Level Expert Group on Artifcial Intelligence.

Hoffmann, Anna. 2019. "Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse," *Information, Communication & Society*, 22(7): 900–915. DOI: https://doi.org/10.1080/1369118X.2019.1573912.

Holton, Richard. 1994. "Deciding to Trust, Coming to Believe," *Australasian Journal of Philosophy*, (72)1:63–76.

Horsburgh, H. J. N. 1960. "The Ethics of Trust," *The Philosophical Quarterly*, 10:41, 343–354. DOI:10.2307/2216409

Howard-Snyder, Daniel and Daniel J. McKaughan. Manuscript. "Relying on someone to do something"

Intemann, Kristen. 2001. "Science and Values: Are Value Judgments Always Irrelevant to the Justification of Scientific Claims?" *Philosophy of Science*, (68)3: S506–S518.

Intemann, Kristen. 2015. "Distinguishing between Legitimate and Illegitimate Values in Climate Modeling," *European Journal for Philosophy of Science*, 5: 217–32.

Intemann, Kristen. 2005. "Feminism, Underdetermination, and Values in Science," *Philosophy of Science*, 72: 1001–12.

IPCC. 2023. "Summary for Policymakers," in *Climate Change 2023: Synthesis Report.* Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, pp. 1-34, doi: 10.59327/IPCC/AR6-9789291691647.001.

IPCC, 2001. *Third Assessment Report of the Intergovernmental Panel on Climate Change,* Geneva: IPCC.

IPCC. 2007. *Fourth Assessment Synthesis Report of the Intergovernmental Panel on Climate Change,* Geneva: IPCC.

Irzik, Gürol and Faik Kurtulmus. 2021. "Well-ordered science and public trust in science," *Synthese*, 198(Suppl 19): 4731–4748.

Irzik, Gürol and Faik Kurtulmus. 2019. "What Is Epistemic Trust in Science?" *British Journal for Philosophy of Science*, 70 (4), 1145–1166.

Jaswal, Vikram and Leslie Neely. 2006. "Adults Don't Always Know Best: Preschoolers Use Past Reliability over Age When Learning New Words," *Psychological Science* 17: 757–8.

John, Stephen. 2017. "From Social Values to P-Values: The Social Epistemology of Intergovernmental Panel on Climate Change," *Journal of Applied Philosophy*, (34)2: 157–171.

John, Stephen. 2018. "Epistemic Trust and the Ethics of Science Communication: Against Transparency, Openness, Sincerity and Honesty." *Social Epistemology* 32: 75–87.

John, Stephen. 2019. "Science, Truth, and Dictatorship: Wishful Thinking or Wishful Speaking?" *Studies in History and Philosophy of Science*, 78: 64–72.

John, Stephen. 2015. “The Example of the IPCC Does Not Vindicate the Value Free Ideal: A Response to Gregor Betz.” *European Journal for Philosophy of Science*, 5: 1–13.

Jones, Karen. 2004. “Trust and Terror,” in *Moral Psychology: Feminist Ethics and Social Theory*, eds. Peggy DesAutels and Margaret U. Walker, 3–19.

Jones, Karen. 1996. “Trust as an Affective Attitude,” *Ethics*, (107)1: 4–25.

Jones, Karen. 2012. “Trustworthiness,” *Ethics*, 123(1): 61–85.

Jones, James. 2008. “The Tuskegee Syphilis Experiment,” In (eds) Emanuel, Ezekiel J.; Grady, Christine; Crouch, Robert A.; Lie, Reidar K.; Miller, Franklin G.; Wendler, David, T*he Oxford Textbook of Clinical Research Ethics*. Oxford; New York: Oxford University Press, 2008: 86–96.

Kaplan, Andreas and Michael Haenlein. 2019. “Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence,” *Business Horizons*, 62(1): 15–25. DOI: https://doi.org/10.1016/j.bushor.2018.08.004.

Kaul, Vivek, Sarah Enslin, and Seth Gross. 2020. “History of artificial intelligence in Medicine,” *Gastrointestinal Endoscopy*, 92(4): 807–812.

Kearns, Michael and Aaron Roth. 2019. *The Ethical Algorithm*. Oxford, UK: Oxford University Press.

Keohane, Robert, Melissa Lane, and Michael Oppenheimer. 2014. “The ethics of scientific communication under uncertainty,” *Politics, Philosophy & Economics*, (13)4: 343–368.

Keren, Arnon. 2020. “Trust and Belief,” in *Routledge Handbook of Trust and Philosophy*, (ed) Judith Simon, 109–120.

Keren, Arnon. 2014. “Trust and Belief: A Preemptive Reasons Account,” *Synthese*, (191)12: 2593–615.

Kitcher, Philip. 2011. *Science in a Democratic Society*. Amherst, NY: Prometheus Books.

Kitcher, Philip. 2001. *Science, Truth, and Democracy*. New York: Oxford University Press.

Kourany, Janet. 1998. “Philosophy of Science: A New Program for Philosophy of Science, in Many Voices,” *In Philosophy in a Feminist Voice: Critiques and Reconstructions*, (ed.) by Janet A. Kourany, Princeton, NJ: Princeton University Press, 231–62.

Kourany, Janet. 2010. *Philosophy of Science after Feminism*. Oxford: Oxford University Press.

Kreps, David. 1990. “Corporate Culture and economic theory,” in *Perspectives on Positive Political Economy*, (eds.) J. E. Alt and K. A. Shepsle, Cambridge: Cambridge University Press, 90–143.

Kvanvig, Jonathan. 2018. *Faith and Humility*, Oxford: Oxford University Press.

Lackey, Jennifer. 2020. *The Epistemology of Groups*. New York: Oxford University Press.

Lee, John and Katrina See. 2004. “Trust in Automation: Designing for Appropriate Reliance,” *Human Factors*, (46)1: 50–80.

Levy, Neil. 2022. *Bad Beliefs: Why They Happen to Good People*. New York: Oxford University Press.

Lewicki, Roy, Daniel McAllister, and Robert Bies. 1998. "Trust and Distrust: New Relationships and Realities," *The Academy of Management Review*, (23)3: 438–458.

Longino, Helen. 2002. The Fate of Knowledge. Princeton, NJ: Princeton University Press.

Longino, Helen. 1990. Science as Social Knowledge. Princeton, NJ: Princeton University Press.

Loury, Glenn. 1977. "A Dynamic Theory of Racial Income Differences," in *Women, Minorities, and Employment Discrimination*, 153–188.

Luhmann, Niklas. 1979. *Trust and Power.* Chichester, England: Wiley.

Marsh, Stephen and Mark Dibben. 2005. "Trust, untrust, distrust and mistrust–an exploration of the dark(er) side," In Trust management: Third international conference, (eds.) P. Herrmann, V. Issarny, & S. Shiu, 17–33.

Marušić, Berislav. 2017. "Trust, Reliance and the Participant Stance," *Philosophers' Imprint*, (17)17: 1–10.

Mayer, Roger, James Davis, and David Schoorman. 1995). "An integrative model of organizational trust," *Academy of Management Review,* 20, 709–734.

McEvily, Bill and Marco Tortoriello. 2011. "Measuring trust in organisational research: Review and recommendations," *Journal of Trust Research*, 1(1), 23–63.

McGeer, Victoria. 2008. "Trust, hope and empowerment," *Australasian Journal of Philosophy*, 86:2, 237-254, DOI: 10.1080/00048400801886413

McGreer, Victoria and Philip Petit. 2017. "The Empowering Theory of Trust," in *The Philosophy of Trust*, ed. Paul Faulkner and Thomas Simpson, Oxford: Oxford University Press: 14–34.

McKaughan, Daniel, and Kevin Elliott. 2013. "Backtracking and the Ethics of Framing: Lessons from Voles and Vasopressin," *Accountability in Research*, 20: 206–226.

McKnight, D. Harrison and Vivek Choudhury. 2006. "Distrust and trust in B2C e-commerce: do they differ?" *ICEC '06: Proceedings of the 8th international conference on Electronic commerce*, 482–491, DOI: https://doi.org/10.1145/1151454.1151527.

McLeod, Carolyn. 2015. "Trust," in *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/trust/.

McMullin, Ernan. 1983. "Values in Science," In *PSA 1982: Proceedings of the 1982 Biennial Meeting of the Philosophy of Science Association*, vol. 2, ed. Peter D. Asquith and Thomas Nickels, 3–28. East Lansing, MI: Philosophy of Science Association.

McMyler, Benjamin. 2017. "Deciding to Trust," in *The Philosophy of Trust*, ed. Paul Faulkner and Thomas Simpson, Oxford: Oxford University Press: 161–76.

McMyler, Benjamin. 2011. *Testimony, Trust, and Authority*, New York: Oxford University Press.

Midden, Cees and Nicole Huijts. 2009. “The Role of Trust in the Affective Evaluation of Novel Risks: The Case of CO2 Storage,” *Risk Analysis*, (29)5: 743–751. DOI: 10.1111/j.1539-6924.2009.01201.x

Mondschein, Ken (ed.). 2023, *Aesop’s Fables Illustrated.* San Diego: Canterbury Books.

Mostafa, Bossy, Moha El-Attar, Samy Abd-Elhaffez, Wael Awad. 2020. “Machine and Deep Learning Approaches in Genome: Review Article,” *Alfarama Journal of Basic and Applied Sciences*, 2(1): 105–113.

National Academies of Sciences, Engineering, and Medicine (NAS). 2018. *Open Science by Design: Realizing a Vision for 21st Century Research.* Washington, DC: The National Academies Press.

Nelson, Lynn Hankinson. 1993. “A Question of Evidence,” *Hypatia,* (8)2: 172–89.

Nickel, Philip. 2017. “Being Pragmatic about Trust,” in *The Philosophy of Trust*, (eds) Paul Faulkner and Thomas Simpson, 195–213.

Nickel, Philip. 2007. “Trust and Obligation-Ascription,” *Ethical Theory and Moral Practice* 10, 309–19.

Nickel, Philip. 2011. “Ethics in e-Trust and e- Trustworthiness: The Case of Direct Computer-Patient Interfaces,” *Ethics and Information Technology* 13: 355–63

Nickel, Philip J. 2013. “Trust in Technological Systems,” in *Norms in Technology*, ed. by M.J. De Vries, S.O. Hanson, and A.W.M. Meijers, 223–37.

Nickel, Philip J., Maarten Franssen, and Peter Kroes. 2010. “Can We Make Sense of the Notion of Trustworthy Technology?” *Knowledge, Technology & Policy*, 23: 429–44.

Nguyen, C. Thi. 2022. “Transparency is Surveillance,” *Philosophy and Phenomenological Research*. 2022; 105: 331–361. https://doi.org/10.1111/phpr.12823

Nguyen, C. Thi. 2022. “Trust as an unquestioning attitude,” *Oxford Studies in Epistemology*, 7: 214–244.

Nosek, B. A., G. Alter, G.C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, and M. Contestabile. 2015. “Promoting an Open Research Culture,” *Science* 348(6242): 1422–25.

O'Connor, Cailin and James Weatherall. 2019. *The misinformation age: How false beliefs spread.* Yale University Press. https://doi.org/10.2307/j.ctv8jp0hk

O’Neill, Onora. 2018. “Linking Trust to Trustworthiness,” *International Journal of Philosophical Studies*, 26(2): 293–300.

O’Neill, Onora. 2002. *A Question of Trust*. Cambridge: Cambridge University Press.

O’Reilly, Jessica, Naomi Oreskes, Michael Oppenheimer. 2012. “The rapid disintegration of projections: The West Antarctic Ice Sheet and the Intergovernmental Panel on Climate Change,” *Social Studies of Science*, (42)5: https://doi.org/10.1177/0306312712448130

Oreskes, Naomi, and Erik M. Conway. 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. New York: Bloomsbury Press.

Oreskes, Naomi. 2019. *Why trust science?* Princeton, NJ: Princeton University Press.

Origgi, Gloria. 2020. "Trust and Reputation," in *Routledge Handbook of Trust and Philosophy*, (ed.) Judith Simon, New York: Routledge, 88–96.

Pace, Michael. 2021. "Trusting in order to inspire trustworthiness," *Synthese* **198**, 11897–11923, https://doi.org/10.1007/s11229-020-02840-8

Parker, Wendy and Greg Lusk. 2019. "Incorporating User Values into Climate Services," Bulletin of the American Meterorological Society, (100)9:1643–1650. DOI: https://doi.org/10.1175/BAMS-D-17-0325.1

Peirce, Charles S. 1931–1966. *The collected papers of Charles S. Peirce*, vol. 8. C. Hartshorne, P. Weiss & A. W. Burks (eds.). Cambridge: Harvard University Press.

Peirce, Charles. 1982. *Writings of Charles S. Peirce*, vol. 6. M. Fisch, E. Moore & C. Kloesel (eds.). Bloomington: Indiana University Press.

Pirson, Michael and Deepak Malhotra. 2011. "Foundations of Organizational Trust: What Matters to Different Stakeholders?" *Organization Science*, (22)4: 1087–1104.

Poorting, Wouter and Nick Pidgeon. 2003. "Exploring the Dimensionality of Trust in Risk Management," *Risk Analysis*, (23)5: 961–972.

Potter, Nancy. 2002. *How Can I be Trusted? A Virtue Theory of Trustworthiness*, New York: Rowan and Littlefield.

Putnam, Robert. 2000. *Bowling Alone: The Collapse and Revival of American Community*, New York: Simon and Schuster.

PytlikZillig, Lisa, Joseph A. Hamm, Ellie Shockley, Mitchel N. Herian, Tess M.S. Neal, Christopher D. Kimbrough, Alan J. Tomkins & Brian H. Bornstein. 2016. "The dimensionality of trust-relevant constructs in four institutional domains: results from confirmatory factor analyses," *Journal of Trust Research*, (6)2: 111–150.

PytlikZillig, Lisa, and Christopher Kimbrough. 2016. "Consensus on conceptualizations and definitions of trust: Are we there yet?" In *Interdisciplinary perspectives on trust: Towards theoretical and methodological integration*, (eds.) E. Shockley, T. M.S. Neal, L. M. PytlikZillig, & B. H. Bornstein (Eds.), New York, NY: Springer, 17–47.

Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Erik Topol. 2022. "AI in health and medicine," *Nature Medicine*, 28: 31–38.

Rolin, Kristina. 2015. "Values in Science: The Case of Scientific Collaboration," *Philosophy of Science*, (82)2: 157–177.

Rooney, Phyllis. 2017. "The Borderlands between Epistemic and Non-Epistemic Values." In ed. K. Elliott and D. Steel, *Current Controversies in Values and Science*, 31–45. New York: Routledge.

Rooney, Phyllis. 1992. “On Values in Science: Is the Epistemic/Non-epistemic Distinction Useful?” In *PSA 1992: Proceedings of the 1992 Biennial Meeting of the Philosophy of Science Association,* vol. 1, ed. David Hull, Micky Forbes, and Kathleen Okruhlik, 13–22. East Lansing, MI: Philosophy of Science Association.

Rousseau, Denise, Sim Sitkin, Ronald Burt, and Colin Camerer. 1998. “Introduction to Special Topic Forum: Not so Different after All: A Cross-Discipline View of Trust,” *The Academy of Management Review*, (23)3: 393–404.

Royal Society. 2012. *Science as an Open Enterprise*. London: The Royal Society.

Rotter, Julian. 1967. “A new scale for measurement of interpersonal trust,” *Journal of Personality*, 35(4): 651–665.

Rotter, Julian. 1971. “Generalized expectancies for interpersonal trust,” *American Psychologist*, 26, 443–452.

Rotter, Julian. 1980. “Interpersonal trust, trustworthiness, and gullibility,” *American Psychologist,* 35, 1–7.

Rudner, Richard. 1953. “The Scientist qua Scientist Makes Value Judgments.” *Philosophy of Science*, 20(1): 1–6.

Russell, Stuart and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach*, Fourth Edition.

Ryan, Mark. 2020. “In AI We Trust: Ethics, Artificial Intelligence, and Reliability,” *Science and Engineering Ethics*, 26(5): 2749–67. https://doi.org/10.1007/s11948-020-00228-y

Schmidhuber, Jürgen. 2015. “Deep leraning in neural networks: An overview,” *Neural Networks*, 61: 85–117.

Schoorman, David, Roger Mayer, and James Davis. 2007. “An Integrative Model of Organizational Trust: Past, Present, Future,” *The Academy of Management Review*, (32)2: 344–354.

Schroeder, S. Andrew. 2022a. “An Ethical Framework for Presenting Scientific Results to Policy-Makers.” *Kennedy Institute of Ethics Journal*, 32: 33–67.

Schroeder, S. Andrew. 2021. “Democratic Values: A Better Foundation for Public Trust in Science,” *British Journal for the Philosophy of Science*, 72:511–43. http://doi.org/10.1093/bjps/axz023

Schroeder, S. Andrew. 2022b. “The Limits of Democratizing Science: When Scientists Should Ignore the Publc,” *Philosophy of Science*, 89: 1034–1043.

Schroeder, S. Andrew. 2020. “Thinking about Values in Science: Ethical vs. Political Approaches.” *Canadian Journal of Philosophy*. http://doi.org/10.1017/can.2020.41

Schwartz, Shalom and Wolfgang Bilsky. 1987. ‘Toward a Universal Psychological Structure of Human Values,” *Journal of Personality and Social Psychology* 53: 550–62.

Scriven, Michael. 1974. “The Exact Role of Value Judgements in Science,” in *PSA 1972*, (eds.) K. Schaffner and R. Cohen, Dordrecht: Reidel, 219–47.

Seligman, Adam. 1997. *The Problem of Trust*. Princeton: Princeton University Press.

Siegrist, Michael. 2021. "Trust and Risk Perception: A Critical Review of the Literature," *Risk Analysis*, 41:3, 480–90.

Siegrist, Michael, George Cvetkovich, and Claudia Roth. 2000. "Salient value similarity, social trust, and risk/benefit perception," *Risk Analysis*, 20(3), 353–362.

Siegrist, Michael, Melanie Connor, and Carmen Keller. 2012. "Trust, confidence, procedural fairness, outcome fairness, moral conviction, and the acceptance of GM field experiments," *Risk Analysis*, 32, 1394–1403.

Siegrist, Michael and Alexandra Zingg. 2014. "The Role of Public Trust During Pandemics: Implications for Crisis Communication," *European Psychologist,* 19:1, 23–32.

Simion, Mona and Christoph Kelp. 2023. "Trustworthy artificial intelligence," *Asian Journal of Philosophy,* (2) 8 https://doi.org/10.1007/s44204-023-00063-5

Simpson, Thomas. 2023. "Faith as Trust," *The Monist*, 106: 83–93. https://doi.org/10.1093/monist/onac025

Simpson, Thomas. 2023. *Trust: A Philosophical Study*, Oxford: Oxford University Press.

Simpson, Thomas. 2018. "Trust, Belief, and the Second-Personal," *Australasian Journal of Philosophy*, 96:3, 447–459, DOI: 10.1080/00048402.2017.1382545

Simpson, Thomas. 2017. "Trust and Evidence," in *The Philosophy of Trust*, ed. Paul Faulkner and Thomas Simpson, Oxford: Oxford University Press: 177–94.

Simpson, Thomas. 2012. "What is Trust?" *Pacific Philosophical Quarterly,* 93(4): 550–69.

Smith, Jordan, Jessica Leahy, Dorothy Anderson, and Mae Davenport. 2013. "Community/Agency Trust and Public Involvement in Resource Planning," *Society and Natural Resources*, (26)4: 452–471.

Stack, L. C. 1988. "Trust," In *Dimensionality of personality,* (eds.) H. London & J. E. Exner, Jr. (Eds.), New York: Wiley, 561–599.

Slovic, Paul. 1993. "Perceived risk, trust, and democracy," *Risk Analysis*, 13(6): 675–682.

Smith, Matthew. 2010. "Reliance," *Noûs* 44, 135–57.

Sparkes, Matthew. 2023. "Game-playing DeepMind AI can beat top humans at chess, Go and poker," *New Scientist*, URL: https://www.newscientist.com/article/2402645-game-playing-deepmind-ai-can-beat-top-humans-at-chess-go-and-poker/.

Staley, Kent. 2017. "Decisions, Decisions: Inductive Risk and the Higgs Boson," In eds. K. C. Elliott and T. Richards, *Exploring Inductive Risk: Case Studies of Values in Science*, 37–55. New York: Oxford University Press.

SteelFisher, G. K., Findling, M. G., Caporello, H. L., McGowan, E., Espino, L., & Sutton, J. (2023). "Divergent Attitudes Toward COVID-19 Vaccine vs Influenza Vaccine," *JAMA network open*, 6(12), e2349881. https://doi.org/10.1001/ jamanetworkopen.2023.49881

Steinhardt, H. Christopher. 2012. "How is High Trust in China Possible? Comparing the Origins of Generalized Trust in Three Chinese Societies," *Political Studies*, (60)2, DOI: https://doi.org/10.1111/j.1467-9248.2011.00909.x

Stewart, Elizabeth. 2024. "Negotiating domains of trust," *Philosophical Psychology*, 37:1, 62–86, DOI: 10.1080/09515089.2022.2144190

Strawson, Peter. 1974. "Freedom and Resentment," in *Freedom and Resentment and Other Essays*. London: Methuen, 1–25.

Sztompka, Piotr. 1999. *Trust: A Sociological Theory*. Cambridge: Cambridge University Press.

Taddeo, Mariarosaria. 2009. "Defining Trust and E-Trust: From Old Theories to New Problems," *International Journal of Technology and Human Interaction*, DOI: 10.4018/jthi.2009040102.

Taddeo, Mariarosaria and Luciano Floridi. 2011. "The case for e-trust," *Ethics and Information Technology*, 13: 1–3.

Tamasello, Michael, Alicia Melis, Claudio Tennie, Emily Wymann, and Esther Herman. 2012. "Two Key Steps in the Evolution of Human Cooperation: The Interdependence Hypothesis," *Current Anthropology*, (53)6: 673–692.

Tardy, Charles. 1988. "Interpersonal evaluations: Measuring attraction and trust," In *A handbook for the study of human communication*, (ed.) C. H. Tardy, Norwood, NJ: Ablex Publishing., 269-283.

Thompson, Christopher. 2017. "Trust without Reliance," *Ethical Theory and Moral Prac*tice, **20**: 643–655. https://doi.org/10.1007/s10677-017-9812-3

Trumbo, Craig, and Katherine McComas. 2003. "The function of credibility in information processing for risk perception," *Risk Analysis,* (23)2: 343–53.

Tsujikawa, Norifumi, Shoji Tsuchida, Takamasa Shiotani. 2016. "Changes in the Factors Influencing Public Accpetance of Nuclear Power Generation in Japan Since the 2011 Fukushima Daiichi Nuclear Disaster," *Risk Analysis*, (36)1: 98–113.

Tummeltshammer, Kristen, Rachel Wu, David Sobel, and Natasha Kirkham. 2014. "Infants Track the Reliability of Potential Informants," *Psychological Science*, (25)9: 730–1738.

Tutić, Andreas and Thomas Voss. 2020. "Trust and Game Theory," in *Routledge Handbook of Trust and Philosophy*, (ed.) Judith Simon, New York: Routledge, 175–188.

Queloz, Matthieu. 2021. *The Practical Origins of Ideas: Genealogy as Conceptual Reverse-Engineering*. Oxford: Oxford University Press.

Vallier, Kevin and Michael Weber. 2021. *Social Trust.* New York: Routledge.

Van De Walle, Steven and Frédérique Six. (2013). "Trust and distrust as distinct concepts: Why studying distrust in institutions is important," *Journal of Comparative Policy Analysis: Research and Practice,* 1–17. doi:10.1080/13876988.2013.785146

Ward, Zina. 2021. "On Value-Laden Science." *Studies in History and Philosophy of Science*, 85: 54–62.

Wilholt, Torsten. 2016. “Collaborative Research, Scientific Communities, and the Social Diffusion of Trustworthiness,” in *The Epistemic Life of Groups: Essays in the Epistemology of Collectives*, (eds.) Michael Brady and Miranda Fricker, 218–234.

Wilholt, Torsten. 2013. “Epistemic Trust in Science.” *British Journal for Philosophy of Science,* 64:233–53.

Williams, Bernard. 2002. *Truth and Truthfulness*. Princeton: Princeton University Press.

Williamson, Oliver. 1993. “Calculativeness, trust and economic organization,” *Journal of Law and Economics,* 30: 131–145.

Wintterlin Florian, Friederike Hendriks, Niels Mede, Rainer Bromme, Julia Metag, and Mike Schafer. 2022. “Predicting public trust in science: The role of basic orientations toward science, perceived trustworthiness of scientists, and experiences with science,” *Frontiers in Communication* 6: 822757.

Wood, B.J. and R.A. Duggan. 2000. “Red Teaming of advanced information assurance concepts,” *Proceedings of DARPA Information Survivability Conference and Exposition*, 2: 112–118.

Wylie, Alison, and Lynn Hankinson Nelson. 2007. “Coming to terms with the values of science: Insights from feminist science studies scholarship,” in *Value-Free Science? Ideals and Illusions*, (eds.) H. Kindcaid, J. Dupré, & A. Wylie, Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195308969.003.0005,

Yamagishi, Toshio, Karen Cook, and Motoki Watabe. 1998. “Uncertainty, trust, and commitment formation in the United States and Japan,” *American Journal of Sociology*, 104: 165–194.