Spring 2022

Alexa, Should I Trust You? A Theory of Trustworthiness for Artificial Intelligence

Elizabeth K. Stewart

Alexa, Should I Trust You?
A Theory of Trustworthiness for Artificial Intelligence

by

Elizabeth K. Stewart

Master of Science
University of Edinburgh 2014

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Philosophy

College of Arts and Sciences

University of South Carolina

2022

Accepted by:

Brett Sherman, Major Professor

Anne Bezuidenhout, Committee Member

Michael Dickson, Committee Member

Pooyan Jamshidi, Committee Member

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

Dedication

In loving memory of my grandmother, Helen Stewart, who wanted to study theology, but had to settle for something more "practical".
Her unwavering love of knowledge paved the way for me to pursue the impractical.

Acknowledgments

Getting a PhD is no joke and I couldn't have completed this dissertation without the support and input of so many people. My advisor, Brett Sherman, not only read this work throughout its many iterations and offered invaluable comments, criticisms, and thoughts, but also recognized the value of the project when I doubted it myself. I am incredibly grateful for all the feedback, the conversations, and the encouragement. I am indebted also to the other members of my committee: Anne Bezuidenhout, who has supported my career in philosophy from the day I visited as a prospective student to the day I defended my dissertation, Michael Dickson, who agreed to teach me about complexity theory and has been stuck with me ever since, and Pooyan Jamshidi, who let an under-qualified philosophy student join his AI class and who is always interested to hear a philosopher's perspective on computer science.

The Russel J. and Dorothy S. Bilinski Fellowship allowed me to spend my final year dedicated entirely to researching and writing this dissertation. I am deeply thankful for the financial support, without which, finishing this work would have been incredibly difficult.

I also owe a debt of gratitude to the intellectual community of the philosophy department, in particular my fellow graduate students, Marco, Apurva, Mike, Emily, Dustin, Justin, Kelly, Nic, Brady, Brittany and all the rest, who sat in classes alongside me, proofread my work, challenged my ideas, and overall, helped me become a better thinker and more compassionate human.

There are others without whom I would not have embarked on or completed this journey. My family, who instilled in me a love for learning that has never faded.

Glenn Gentry, who was the first person who told me I should study philosophy and to whom I am grateful for many philosophy conversations had around campfires over the years since. I would also not have maintained the confidence to keep working on this project without the enduring support of my partner, Dan, who made sure I ate food, slept, and had a social life, all of which, apparently, are important for functioning humans. There are countless others, friends and colleagues, who have made the past six years, and subsequently, this dissertation, possible. Thank you.

Abstract

As people turn to AI driven technologies for help with everything from meal planning to choosing a mate, it is increasingly important for individuals to gauge the trustworthiness of available technologies. However, most philosophical theories of trustworthiness focus on interpersonal trust and are inappropriate for non-agents. What, then, does it mean for non-agents such as AI driven technologies to be trustworthy? I distinguish two different forms of trustworthiness: naive trustworthiness and robust trustworthiness. An agent is naively trustworthy to the extent that it would be likely to meet the truster's expectations with respect to a given domain. An agent is robustly trustworthy to the extent that it would be likely to meet the truster's needs with respect to a given domain. I argue that it is possible for AI driven technologies to be both naively and robustly trustworthy, but this trustworthiness is not a stable feature of trustees, but is relative to a trusters' expectations and vulnerabilities.

In chapter one, I argue that current accounts of trustworthy AI obscure the dynamics of trust relationships that are important for individual decision-making. I then argue that unquestioning trust characterizes many of the risks associated with human-AI relationships and sets the bar for trustworthiness at the appropriate level.

In chapter two, I argue that vulnerability both precedes and follows from trust relationships. I demonstrate how these different sources of vulnerability create a tension for people seeking to be trustworthy: sometimes, the actions that mitigate the vulnerabilities that follow from trust reinforce the vulnerabilities that precede trust. Does the trustworthy agent do what they have been trusted to do, even if that

reinforces harmful vulnerabilities that motivated the trust? Or does the trustworthy agent break trust when that trust is ill-conceived? In addressing this tension, I distinguish two kinds of trustworthiness. *Naive trustworthiness* requires an agent to act as entrusted, regardless of the context or consequences. *Robust trustworthiness* requires an agent to act so as to minimize vulnerabilities that both precede and follow from trust relationships when those vulnerabilities are harmful.

In chapter three, I argue that robust trustworthiness is not a stable feature, but is sensitive to the particular trust-context. When people trust, that trust is limited to a particular domain, determined by the truster's expectations of the trustee. Sometimes, however, a truster's expectations are misplaced or too vague and meeting them may harm the truster. In cases like this, the robustly trustworthy agent may break trust in order to avoid such harm. However, it is not an easy matter to determine when the expectations comprising a trust domain are misplaced or otherwise inappropriate. In cases where the truster and trustee disagree regarding what the appropriate expectations are and how they should be met, I argue that it is not necessarily the case that either is making a mistake. I call these cases of "faultless broken trust". When faultless broken trust occurs, the truster should not continue trusting the trustee. The robustly trustworthy agent, then, is trustworthy in contexts where trust breaking is not faultless.

In chapter four, I demonstrate how the concepts of naive trustworthiness and robust trustworthiness apply to human relationships with AI-infused technologies. I argue that black-box AI technologies pose a particular problem for naive trustworthiness. I then argue that robustly trustworthy AI must be aimed at legitimate needs and must not require that people adapt to the limitations of AI technologies in harmful ways.

Table of Contents

Chapter 1

Introduction

In 2018, Gizmodo reporter Kashmir Hill converted her one-bedroom apartment to a "smart home". "Smart homes" use wi-fi enabled devices to automate processes like temperature regulation, lighting, and cleaning. These devices are marketed as streamlining their owner's lives through taking care of processes that owners would otherwise have to do themselves. Hill integrated as many wi-fi devices as possible into her daily routine, including a smart speaker, lights, coffee maker, baby monitor, kid's toys, vacuum, TV, toothbrush, a photo frame, sex toy and bed (Mattu and Hill, 2018). The goal of this transformation was to discover what kind of data these devices collected and shared, and who it was shared with. It turned out that a lot of data was collected, including what episodes of their favorite shows they were watching, when they were listening to music, a map of their house, and video recordings from inside their home. Some of this information was stored remotely and then deleted, but other information was sold to third-party companies which buy data for marketing purposes. However, over the course the two-month experiment, the author found that, in addition to privacy concerns, integrating these devices into her home was *annoying*. She had to download 14 different apps to control everything. The different devices weren't all compatible with each other, her smart speaker didn't always understand her commands and she was bombarded with notifications from all her various devices. At the conclusion of her experiment, she reported that "I thought the house would take care of me but instead everything in it now had the power to ask me to do things. Ultimately, I'm not going to warn you against making everything in your

home smart because of the privacy risks, although there are quite a few. I'm going to warn you against a smart home because living in it is annoying as hell." (Mattu and Hill, 2018).

Hill's experience highlights the risks and opportunities that the future of Artificial Intelligence (AI) offers us. The promise of care-free home maintenance is delightful. After all, who wouldn't enjoy freedom from vacuum duties or turning off the lights with a simple verbal command? However, this promise does not come without perils, which include both practical problems getting devices to function as intended as well as more troubling issues related to user's privacy, autonomy and social well-being. As individuals, organizations, and governments increasingly integrate AI into their daily operations and decision making processes it is of utmost importance that we develop an understanding of when and why such integration is appropriate. What role should AI play in our individual lives and collective societies? That is, how much and in what ways should we trust AI? When is AI *trustworthy*? The aim of this dissertation is to provide a framework for answering this question.

Artificial intelligence is broadly understood as machine behavior that would be called intelligent if a human exhibited it (McCarthy et al., 1955). Thus understood, artificial intelligence encompasses a wide variety of technologies that display behaviors ranging from more to less stereotypically human. Consider, for example, how AI is increasingly replacing other humans in social relationships, including companions and marriage partners. In 2017, over 1 million people asked Amazon's personal assistant Alexa if she would marry them (Leskin, 2018). While most of these proposals were probably offered in jest, there are individuals who seriously seek committed relationships with digital or robotic entities. Akihiko Kondo made news in 2018 after deciding to marry a hologram of a cyber celebrity, reporting that the hologram "kept me company and made me feel like I could regain control over my life" (Jozuka, 2018). Kondo is not alone in pursuing romance or friendship with an AI implemented as a

hologram or robot, as the increase in "digisexuality" demonstrates. Neil McArthur and Markie L.C. Twist introduced the term "digisexuality" to describe people whose primary sexual identity comes through the use of technology, such as VR or sexbots. Digisexuality, McArthur and Twist argue, could provide a helpful alternative for individuals who struggle to trust other humans as partners due to previous abuse. Additionally, they argue that AI may be helpful to those who cannot find partners that share an interest in sexual behavior that is considered taboo in their culture, such as BDSM or multi-partner sexual behavior (McArthur and Twist, 2017).

AI is not only an alternative for a human intimate partner. Tech companies have also developed a raft of options for robots that act as companions. ELLIQ, from Intuition Robotics, is just one option developed specifically for older adults. ELLIQ "helps you get things done and keeps you engaged" (Intuition Robotics, 2017). She offers reminders regarding daily medications and appointments, arranges rides, receives calls and texts, and can share information relevant to well-being with distant loved ones. She also encourages individuals to engage in physical activities and suggests personalized digital content. Users reported that they felt that ELLIQ was "like a friend or a person who was actually there" and that "if I'm feeling blue, she can pick me right up" (Intuition Robotics, 2019).

However, not all AI is designed to interact in human-like ways with other people. Indeed, artificial intelligence more often shows up in less obviously human forms. Unlike C-3PO and other science fiction portrayals of personal assistants, real life artificial assistants accompany us on our cell phones, laptops and other portable devices. They help look after us from dawn till dusk. They wake us up and offer reminders of the day's schedule. When we head to work for the day, they turn the lights out behind us, track the temperature of our house and alert us if anyone enters while we are away. At work, they can look up information for us, make appointments and send automated messages for us. As we leave, they play music for us on our

drive home and offer restaurant recommendations for dinner. Such examples of AI, while not themselves particularly human-like, do things that were previously, and often still are, done by humans, such as keeping track of a schedule or offering film recommendations.

AI is also used, however, to do things that humans are unable to do, primarily processing large amounts of data in order to identify subtle patterns that can be used to generate new information. In fields as diverse as chemical engineering, the financial sector, and criminal justice, AI is revolutionizing the way that information is collected, processed and used. In the sciences, researchers use machine learning to predict the results of empirical experiments, thus decreasing the time and financial resources used in experimentation. If AI predicts that the experiment will not provide the desired result, researchers can avoid spending the resources to conduct that experiment. This is especially useful in fields such as chemical engineering, where AI can predict which environments are suitable for certain kinds of reactions and how different chemicals will behave in different circumstances. Whereas previously, researchers had to spend time in a lab testing, making observations and manually sorting through the generated data, they can now limit their empirical work to experiments that AI has predicted will have desired results. This vastly decreases the time and resources required. Chemical engineering researchers Spelling and Glotzer note "As computers get faster, researchers - not hardware or algorithms - become the bottleneck for scientific discovery" (Spellings and Glotzer, 2018). The banking industry also benefits from AI's ability to sort through massive amounts of data quickly. Whether predicting credit-risk, the success and failure of certain businesses, or detecting fraud, AI informs how people conduct their financial affairs. Additional uses of machine learning in criminal justice include using facial recognition technology to identify suspicious persons and security threats in public spaces and machine learning for determining whether a person is eligible for parole. What we need in the face of

the diversity of AI applications, is a general account of trustworthiness that can guide us in evaluating particular applications of AI.

The matters of trust in AI and trustworthy AI have recently gained a lot of attention in both philosophy and in computer science, but it is often unclear what is meant by "trusting AI" or what it would take for AI to be worthy of this trust. Some philosophers deny that trust is even the kind of thing that people can have toward machines, arguing that trust is necessarily directed toward other humans (Holton, 1994; Hawley, 2014; Jones, 1996; Pitt, 2010; Nickel et al., 2010). If this is correct, then the entire project of understanding when AI is trustworthy is wrongheaded. Thus, in this chapter, I will review several influential ways in which trust is understood and argue that, in at least some senses of the term, trust is something that can be, and often is, directed toward AI. In section one, I will review the philosophical literature on various kinds of trust. In section two, I will argue that these trust concepts can be sensibly applied to talk about the relationship between humans and AI. In the following chapters, I will describe what I think it takes to be a worthy recipient of this trust.

1.1 What is Trust?

People use the term "trust" to refer to a wide variety of concepts. We talk about trusting intimate relations, like family and close friends. We also talk about trusting the government or trusting that the sun will rise. Surely we do not mean the same thing when we assert that we trust our parents as when we assert that we trust that our car will start. If trustworthiness indicates when trust is well placed, what trustworthiness consists of will look very different depending on which concept of trust is used.

Consider concepts such as "love" or "faith", which similarly have a variety of meanings and which similarly can be well or ill placed. If someone loves a pizza in

the same way that they love their spouse, one or the other of these loves is not well placed. In order to know which, we would have to know what "love" means. If it means something like "lifelong commitment to affirming the other", then it is not clear that pizza is a worthy object of such an attitude. Likewise, a spouse should not be the object of "love" when it means something like "satisfies one's literal appetite". In order to understand the conditions under which trust is well placed, we should begin with an analysis of what trust might mean. As will become evident in what follows, there is a great deal of diversity in what people take "trust" to mean. Given this diversity, there is an important question as to what, if anything, these concepts share that make them concepts of "trust", rather than something else. While I am not interested in arguing that one or another concept of trust is the "right" concept, I will note that generally trust is understood as involving reliance and making oneself vulnerable to the actions or inactions of another. Instead of arguing that one of these trust concepts is the right one, I simply wish to analyze how they differ in order to better understand the different ways in which we interact with other agents, human and non-human, and thereby make sense of how the demands of trustworthiness vary accordingly.

As the aim of this dissertation is to develop an account of trustworthiness applicable to artificial intelligence, I will pay particular attention to trust concepts that can meaningfully take AI as an object of trust. Thus, there are several accounts of trust that I will not consider. In particular, I will not consider accounts that view trust as a one-place or two-place predicate. According to one-place accounts, A trusts if A holds a particular kind of attitude and according to two-place accounts, A trusts B if that attitude is directed toward B. In both cases, this attitude is not limited to a particular domain (Faulkner, 2015; Domenicucci and Holton, 2017). Instead I will focus on accounts of trust according to which trust attitudes are domain limited. According to three-place accounts of trust a truster A trusts a trustee B with respect

to some domain x. I focus on three-place accounts of trust because when we trust AI, whatever that means, we generally trust it with respect to some domain. For example, we might trust AI to meet some social need, to predict the likelihood of some event or to discover some new chemical property.

Even with limiting the discussion to three-place accounts of trust, there are still a large number of accounts of trust. In order to better frame the discussion, I will introduce some distinctions made by Ivana Markovà, Per Lineel, and Alex Gillespie (Markovà et al., 2008). They characterize different forms of trust as ranging across two axes. Along one axis, trust concepts range from taken-for-granted or unquestioned trust to reflective and calculating trust. This dimension tracks the degree to which trust involves conscious awareness, rational decision-making, or the calculation of costs and benefits. Along the other, trust concepts range from micro-social relations between individuals to macro-social relations involving groups, institutions or society as a whole. This second dimension tracks the scale of the relationship, moving from trust in another individual to trust in increasingly large groups. I am particularly interested in the kinds of trust that an individual might have toward a particular other, and so will focus only on micro-social forms of trust that range from unreflective to reflective.

There is, however, a third dimension according to which we can characterize different accounts of trust: the degree to which trust involves belief. On doxastic accounts of trust, trust just is a belief. On non-doxastic accounts, however, trust can be chosen for pragmatic reasons. In order to understand this dimension, I will briefly take a detour into the literature on reliance.

Discussions of the nature of trust often begin by distinguishing it from other similar concepts, in particular, the concept of reliance. While trust and reliance are frequently used interchangeably in colloquial use, many philosophers find that distinguishing one from the other is a useful starting point in developing an account

of trust. Following Baier (1986), I take the kind of trust I am interested in to be a species of reliance. Facundo Alonso presents an account of reliance in which reliance is a mental state aimed at "providing cognitive guidance that is sensible or correct from the standpoint of relevant ends, values and so on - where these include, among others, prudential, moral and intellectual ends and values" (Alonso, 2014, pg. 169).

Let us imagine a manager in a company tasked by their CEO with a large project. That manager may worry about getting the project finished within the deadline and needs to figure out what to do in order to complete it. When they rely on one employee to complete one part of the project, this reliance allows them to plan future actions accordingly. They will not, for example, do this part of the project themselves or ask a different employee to complete that part. In this way, reliance guides actions.

For Alonso, the distinctive function of reliance is providing sensible guidance. This function is connected to the way this attitude is constitutively regulated. That is, the considerations that regulate what things are relied upon depends on its function. The considerations that Alonso takes central to regulating reliance are pragmatic. Furthermore, there is a norm accepted, to the exclusion of other competing norms, associated with the action guiding function of reliance, namely that reliance guide actions sensibly.

Alonso sketches out different types of sensible guidance that reliance provides. These include prudentially sensible guidance, in which reliance depends on contingent features of the context. In the above manager example, the manager might rely on the consideration that the employee with the smallest workload at that moment in time will have the time and energy required to help with the project. Reliance can also provide morally sensible guidance, in which case reliance depends on moral considerations. Perhaps the manager made a promise to an employee to involve them with the next big project. The manager might rely on honoring that promise in making the decision regarding who to involve in the big project.

Note that in the scenario where the manager relies on the promise in choosing the employee, this need not require that the manager also believe that said employee will actually be helpful. Reliance does not, on Alonso's account, need to involve accurate representations of the world. Alonso provides a similar example in which a manager knowingly relies on an employee who she knows will not finish the project in time. Yet she chooses to rely on this employee because it might help him gain confidence and consequently improve future contributions to the company. While she knows that he will not finish in time, she acts as though he will. This latter fact is important because if she knew he would not finish in time and acted as though he wouldn't, perhaps having a back up plan or someone else secretly doing the project as well just in case, then the boss would not properly be relying on the employee because her actions are not guided by the reliance. While reliance need not involve accurate representations of the world, it may, and it may also involve unknowns. Alonso provides the example of someone stuck on a mountain cliff. The only way down is a questionable rope, which they do not know will hold them. Yet, it is the only way down and so the individual relies on it, and plans his actions accordingly, slowly easing his way over the edge.

Thus reliance can involve situations where we expect that what we are relying on will fail us, situations where we simply don't know the outcome of the reliance, and situations where we expect a positive outcome from reliance. The dimension of doxasticity thus helps us distinguish between the kinds of trust that arise in these various situations. At one end of the spectrum are accounts of trust that require trusters to actually believe that the trustee will behave in a particular way, on the other are accounts of trust according to which trusters may choose to trust for pragmatic reasons, even when they are uncertain of a positive outcome. We can represent the dimensions of unreflective-reflective and doxastic-non-doxastic graphically (See Figure 1.1).

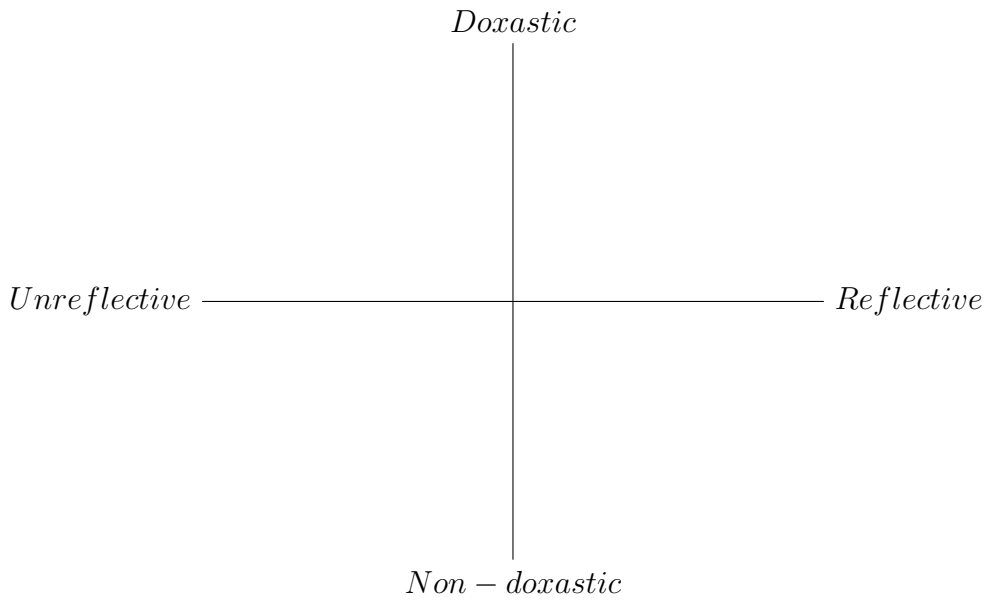

Figure 1.1 Axes of Trust

This is not not the only way to taxonomize different accounts of trust. As noted earlier, Markovà et al. include a dimension of micro- and macro- social trust. Karen Jones also offers an alternative taxonomy which characterizes trust according to two different axes (Jones, 2015). Along one axis, trust is plotted according to the attitude or stance that trust is taken to be (actionlike, belieflike, affective attitude-like). Along the other, trust is plotted according to the motivation that trust tacitly ascribes to the trustee. Loose accounts of trust may not ascribe any particular motivation to the trustee, while others may ascribe motivations like goodwill or moral integrity. I have chosen to focus on the reflectiveness and doxasticity because I believe these dimensions of trust are particularly relevant to human-AI trust relationships, a thought I will return to later in this chapter.

1.1.1 Non-Doxastic, Reflective Trust

It may, at first, seem strange to think of trusting someone despite remaining doubtful that they will meet our trust. Indeed, on most accounts, trust requires the belief

that the trusted person will act as expected. However, Karen Frost-Arnold argues that sometimes we trust others for practical reasons (Frost-Arnold, 2014). On her account, trust is a rational cognitive attitude that involves "taking the proposition that someone will do something as a premise in one's practical reasoning". Importantly, this need not also involve actually believing that proposition to be true. Instead, we may choose to trust for strategic or practical reasons, even when we believe our trust will not be met. She argues that "One can choose to engage in therapeutic trust to inspire trustworthiness, coping trust to simplify one's planning, or corrective trust to avoid doing a testimonial injustice" (Frost-Arnold, 2014, p. 1957).

For example, a person may choose to trust another not because they believe that the other will act as expected, but in hopes that the trust will inspire some positive change in the trustee. Karen Jones describes this as therapeutic trust, or "trust undertaken with the aim of bringing about trustworthiness" (Jones, 2004, p. 5). In her discussion of therapeutic trust, Jones distinguishes expectations of actual behavior from normative expectations regarding hoped-for behavior. Jones argues that a parent may trust their child to take care of the house while they are away, in hopes that this trust will bear fruit in the long run while not believing that their child will actually be responsible in their absence.

I will not dwell long on strategically chosen trust because it is quite similar to Alonso's account of reliance. Both involve taking some proposition as fixed and acting accordingly, despite one's beliefs or doubts. I include it here as the foil against which I will compare more demanding accounts of trust.

1.1.2 Reflective, Doxastic Trust

Many accounts of trust involve holding some set of beliefs about the trustee. What differs among these accounts is the content of these beliefs and the degree to which these beliefs are reflected upon. In this section, I will focus on accounts in which trust

is the result of some kind of deliberation. This is notably evident in Russell Hardin's encapsulated interest account of trust.

According to Hardin, we trust when we believe the other encapsulates our interests. For Hardin, if I trust you, that means that I expect that you will meet my trust and that expectation is "grounded in an understanding (perhaps mistaken) of your interests specifically with respect to me" (Hardin, 2001, p. 3) He notes several reasons why such a belief or expectation might be justified, which include a trustee's interest in maintaining a relationship or a reputation, an interest in the truster's well-being, and an interest in keeping moral commitments. For example, when I purchase something online, I trust the online seller to send me what I have purchased because they have an interest in meeting my interests because they may wish to maintain my loyalty as a customer or to uphold their good reputation. Thus, their business interests encapsulate my consumer interests.

Unlike Frost-Arnold, for Hardin, trusting is not a matter of deciding to act: "Commonplace claims that one *chooses* to trust mistakenly imply that trusting is a matter of acting.... All of this is wrong. Usually I just do or do not trust to some degree, depending on the evidence I have" (Hardin, 2001, p. 11-12). Instead, trust is a kind of belief that "commonly grows out of an ongoing exchange or iterated prisoners' dilemma interaction" (Hardin, 2001, p. 18). Importantly, these beliefs are grounded in evidence of the trustee's trustworthiness. It is through reflection on the evidence that we come to trust others.

Hardin accepts a broad range of beliefs as grounds for trust. Others, however, ground trust in beliefs about particular characteristics of the trustee. Annette Baier, whose account of trust has been hugely influential in the philosophical literature on trust, is a good example of this. For Baier, trust is grounded on beliefs about the trustee's goodwill. On this account, I trust when I depend on another person's goodwill towards me. In doing so, I make myself vulnerable "to another's possible

but not expected ill will (or lack of goodwill)" (Baier, 1986, p. 235). Like Hardin, Baier does not think that trust is the result of an act of will. It can, however, be encouraged through experiences that provide reasons for confidence in the trustee's goodwill. It can also be undermined by evidence that the trustee does not care about the truster or what the truster has entrusted to them.

There are other similar belief-based accounts of trust wherein trust involves believing that the trustee has some particular set of characteristics, such as believing the trustee has the right intention, right reason, and competence (Ullmann-Margalit, 2004), believing that the trustee has a commitment to meet the trust (Hawley, 2014), or having evidence that the trustee will do what they've been trusted to do (Simpson, 2017).

Additionally, some belief-based accounts view trust as belief plus something extra. For example, Arnon Keren argues that trust involves beliefs about the trustee plus the idea that the reasons for these beliefs are preemptive reasons. They argue that a truster cannot trust someone to do something while taking precautions against their failing to do that thing. In order to accommodate this, they argue that the reasons for trust must be preemptive (Keren, 2014; Keren, 2020). That is, trust involves higher order reasons against believing or acting for certain other reasons.

1.1.3 Non-Doxastic, Unreflective Trust

Alternatively, some accounts of trust may involve beliefs, but these beliefs are not understood as central to trust. Karen Jones, for example, argues that trust is an affective-attitude, not primarily a belief. Keren (2014) argues that Jones' account could be characterized as a "mixed doxastic account" because of the way beliefs, while not central to trust, still play a role in trust relationships. Jones rejects the idea that beliefs are central to trust because she rejects the idea that trust is justified only if the beliefs that constitute that trust are justified. Additionally, she prefers

the idea that trust is an affective attitude because it avoids the issue of whether we have enough evidence to justify our trust. While beliefs are not central to Jone's account, she does note that trust, as an affective attitude, does often give rise to beliefs. This happens because trust restricts which interpretations of another person's behavior we consider. To trust another, according to Jones, is "to have an attitude of optimism about her goodwill and to have the confident expectation that, when the need arises, the one trusted will be directly and favorable moved by the thought that you are counting on her" (Jones, 1996, p. 5-6). The latter feature is what Jones calls "trust-responsiveness", which is the disposition to be moved by the fact of another's trust. When we have such a positive attitude about another, we are less likely to believe certain things about them. As Jones state: "Trusting thus functions analogously to blinkered vision: it shields from view a whole range of interpretations about the motives of another and restricts the inferences we will make about the likely actions of another. Trusting thus opens one up to harm, for it gives rise to selective interpretation, which means that one can be fooled, that the truth might lie, as it were, outside one's gaze.... The harms they might cause through failure of goodwill are not in view because the possibility that their will is other than good is not in view" (Jones, 1996, p. 12). This idea of trust acting as a kind of blinker places Jones' account of trust closer to the *unquestioning* end of the unquestioning/reflective spectrum.

Thi Nguyen presents an account of trust that similarly views trust as an attitude that places the thought of untrustworthiness out of view. It is non-doxastic insofar as trust involves putting something outside the space of evaluation and deliberation; "to rely on it without pausing to think about whether it will actually come through for you" (Nguyen, 2019, p. 2). He argues that this does not mean that trust cannot involve seeking evidence or good reasons to trust. He states, "Trust can certainly arise out of deliberation and it can certainly be called into question. But it is to

say that when one has come to trust, one has adopted an unquestioning attitude. And limited beings like us must often take up such unquestioning attitudes as part of a reasonable strategy for coping with the overwhelming cognitive onslaught of the world" (Nguyen, 2019, p. 2). While Nguyen himself does not make this explicit, I think it is compatible with his account that trust *can* be understood as doxastic in that we can adopt the unquestioning attitude toward beliefs. I can, for example, trust that the sun will rise tomorrow, which involves believing, without question, that the sun will rise tomorrow. In this way, Nguyen's account can also be seen as a "mixed doxastic" account of trust.

However, his account differs because from Jones' in that he does not view trust as something that need be directed at other agents. While Jones argues that the attitude of trust involves optimism about another's goodwill and responsiveness to trust, Nguyen argues that we can adopt an unquestioning attitude toward objects and artefacts, as well as other people. For Nguyen, trust has to do with agency. In particular, trust allows us to "integrate other people and objects into our own functioning" (Nguyen, 2019, p. 4).

Nguyen also uses the example of a rock climbing rope in his discussion of trust, which ties in well with Alonso's rope example to demonstrate the difference between mere reliance and trust. In Alonso's example, a person relies on the rope when they incorporate the rope holding them into their plans, regardless of whether they actually believe that it will. I might have serious doubts about the rope, but out of desperation rely on it, using it to guide my future actions, but experience tremendous anxiety the entire climb down the cliff. This is mere reliance. In Nguyen's example, he discusses the difference between someone who trusts the rope and someone learning to trust the rope as follows:

> "Consider the experienced rock climber's attitude toward the rope and gear. A novice rock climber tests the rope gingerly, occasionally

> weighting it, telling themselves over and over again to trust it. This is part of the process of coming to trust - but it is only the beginning of the process. While they are engaged in this process of self-negotiation and self-reassurance, we would say that they do not yet fully trust the rope. It is the experienced rock climber who truly trusts their rope. Their trust is reflected in the fact that concerns about the rope's reliability occupy no mental space for them at all." (Nguyen, 2019, pg. 12)

Trust allows us to set aside deliberation and devote our mental energies elsewhere. Nguyen's concept of trust has a two-tiered structure. Trusting involves a first-order disposition to immediately accept that X will P and second-order disposition to resist questioning about the first-order disposition. Individuals who trust, therefore, do not deliberate about whether the trusted will come through for them. Additionally, while trust is defeasible, the reasons for defeating trust must be "significantly stronger" (pg. 15) than reasons that would defeat beliefs. Furthermore, Nguyen views trust as a spectrum concept. People can have varying degrees of trust and trust can be restricted to particular functions.

Importantly for my purposes, Nguyen argues that this unquestioning attitude can be directed at other people, as well as organizations, governments, and non-agents, such as AI and inanimate objects. Some have argued that trust, by definition, cannot be directed at AI or that in trust in AI reduces to trust in the people who develop AI. In the next section, I will review some of these arguments and argue that, if we consider Nguyen's account of trust as an unquestioning attitude, we can, indeed, trust AI and not just those who make it.

1.2 Can We Trust AI?

Can these existing accounts of trust and trustworthiness that were developed with interpersonal, human relationships in mind, translate into interactions with AI? The

general argument to the contrary, while presented slightly differently by different philosophers, runs something like:

Argument 1

Premise 1: An entity can be trusted only if the entity has some feature (or set of features) F.

Premise 2: Machines do not have feature (or features) F.

Conclusion: Therefore, machines cannot be trusted.

We have already seen the ways in which philosophers view trust as grounded in features that machines do not currently possess. These features include goodwill (Baier, 1986; Jones, 1996), trust-responsiveness (Jones, 1996), encapsulated interests (Hardin, 2001). Others have presented alternative features related to persons that form the basis of trust, such as the ability to make promises (Pitt, 2010) or have motivations (Nickel et al., 2010). The argument then proceeds by stating that machines do not have goodwill, interests that can encapsulate our own, the ability to make promises, etc. Thus, people cannot trust machines.

For example, consider what Karen Jones says about trust and machines:

> "Trusting is not an attitude that we can adopt toward machinery. I can rely on my computer not to destroy important documents or on my old car to get me from A to B, but my old car is reliable rather than trustworthy. One can only trust things that have wills, since only things with wills can have goodwills - although having a will is to be given a generous interpretation so as to include, for example, firms and government bodies. Machinery can be relied on, but only agents, natural or artificial, can be trusted." (Jones, 1996, p. 14)

Jones' argument and those similar to it, however, conflate several issues that are worth separating. When someone argues that AI cannot be trusted, there are several claims they might be making:

> Conceptual claim: trust, by definition, just doesn't apply to human-AI relationships.
>
> Empirical claim: trust, in fact, just isn't extended to machines.
>
> Normative claim: trust, as a characteristic of interpersonal relationships, should not be extended to machines because machines lack the relevant features (goodwill, trust-responsiveness, etc).

In the literature, these claims are often not clearly distinguished.

As another example, Richard Holton states the following:

> "When you trust someone to do something, you rely on them to do it, and you regard that reliance in a certain way: you have a readiness to feel betrayal should it be disappointed, and gratitude should it be upheld. In short, you take a stance of trust towards the person on whom you rely. It is the stance that makes the difference between reliance and trust. When the car breaks down we might be angry; but when a friend lets us down we feel betrayed." (Holton, 1994, p.67)

In this quote, Holton appears to make an empirical claim about how people interact differently with humans versus machines. This empirical claim is then used to support a claim about the definition of trust: that trust involves reactive attitudes. Hawley, in understanding trust's counterpart distrust, similarly relies on a claim about what people do to support a claim about the nature of distrust:

> "Let's understand nonreliance as a refusal to accept vulnerability, and/or a continuing attempt to reduce such vulnerability. One might have this

> attitude to a machine one takes to be unreliable. What more is needed for distrust? Not just the absence of normative expectations, since such expectations are absent from our attitudes towards inanimate objects." (Hawley, 2014, p. 8)

Holton and Hawley both seem to be making a conceptual claim about the nature of trust – that it involves reactive attitudes, and support this claim by making an empirical claim about people, namely that they do not trust machines. However, their conceptual claim about the nature of trust is not incompatible with people actually trusting AI. It might well be be true that people, as a matter of fact, do not trust AI. However, this doesn't mean that AI is necessarily un-trustable. It could mean that AI hasn't yet demonstrated the features that would elicit trust from a human, or that a human and an AI haven't yet been in a context in which a human would be likely to extend trust. It is not particularly difficult to imagine the kind of context in which a person might respond to an AI's failure with feelings of betrayal, rather than disappointment. If, for example, it turns out to be possible to develop sentient machines that are capable of feeling emotions and taking a participant stance in relationships, then it certainly does not seem out of the question that people might trust such machines. Furthermore, even if it turns out to be impossible to create such machines, it still does not seem impossible that people can, and perhaps often do, adopt such trust attitudes toward machines.

Consider an analogy with empathy. Empathy is typically conceived of as putting oneself in another's shoes. Thus, if there is no "other", we cannot have empathy. In the case of inanimate objects, like robots, there is no other to have empathy with. However, empirical studies reveal that people do feel something like empathy toward virtual entities and robots, although perhaps not to the extent that they feel for other humans (Bartneck et al., 2005; Slater et al., 2006). Why would humans feel empathy for something that clearly does not have shoes into which they can put

themselves? Misselhorn (2009) proposes that people feel empathy with humanoid robots because they imaginatively perceive these inanimate objects as undergoing some empathy-inducing experience. They define empathy with inanimate objects as: "S empathizes with an inanimate object's imagined experience of emotion E if S imaginatively perceives the inanimate object's T-ing and this imaginative perception causes S to feel E for the inanimate object" (Misselhorn, 2009, p.352). Thus, when a person imaginatively perceives an inanimate object as undergoing some empathy-inducing event, T, then they can feel empathy toward that object. Redstone (2016) argues that what Misselhorn calls an "imaginative perception" is better described as a perceptual illusion. Perceptual illusions, unlike imaginative perceptions, persist despite knowing that the perception is not real. For example, "even if one *knows* that the robot does not experience emotions, one might, for example, still cringe upon seeing the emotive expressions a robot makes when it is being mistreated" (Redstone, 2016).

We can make sense of trust, as understood by Holton and Hawley, in AI if we suppose that those who take the required participant stance toward AI are imagining or misperceiving that AI as having something like personhood. One way we might understand trust in AI, then, is as a result of an illusion or a product of imagination. If this is right, then Misselhorn and Redstone's accounts do not only make sense of empathy toward machines, but also trust. We thus could trust AI because we perceive it as having properties, such as sentience or the capacity for emotions, which it lacks. We perceive it as the kind of thing toward which we might feel reactive attitudes, such as betrayal, although it is not. Someone might, for example, trust a robot to provide companionship and, despite knowing that it doesn't *really* understand them, it might still feel like it. Thus, Holton and Hawley might be right about their understanding of trust, but wrong about the empirical fact that people don't trust machines.

A similar argument can be made regarding other accounts of trust that require the trustee to demonstrate some feature that AI does not currently possess. A person can perceive the AI to have features like interests, goodwill, the ability to make promises, etc, even if the AI does not have those features. If someone perceives the AI as having these features, then they can extend trust to them. This does not, however, mean that this trust is well-placed.

Consider again what Jones says about trust and machines:

> "Trusting is not an attitude that we can adopt toward machinery. I can rely on my computer not to destroy important documents or on my old car to get me from A to B, but my old car is reliable rather than trustworthy. One can only trust things that have wills, since only things with wills can have goodwills - although having a will is to be given a generous interpretation so as to include, for example, firms and government bodies. Machinery can be relied on, but only agents, natural or artificial, can be trusted." (Jones, 1996, p. 14)

While Jones does not make claims about what people actually do, she does seem to conflate two different issues: (1) what is possible given the nature of trust, and (2) what people ought to do. We have already seen how it is possible that people might come to trust things that do not have wills. Thus, Jones' statement is perhaps better understood as a statement about what people *ought* to do. Perhaps people may, in fact, trust AI via imaginative perception or perceptual illusion, however, it may be the case that they should not (unless the AI has a will). Jones here seems to suggest a condition for well-placed trust: well-placed trust is directed toward agents with wills. Thus, interpreting arguments against trust in AI as making normative claims about what people *ought* to trust seems like the correct route to take.

Consequently, the question "Can we trust AI?" should be reframed as "Should we trust AI?". However, a blanket answer of "no" seems insufficient for several reasons.

Firstly, that answer is premised on the idea that AI will never have human-like capacities. While the jury is still out on whether they ever will, some researchers are optimistic that someday machines will be just like people (Asada, 2019; Bronfman et al., 2021; Shevlin, 2018). If trust is rooted in some particular capacity of the trustee, then we must pay careful attention to what a given AI is capable of. Even if AI never gains the human ability to feel emotion, they already display several characteristics that may be relevant for trust, including autonomous decision-making. If the argument against trusting AI rests on the limited abilities of machines, then, at the very least, we must accept "no" as an answer open to revision.

Secondly, it is not obvious to me why it is the case that it is wrong to trust in cases where the trustee lacks the relevant characteristic. If I imagine that a machine is capable of having goodwill toward me, and trust it to predict the results of a chemical reaction and it does so accurately, then, according to the above argument, I have trusted poorly. However, the machine has done exactly what I trusted it to do, the only problem being that I believe or feel like it has goodwill toward me when it does not. Recall that trust is frequently understood as a species of reliance and reliance has to do with guiding our actions with respect to relevant ends, values, etc. In the case of trusting the machine to predict the results of an experiment, my reliance has guided me well because it has resulted in me achieving the relevant end of an accurate prediction. My wrong belief about the machine's goodwill has no negative consequences for this end. That belief only has negative practical effects insofar as it may cause me to trust the AI to do things that it cannot do. But if I only trust it to do things that it is capable of, then my false belief about its goodwill is silly and wrong, but otherwise harmless.

However, if goodwill were necessary with respect to the relevant ends, then my faulty belief could be quite harmful. If, for example, I trust the AI to be my friend, a role for which goodwill is necessary, then my false belief is more obviously harmful.

The danger in trusting, it seems to me, is importantly related to what a person is trusting an agent to do or be. Pitt (2010) argues that the question "Can we trust technology?" is incoherent because trust is always limited to certain domains. Instead, he recommends that we ask specific questions about technology: Can we trust technology to work properly? To always improve? To always benefit humanity? To always fail? Limiting our questions regarding trustworthy AI to specific domains seems to me a better direction to take in addressing whether to trust AI than adopting a one-size-fits-all answer.

However, taking this route leads us back to the question that philosophers of trust initially raised: what distinguishes trust from reliance? If goodwill and the like are not relevant for a given end and are thus not counted as necessary for being trusted, then how should we understand what makes trust special? One answer is perhaps just to say that we only trust in contexts where goodwill and the rest are necessary. If we take this route, then it would be wrong headed to talk about trusting someone to run a calculation, provide some piece of information, or any other number of tasks that can be done without any goodwill. Instead, we could only speak rightly of trust in contexts where what is trust requires goodwill - like friendship. This seems unnecessarily restrictive.

Instead, we could adopt a view wherein goodwill, trust-responsiveness, etc. are not essential to placing trust well, except in those circumstances where goodwill is important for the relevant end. Nguyen's account of trust as an unquestioning attitude suits this need well. While the other accounts of trust are useful in understanding trust in those contexts where certain capacities and attitudes are important, Nguyen's account provides a way of thinking about trust that encapsulates these contexts as well as other contexts which do not require them. The danger in trusting, on the unquestioning account, is not that the trustee doesn't hold the right attitude toward the truster, but that the truster has outsourced their agency to something unable to

act accordingly. They have relied on something for some need without questioning whether it can meet that need. Sometimes, the failure to meet the need is related to the trustee's attitude, but it need not be. Thus, I think that viewing trust as an unquestioning attitude is often quite compatible with views of trust for which a certain attitude, like goodwill, is necessary. What counts as well-placed trust in those accounts remains the same in the unquestioning account with the caveat that the conditions of goodwill, etc., only apply to cases where those conditions are important to the relevant ends of the trust relationship.

Understanding trust as an unquestioning attitude allows us to answer the question "Should we trust AI?" without resorting to "no" as a blanket answer. Instead, it offers us the chance to explore how the relevant ends which trust aims towards affect judgements of trust. This enables us to better characterize the risks involved in extending our agency through utilizing AI and thus better helps us understand when and how to integrate AI into our lives. Thus, I would like to reframe the argument that began this section as follows:

Argument 2

Premise 1: An entity should only be trusted if it possesses the feature (or set of features), F, necessary for achieving the relevant ends.

Premise 2: Machines do not possess features necessary for achieving end, G.

Conclusion: Therefore, people should not trust machines with end, G.

The variable "G" stands in for ends that require features like personhood, goodwill, motivations, etc. This argument, however, leaves open the possibility that people can and should trust machines in contexts where the relevant end doesn't require the machine to possess features that it currently lacks, such as goodwill and the like. If people do unquestioningly integrate AI into their lives, when is this appropriate?

Why is it appropriate? And how do we tell the difference? That is, *when is AI worthy of this kind of trust?* That is the question that I wish to take up in this dissertation.

1.3 Conclusion

We have already seen that many existing philosophical accounts of trustworthiness are going to be unhelpful in addressing this question because they apply to only a limited range of cases - those in which some feature of persons is essential for the relevant end. We could, instead, look at what has been said specifically about trustworthy technology. However, accounts of trustworthiness developed specifically for AI typically consist of guidelines for developers and policy makers on creating ethically sound AI without reference to what trust is or what it would take to be a worthy recipient of that attitude. Thus, instead of providing an account of trustworthiness that addresses the attitudes that people have toward AI, these guidelines often consist either of a checklist of features that are too rigid to generalize to all uses of AI or consist of high level principles that are so vague that they fail to actually guide industry decisions (Mittelstadt, 2019). Neither case provides a definition of what trustworthiness is, such that AI ought to have the recommended features.

For example, the European Commission's High Level Expert Group (HLEG) on Trustworthy AI describe trustworthy AI as having three components: it must be lawful, ethical and robust from a technical and social standpoint. They include four ethical principles "which must be respected in order to ensure that AI systems are developed, deployed and used in a trustworthy manner" (High Level Expert Group on Artificial Intelligence, 2019). These principles include the respect for human autonomy, prevention of harm, fairness, and explicability.

These principles are realized in seven requirements that AI systems should meet:

1. AI systems should be subject to human agency and oversight,
2. AI systems should exhibit technical robustness and safety,
3. AI systems should preserve privacy and data governance,
4. AI systems should be transparent,
5. AI systems should recognize diversity and exhibit non-discrimination and fairness,
6. AI systems should promote environmental and societal well-being,
7. AI systems must be accountable.

While these are important principles and practices for the ethical use of AI, they are less helpful for particular end users considering whether to trust a given AI. To understand why, let us consider the case of privacy. According the HLEG, "AI systems must guarantee the privacy and data protection throughout a system's entire lifecycle. This includes the information provided by the user, as well as the information generated about the user over the course of their interaction with the system (e.g. outputs that the AI system generated for specific users or how users responded to particular recommendations). Digital records of human behavior may allow AI systems to infer not only individuals' preferences, but also their sexual orientation, age, gender, religious or political views. To allow individuals to trust the data gathering process, it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them." (High Level Expert Group on Artificial Intelligence, 2019, p. 17). However, what counts as guaranteeing user privacy is not universally agreed upon (Ess, 2006; Capurro, 2005).

Importantly, the view of privacy that dominates Western thinking and policy making is centered on the individual. On this view, privacy involves having control over our personal information, which allows us to develop and pursue our own individual interests and goals without interference from the government or community. In this way, privacy is seen as important and valuable because of its connection to individual

autonomy (Henkin, 1974; Ortiz, 1989; Dworkin, 1982; Westin, 1967). The view of privacy as promoting autonomy, however, is not shared across cultures. Theories of privacy not grounded in individual autonomy have been put forward by researchers working in a variety of traditions, including Confucian (Ma, 2019), Ubuntu (Reviglio and Alunge, 2020), Indian (Basu, 2012), Buddhist (Hongladarom, 2016), and Japanese (Capurro, 2005) traditions.

Given the diversity of interpretations of fundamental values, such as privacy, it can be problematic to rely on a checklist of ethical principles when identifying where to place one's trust. What counts as trustworthy according to the HLEG, which is constituted by European members, may not be counted as trustworthy in a different cultural context. Additionally, I believe that this applies not only on the cultural level, but also on the individual level. A technology might be wisely trusted by one individual, but not another, a thought which I will return to in detail later. In order to place our trust well when interacting with AI, we need something different than a checklist of high-level ethical principles. These may be useful in guiding policy-making and development at the level of a population within a given culture, but are less useful at identifying trustworthy AI when applied at the individual level or across cultures.

The perspective of the end user has thus been obscured in the literature on trustworthy AI. Meanwhile, much of the literature on trust at the individual level suggests that AI should not be trusted. I believe that ignoring the trust of end users and ruling that any trust in AI is misplaced are both mistakes. In this dissertation, I will argue that AI can be the appropriate recipient of trust. I begin this argument by distinguishing two different kinds of trustworthiness. The first prioritizes meeting the truster's expectations, while the second prioritizes meeting the truster's needs. In shifting the focus of trustworthiness to the truster's needs, however, debates emerge regarding what needs are legitimate and what constitutes the appropriate response

to those needs. These debates can generate broken trust or disputes about the trustworthiness of an agent. This is particularly important for understanding trust in technology, where a given AI application may meet legitimate needs in one context (or for one individual), but generate or reinforce harms in a different context (or for a different individual).

Thus, I aim to provide a definition of trustworthiness that (a) corresponds to the attitudes of trust that people display toward AI, (b) is able to generate reasons for why certain features are required in order for a person to count AI trustworthy in a given context, and (c) allows that an AI may be counted trustworthy in one context, but not another. In Chapters Two and Three, I focus on understanding what trustworthiness *in general* is, without much reference to the specific problem of identifying trustworthy AI. I have chosen to do this because I view identifying trustworthy AI as just an instance of a task that we do in a wide variety of contexts, many of which, I assume, will be more familiar to most people. Once I've explored the nature of trustworthiness in more familiar contexts, I will then apply it in the less familiar context of AI in Chapter Four.

In Chapter Two, I introduce two different ways of thinking about the relationship between trust and trustworthiness. On the one hand, trustworthiness might involve doing what the truster expects. On the other hand, trustworthiness might involve doing what the truster needs. These are often one and the same, but they do come apart, which has important consequences for how we determine the trustworthiness of an agent.

In Chapter Three, I argue that the domain specificity of trust creates interesting challenges for establishing robust trustworthiness. In particular, when trust is broken, it is often difficult to determine who is at fault. This makes it difficult to determine whether the trustee is untrustworthy or whether the truster has trusted poorly. This difficulty arises because the appropriate boundaries of the domain are

not always obvious. Indeed, sometimes the appropriateness of a domain varies across perspectives. This means that judgements of trustworthiness are determined relative to the truster's and trustee's perspectives.

In Chapter Four, I return to the question of how to determine the trustworthiness of an AI. I argue that judgements of trustworthiness are best determined relative to a particular individual's vulnerabilities and needs. I explore the consequences of this view in the context of apps used for healthcare.

Chapter 2

Trust and Vulnerability

Introduction

We trust others because we are vulnerable. We are not all-powerful, all-knowing beings. We cannot be everywhere or do everything. Our finite, embodied nature renders us inherently vulnerable. So we reach out to others to do or be what we cannot. When we reach out, however, we become vulnerable in a more narrow sense – vulnerable to the actions or inactions of those to whom we reach. Vulnerability thus plays into trust relationships in two different ways. On the one hand, our vulnerabilities precede trust relationships and on the other, vulnerabilities follow from specific trust relationships.

Typically, when considering trustworthiness, the focus is on responding to the vulnerabilities that follow from the trust relationship rather than those that precede it. For example, if I trust someone to deliver a letter, I am vulnerable to them not doing so. The trustworthy person, then, is the one who reliably delivers the letter. The emphasis, in such accounts of trustworthiness, is on the person's ability to fulfill the truster's expectations regarding the trust relationship. Being trustworthy, then, is akin to fulfilling a contract.

However, some vulnerabilities that precede trust cause people to pursue unhealthy or morally questionable trust relationships. Such relationships give rise to further morally unacceptable vulnerabilities. For example, suppose a child entrusts someone with the secret of their abuse. In doing so, they become vulnerable to the trustee

sharing their secret. It isn't immediately clear what the trustworthy person ought to do in this case. If the trustee keeps the secret, the child may be subject to further abuse. If the trustee shares the secret, they may break the child's trust. The trustee is in a difficult position. In honoring the child's trust, they fail to protect the child and in protecting the child, they fail to honor the child's trust. What does trustworthiness require in this case?

If we focus only on responding to the vulnerability that arises from the trust relationship, that is, the vulnerability to the secret being shared, then the trustworthy person should keep the secret. However, while the child may think it is better to trust the secret keeper, there are good reasons to suppose that they are mistaken about this. Importantly, the child's motivation for wanting their abuse kept secret is undergirded by the abuse itself (threats, coercion, etc). When the motivations for trust are themselves are rotten, as when motivated by threats or coercion, then the results of honoring that trust may also be rotten.

In this case, it is vulnerability to abuse that precedes the child's trust and if the trustee keeps this trust, then they reinforce that vulnerability. If the trustee is in a position to intervene to end the abuse and is willing to do so, then the child would be better off trusting the secret-sharer, other things being equal. I believe that the child ought to trust the person who is able to promote their well-being in the long-term, even if they fail to recognize which person this is. Thus, in this case, I believe the secret-sharer is more trustworthy than the secret-keeper. In order to defend this claim, I will sketch out and defend a novel account of trustworthiness which attends both to vulnerabilities that precede a trust relationship, as well as those that follow from it. According to this account of trustworthiness, the person who reports the abuse will count as more trustworthy, even though their actions may break the child's trust.

Trustworthiness, I believe, has to do with who a truster *ought* to trust in a given context. Identifying the extent of trustworthiness, however, requires that we look beyond the vulnerabilities that arise within a particular trust relationship to the vulnerabilities that precede it and, indeed, to the vulnerabilities of the person as a whole. Trustworthiness, on my account, has less to do with fulfilling a contract and more to do with attending to vulnerabilities - with navigating through the vulnerabilities that shape another person's motivations, expectations, fears, abilities, and desires while sorting out whether and how to respond.

In this chapter, I explore the ways in which vulnerabilities precede and follow from trust relationships and argue that the obligations of the trustworthy are more expansive than traditionally conceived. In section one, I develop an account of the two different sources of vulnerability relevant to trust relationships, building off of recent work from feminist ethics. In section two, I explore a tension that arises when trustworthiness is understood as involving competency in a domain and a commitment to meet expectations. In sections three and four, I resolve this tension through distinguishing between two different kinds of trustworthiness: naive trustworthiness and robust trustworthiness. Naive trustworthiness is blind to the vulnerabilities that precede a trust relationship, while robust trustworthiness requires attention to the broader array of vulnerabilities that shape trust relationships.

2.1 Preceding and Trust-Situational Vulnerabilities

In order to clarify the distinction between vulnerabilities that precede a trust relationship and those that arise within it, I will first discuss an existing distinction in the literature on vulnerability. In their introduction to *Vulnerability: New Essays in Ethics and Feminist Philosophy* (2014), Catriona Mackenzie, Wendy Rogers, and Susan Dodds introduce a taxonomy of types of vulnerability. Their taxonomy helpfully lays the groundwork for the distinction I wish to draw between what I call "trust-

preceding" and "trust-situational" vulnerabilities and thus I will briefly summarize their work before connecting it to my own.

2.1.1 Inherent and situational vulnerabilities

While vulnerability has often been associated with weakness, need, or victimhood, recent work in feminist philosophy challenges this assumption. On such accounts, vulnerability is universal to humans in virtue of our embodiment and corporeality (Fineman, 2010; Mackenzie, 2014). Our human bodies are inherently susceptible to harm in a variety of ways. If we are deprived of any one of our basic needs of food, shelter, meaningful relationships and the like, our quality of life is significantly diminished. These basic needs represent what Mackenzie, Rogers and Dodds describe as *inherent vulnerabilities*, which are vulnerabilities "inherent to the human condition" (Mackenzie et al., 2014, p. 7).

Some of these vulnerabilities are constant across people and lifespan, such as hunger, which all humans must live with from birth till death. Others, however, vary according to factors like age, gender, personality, health status, etc. For example, need for social interaction is inherent to the human condition, but some people may need more social interactions than others to satisfy this need. Such people have a greater vulnerability for diminished quality of life if they are unable to socially engage with others. While prolonged isolation may cause some to experience deep and troubling loneliness due to the lack of social interaction, others may not mind the solitude. As social beings, humans are inherently vulnerable to the effects of social deprivation, but personality differences may make some people experience this vulnerability to a greater extent than others.

While some vulnerabilities are inherent, others are context specific. In an effort to distinguish those vulnerabilities that are part and parcel of human life from those that generate special obligations in a given context, Mackenzie et al. distinguish a

second source of vulnerability. *Situational vulnerabilities* are those sources of vulnerability that arise in specific contexts, such as interpersonal, political, economic, social, or epistemic situations. Some situational vulnerabilities last only for a short time, others intermittently come and go, and others may last a person's whole life. For example, a person may be economically vulnerable for only a short time while temporarily unemployed. Alternatively, if they have gig work, they may find themselves intermittently vulnerable as they move between jobs. Or, if they are unable to work due to long-term medical or cognitive disability, they may spend their whole life economically vulnerable.

It is important to note that there is not a clear categorical difference between inherent and situational vulnerabilities. For example, all people must eat in order to survive and thus are, in a sense, inherently vulnerable to starvation. However, depending on the environment into which one is born, one may not actually ever face the potential of starving to death. Those in unstable environments, who do not have ready access to food, are vulnerable to starvation in a different way than those whose pantries are perpetually full. They are not only inherently vulnerable, but are also situationally vulnerable. The distinction between inherent and situational vulnerabilities, while not categorical, is useful for "understanding the different duties involved in responding appropriately to different kinds of vulnerability" (Mackenzie et al., 2014, p.8).

The blurriness between the categories of inherent and situation vulnerabilities highlights a second kind of distinction regarding vulnerabilities: occurrent versus dispositional. This distinction refers to the state of vulnerable person, whether they are potentially vulnerable or actually vulnerable. Mackenzie et al. provide the following example:

> "For example, all fertile women of childbearing age are dispositionally vulnerable to life-threatening complications in childbirth. But whether

> or not a pregnant woman is occurrently vulnerable to such complications will depend on a range of factors, both inherent and situational, such as her physical health, medical history, socioeconomic status, geographical location, access to health care, and cultural norms relating to pregnancy and childbirth. The dispositional–occurrent distinction serves to distinguish vulnerabilities that are not yet or not likely to become sources of harm from those that require immediate action to limit harm." (Mackenzie et al., 2014, p. 8)

The final distinction that Mackenzie et al. make in their taxonomy of vulnerability is between those vulnerabilities that are "particularly ethically troubling" (p. 9) and those that are not. Such ethically troubling vulnerabilities are the result of harms such as injustice, oppression, or abuse. An individual should not, say, be vulnerable to unjust arrest because of their race. A child should not be vulnerable to abuse because of their relative powerlessness in familial and social settings. These are situational vulnerabilities that ought not exist. Mackenzie et al. term such vulnerabilities *pathogenic vulnerabilities*. Pathogenic vulnerabilities are a subset of situational vulnerabilities which are morally unacceptable and which we should aim to eliminate.

In another essay, Mackenzie elaborates further on this distinction. She notes that even in a perfectly just society, some vulnerabilities, i.e. those inherent to human embodiment, cannot be eliminated. Instead, "what we should expect of a just society is that its social and political structures are responsive to and seek to mitigate the effects of inherent vulnerabilities, ensuring that their burdens do not fall disproportionately on the disadvantaged." (Mackenzie, 2014, p. 39). Pathogenic vulnerabilities, then, are those vulnerabilities that can and should be eliminated, rather than those that we must learn to live and deal with.

Given this taxonomy, Mackenzie et al. discuss several important dynamics between vulnerability and responses to it. They note two sources of pathogenic vulnerability which are especially important for the following discussion of trustworthiness. One source of pathogenic vulnerabilities is when a response intended to ameliorate a vulnerability exacerbates existing vulnerabilities or creates new ones. They provide the example of people with cognitive disabilities who require care and thus are at a heightened risk for sexual abuse by their carers. Similarly, a second source of pathogenic vulnerabilities is when a social policy intervention intended to ameliorate some vulnerability has the opposite affect and increases vulnerability.

2.1.2 Vulnerabilities and trust relationships

While the following is rooted in the taxonomy described above, my use of it is narrowly focused on making sense of the ways that vulnerability both contributes to and arises from trust relationships. Thus I will make several assumptions about the different types of vulnerability that Mackenzie et al. do not make. For one, I am assuming a clear distinction between inherent and situational vulnerabilities and between pathogenic and non-pathogenic vulnerabilities. Secondly, I am assuming that inherent vulnerabilities are not pathogenic. They simply accompany human life as we know it. One might argue that inherent vulnerabilities are simply a subset of situational vulnerabilities - namely those that arise in the situation of being human. If there are aspects of being human that are problematic, then, it would seem that some so-called inherent vulnerabilities are also pathogenic. Death, for example, is frequently treated as a pathogenic vulnerability that should be eliminated. The quest for the fountain of youth, or for the more technologically minded, uploading one's consciousness to the cloud are attempts at eliminating this vulnerability. In this chapter, however, I am going to assume that human bodies and their associated finitude are not things to be overcome and thus inherent vulnerabilities are not pathogenic.

Yet, as Mackenzie et al. note, the distinction between what vulnerabilities are inherent to us and those that are situational is not categorical. Consequently, while Mackenzie et al. do not discuss this in their work, I think it likely that in cases where it is unclear whether a given vulnerability is better understood as inherent or as situational, it may also be unclear whether it is a pathogenic vulnerability and thus whether it requires amelioration. I have said that I don't think that death is necessarily a pathogenic vulnerability, but surely there are many cases where it is. The world's population is currently facing a global pandemic, for example, and many are vulnerable to dying from it. Surely such a situation calls for life preserving intervention. However, if we consider someone in the final stages of a long life, it is less clear whether life preserving intervention is the appropriate response to their vulnerability to death. While the distinction between inherent and situational, pathogenic and non-pathogenic vulnerabilities is not clear cut in practice, for the purposes of developing a framework for how inherent and situational vulnerabilities relate to trust, I will assume that there is a clear distinction. Thus, I will treat inherent vulnerabilities as vulnerabilities that need not be eliminated, but do need to be recognized and attended to, and situational vulnerabilities as those which may, but not necessarily, require amelioration.

As noted above, people cannot meet all their basic needs without the help of others. Even the survivalist living alone in the forest depends on and is vulnerable to others to some extent. After all, they had to learn the skills necessary to survive alone from someone. Additionally, access the land and resources necessary to live alone is possible due to existing social structures. Survival skills will do a person no good if land owners, game wardens or park rangers prevent the opportunity to exercise those skills. Even the "self-sufficient" person needs others - if only the person who taught them how to provide for themselves or the social structures that provide the opportunities to use what they have learned.

As Joel Anderson has argued, vulnerability and autonomy are entwined:

> "Although a person's vulnerability can and often does diminish her autonomy, this is not true in all senses. Indeed, realizing autonomy as an ideal of personal agency requires certain forms of vulnerability. In particular, the acquisition and maintenance of autonomy rely on interpersonal relations in which vulnerability plays a vital role. And a fuller realization of the ideal of autonomy is sometimes possible only within specific social practices from which vulnerability cannot be eliminated." (Anderson, 2014, 135)

Inherent vulnerabilities render humans dependent on others, at least to some extent. Inherent vulnerabilities and the associated dependency on others invites the formation of trust relationships. That is, the contexts in which we are in need are those in which it often makes sense to turn to others for assistance. I am continually in need of food in order to stave off hunger, but I cannot grow or raise all the food that I need and also pursue my career goals. I simply don't have time, energy or material resources to do so. It makes a great deal of sense, then, for me to trust the existing food system - the farmers, delivery infrastructure, and grocery stores - with my food needs, rather than try to meet them myself. While not every dependent relationship is characterized by trust, nor should it be, when appropriate, adopting an attitude of trust can benefit the dependent person in many ways. The benefits and value of trust are frequently discussed elsewhere (Dimock, 2020; Jones, 2015; Locke, 2001; Morrone et al., 2009; Townley and Garfield, 2013), so I will not dwell on them here except to note that trust enables cooperation and thereby the formation and persistence of human society. We need trust in order to have societies that are worth living in; societies in which we recognize and respond to each others' needs.

It is important to note that while inherent vulnerabilities frequently precede trust relationships, that trust does not eliminate those vulnerabilities. As noted above,

inherent vulnerabilities accompany our humanity. Eliminating them would mean changing what we are, whatever that may be. Instead of eliminating inherent vulnerabilities through trust, we manage them. We keep from starving to death through trusting the agricultural sector to continually produce food, we keep from suffering from loneliness through continually seeking out meaningful friendships, and so on. I will always be vulnerable to starvation, loneliness and the like, but trust allows me to depend on others so that those vulnerabilities are not actualized into harms. Inherent vulnerabilities thus do not only invite the *formation* of trust relationships, but also the *maintenance* of trust relationships. I am motivated to continue my trust relationship with farmers and grocers because while they do not eliminate my vulnerability to starvation, they do help me manage that vulnerability.

However, in managing inherent vulnerabilities through trust, people make themselves situationally vulnerable. Situational vulnerabilities, in addition to arising in specific contexts, generate special obligations in that context. In a trust relationship, the truster becomes vulnerable to the actions of the trustee and the trustee, in response to this vulnerability, has a special obligation to the trustee that others outside of the trust relationship do not share. Suppose the survivalist works out a deal with a land owner who allows them access to space to exercise their survival skills undisturbed. This trust relationship makes the survivalist situationally vulnerable to the specific actions of the landowner. For example, the landowner might retract his offer or might alert the game wardens of the survivalist's illegal hunting activities. Whenever someone manages their inherent vulnerabilities through trust, they become situationally vulnerable to the trustee. Thus, while everyone is vulnerable to hunger, they are situationally vulnerable to those specific food systems from which they get their food.

In addition to following from the trust relationships that manage inherent vulnerabilities, situational vulnerabilities can also precede trust relationships and thus can

give rise to new situational vulnerabilities. Economic vulnerability, for example, is a situational vulnerability that might propel an unemployed person to trust a friend to help them find work. However, once that trust relationship is established, the unemployed person becomes more narrowly vulnerable to their friend's actions. The friend may fail to help them find work, or worse, may recommend to them an exploitative work setting, in which case new situational vulnerabilities emerge.

Situational vulnerabilities related to trust relationships, unlike inherent vulnerabilities, can be eliminated. If the friend successfully helps the unemployed person find a good job, then that person is no longer situationally vulnerable to their friend's actions pertaining to job hunting. In cases like these, the situational vulnerability is eliminated because the trust has been fulfilled. There is no longer any need for the person to continue trusting their friend to help them find a job because it has already been done.

Alternatively, the situational vulnerability can be eliminated if the trust relationship is abandoned altogether. If the friend is slow to help or offers recognizably bad advice, the unemployed person may call off the trust relationship and seek help elsewhere. In doing so, they are no longer situationally vulnerable to their friend's actions or inactions. They may however, remain situationally vulnerable in a range of other ways. They still, for example, remain economically vulnerable. Thus, the vulnerabilities that follow from trust can be eliminated through fulfilling or abandoning the trust relationship.

Additionally, these vulnerabilities can be eliminated if the situational vulnerability that motivated the trust in the first place is eliminated. If the job seeker unexpectedly lands their dream job independently of the actions of their friend, then there is no need for the trust relationship to continue. While similar to to cases of trust-fulfillment, this case differs because the trustee didn't contribute anything toward meeting the trust. Instead, the preceding conditions that motivated the trust changed, eliminating

the need for trust. While the job seeker might continue trusting their friend in a range of other ways, they no longer need trust them to help them find a job.

Situational vulnerabilities can thus both precede and follow from trust relationships and can be eliminated in several different ways. The vulnerabilities that follow from trust relationships are what I call "trust-situational vulnerabilities". These are the subset of situational vulnerabilities that arise whenever a trust relationship is established. The vulnerabilities that precede a trust relationship, which may be either inherent or situational, I call "trust-preceding vulnerabilities"'. I distinguish trust-preceding from trust-situational vulnerabilities because discussions of trustworthiness have largely ignored trust-preceding vulnerabilities. In the next section, I will demonstrate why this is a problem.

2.2 Competency and Expectations: A Tension

While trust is a powerful tool for addressing inherent and situational vulnerabilities, it is also opens a person up to experiencing a range of harms including betrayal, disappointment, manipulation, and exploitation. Trust can thus contribute to an increasing spiral of vulnerabilities through creating and reinforcing pathogenic vulnerabilities. Consider the economically vulnerable person who trusted their friend for help finding a job. The trustee may find them a stressful low wage job that offers little chance of upward mobility. While the job eases the immediate economic vulnerability, it also does not allow the vulnerable person the opportunity to develop skills for a better job or the time to find such a job. This ends up actually reinforcing the person's economic vulnerability through keeping them trapped in a cycle of low-wage work.

Trust not only potentially reinforces existing pathogenic vulnerabilities, it can also create new pathogenic vulnerabilities. The stress of the new job may make the person especially vulnerable to trouble maintaining important personal relationships,

substance abuse, and health problems, among others. Thus, trust is a risky business. Placed well, it can reduce the negative impacts of both inherent and situational vulnerabilities. Placed poorly, it can generate and reinforce pathogenic situational vulnerabilities. Determining when trust is well placed or not is clearly of critical importance. Oddly, philosophers have historically prioritized matters of trust over matters of trustworthiness (Hawley, 2014; O'Neill, 2018; O'Neill, 2020). Despite the overall asymmetry in attention given to trust and trustworthiness, several philosophers have addressed the question of what makes someone a worthy recipient of another's trust.

One general view of trustworthiness simply requires that a trustee is willing to do what the truster asks of them and is competent to do so. Versions of this view frequently appear when trust is discussed in contractual terms or game theoretic terms (see (Tutić and Voss, 2020; Hardin, 2001; Simpson, 2017)). The key question here is whether B will do what A trusts them to do, that is, whether B will mitigate trust-situational vulnerabilities by meeting A's expectations. If they do not, the harms associated with the trust-situational vulnerabilities are actualized. I will briefly discuss several views of trustworthiness that roughly take this form. While each view has certain advantages, all illustrate a tension between meeting a trustee's expectations and demonstrating competence in the relevant domain. I will briefly summarize the different views and then highlight several assumptions these viewpoints share which generate a problematic tension.

Russell Hardin's Encapsulated Interest account of trust and trustworthiness emphasizes the importance of the trustee's *commitments* and *competence*. According to Hardin, "the central problem in your trustworthiness is your commitment to fulfill another's trust in you" (Hardin, 2002, p. 28). When you trust someone to do something, how can you be sure that they will keep their commitment to meet your trust? You must first have evidence that they are actually competent to fulfill the trust.

The most committed trustee is hardly trustworthy if they lack the requisite skills or knowledge to meet the trust. Hardin is hardly unique in highlighting the importance of competence. Most accounts of trustworthiness similarly take it as necessary that the trustee has the skills and abilities necessary to do what is trusted.

What makes Hardin's account notable is his emphasis on whether it is in the trustee's interest to act on behalf of the truster's interests. How can we be assured of a trustee's commitments toward meeting our trust? We can be assured that the other will meet our trust if it is in their interest to act in our interest, that is, if their interests *encapsulate* our own. The kinds of cases he has in mind are those in which the truster trusts the trustee with respect to some future action. For example, if I buy something online, how can I be assured that I will receive what I have paid for? The trustworthy other must have good reasons for following through on commitments and these reasons should be apparent to both parties. He suggests that there are three sources of motivation, or three kinds of reasons, for honoring the commitments one makes when accepting trust. The first is an internal source of commitment, such as having a certain disposition or from moral commitments, character or habit. The second source includes external sources such as societal or social devices. These include legal consequences. Finally, there are mixed sources of motivation. For example, people can be induced to meet trust from social norms that have become internalized. Often, these reasons include a desire to maintain the relationship or to preserve a good reputation. The online seller has good reason to send me what I've bought because if they do not then I will not buy from them again and may damage their reputation through leaving a bad review. Ultimately, for Hardin, being trustworthy is a matter of self-interest. When meeting trust ceases to be in the trustee's interests, then they should not be counted trustworthy.

Karen Jones criticizes Hardin's account because, she argues, self-interest is an unstable motive for trustworthiness (Jones, 2012). Instead, we should count a person trustworthy if they take the fact of our dependency as a reason for meeting trust. According to Jones trustworthiness is "competence together with direct responsiveness to the fact that the other is counting on you" (Jones, 2012, p. 62). *Trust-responsiveness*, rather than encapsulated interest, is the relevant reason for meeting trust. Consideration of the other and their trust must directly move the trustworthy trustee, rather than, say, fear of legal or social consequences. Furthermore, trustworthiness is a dispositional property. An individual need not actually be in a trust relationship to be trustworthy, but if they were trusted, they would be moved by that trust. For Jones, the likelihood of following through when trusted is undergirded by the trustee being directly moved by the trust itself, that is, trust-responsiveness.

Both Hardin and Jones highlight the need for competency in the relevant domain. A trustee must have the relevant skills and abilities to do whatever it is they have been trusted to do. This means that trustworthiness is limited to specific domains. An individual might be trustworthy in the domain of online sales, but not trustworthy in the domain of plumbing. Additionally, this competency must be accompanied by relevant reasons for acting. For Hardin, these reasons are self-interested. For Jones, these reasons are rooted in the disposition to respond to trust.

Accounts that focus on mitigating trust-situational vulnerabilities, such as Hardin's and Jones', are useful for understanding certain social and economic dynamics. We readily see the importance of this kind of trustworthiness when we look at everyday interactions like buying products online. We frequently pay for items before we have seen them and before they are actually in our possession, leaving us quite vulnerable to the seller's actions. It is, unfortunately, not an uncommon experience to order something only to have it never arrive. Or, if it does arrive, to open the box and discover a cheap, poorly made approximation of the expected item. What we want,

in these scenarios, is for the trustee to not take advantage of our vulnerability, but to meet our expectations by following through on what they have committed to do. Thus, these views of trustworthiness are useful for discussing things like contracts or economic and political negotiations where the vulnerabilities in question are trust-situational. However, these views have several important limitations.

2.2.1 The Voluntariness Assumption

One limitation of accounts such as Hardin's and Jones' is that they embed an assumption of voluntariness. They are not alone in assuming voluntariness in contexts of trust. For example, in a discussion on trust and game theory, Tutić and Voss make this assumption explicit: "trust relations are *asymmetric* in the sense that the trustor has to decide whether or not to choose the risky option of placing trust. Trust relations generally require sequences of decisions." (Tutić and Voss, 2020, p. 176) When I trust the online seller, we both enter into the trust relationship voluntarily. They chose to market their items online and I chose to purchase their items. This assumption, however, does not hold in many instances of trust. Annette Baier criticizes accounts of trust that focus on contract-keeping for precisely this reason. She cites examples of trust between infants or children and their caretakers, wherein infants do not choose to trust their caretakers (Baier, 1986). While Baier is primarily concerned with how this voluntariness assumption is troublesome for trust, it is also troublesome for trustworthiness. Just as trusters do not always voluntarily choose who to trust, as in the case of infants, so too do trustees not always voluntarily choose to receive another's trust. Trust may be unwanted or even explicitly rejected by the trustee. Still, despite objections, trusters may, unreasonably, persist in trusting the unwilling trustee.

How should we think about the trustworthiness of the recipient of unwanted trust? On the face of it, it seems quite wrong-headed to deem the trustee as untrustworthy

because they are unwilling to meet trust that they did not want in the first place. Suppose I trust the chair of my philosophy department to give me one million dollars upon my graduating with my Phd, despite having no reason to trust him for this and despite his refusal to commit to this. If trustworthiness is simply about trust-situational vulnerabilities, then I have good reason to deem him untrustworthy when he fails to deliver the money on graduation day. But surely there is something wrong with an account of trustworthiness that reaches this conclusion because the chair never accepted my trust and even explicitly rejected it. The chair, in this situation, seems to have no obligation to meet my trust. Whether trust generates obligations, thus, seems important in assessing the trustworthiness of another.

We could solve this issue if we stipulated that in order for trust to generate obligations for the trustee, the trustee must voluntarily accept the truster's trust. Without this voluntary acceptance and the associated obligations, then failure to act does not indicate untrustworthiness. This seems to be the route that Jones takes in her discussion of the importance of the trustee's ability to signal to trusters that they are not willing or capable of accepting trust:

> "We want them [trustees] reliably to signal in, so that we can identify those whose agency is recruitable to extend the effectiveness of our own. We also want them reliably to signal out if they do not have the competencies on which we might base potential dependencies or if they are not willing to be responsive to those dependencies." (Jones, 2012, p. 75)

Yet, there are cases where trust may be unwanted, but an obligation to meet it may still exist; where the agent cannot simply "signal out" and remain worthy of trust. A reluctant parent, for example, may not have wanted the trust of a child, and yet they have some obligation to meet the trust of that child. If the parent neglects the child's trust-situational vulnerabilities, thus breaking the child's trust, should we count them untrustworthy? If trustworthiness requires voluntary acceptance of trust, then no.

But deeming the neglectful parent as trustworthy, even if they reliably signal their unwillingness to meet their child's trust, seems wrong as well.

2.2.2 Competency Versus Meeting Expectations

Cases of non-voluntary trust relationships raise several important questions about trust and trustworthiness. Where does the obligation to meet another's trust come from? Do the expectations that comprise the trust determine what the trustworthy agent ought to do or do the normative constraints for trustworthy action come from somewhere else? And does trustworthiness track meeting the truster's expectations or meeting some set of normative constraints (if the two come apart)? In order to answer these questions, we must understand in what way trustworthiness is normative.

In existing accounts of trustworthiness, it isn't always entirely clear what constrains trustworthy behavior. This lack of clarity arises because the trustworthy agent is supposed to demonstrate competency with respect to the domain in which they are trusted and are also supposed to meet the truster's expectations with respect to that domain. However, the requirement of competency sneaks in normative constraints that may conflict with trusters' expectations. There is a tension between *competency with respect to a domain* and *competency with respect to a set of expectations*. I will examine this tension in both Jones' and Baier's accounts of trustworthiness and then argue that we can resolve this tension by distinguishing two forms of trustworthiness.

For Jones, the central feature of trustworthiness is responsiveness to trust. We have reason to trust another when that other takes our trust as a reason to act. Thus, the obligation to act arises from the trust itself and it is the truster's expectations that determine what the trustee ought to do. Importantly, this means that when trusted to do something bad, the trustworthy person is still obliged to follow through, unless they signal that they do not want to be trusted in this way. Trustworthiness is thus decoupled from moral considerations. This means that when meeting trust is in

tension with other obligations, if we wish to be trustworthy in the relationship, we should prioritize meeting the truster's expectations.

In order to meet said expectations, however, trustees must be competent with respect to the relevant domain. For example, while trustworthiness itself is not necessarily a moral concept according to Jones, she does note that sometimes competency does require the trustee to use moral judgement:

> "When we trust professionals, from plumbers to physicians, we expect of them a technical competence (and minimal decency). However, the competence we expect in trusting need not be technical: when we trust a friend, the competence we expect them to display is a kind of *moral* competence. We expect a friend to understand loyalty, kindness, and generosity, and what they call for in various situations." (Jones, 1996, p. 7)

The domain of trust is thus central to understanding what a trustee ought to do. When the domain requires moral competency, the trustee must be relevantly moral. When the domain requires technical skill, the trustee must have the relevant technical skill. Confusingly, here Jones seems to indicate that the domains not only constrain what a trustee ought to do, but what a truster ought to expect. Friends ought to expect loyalty, kindness, and generosity of each other and ought to respond accordingly.

However, Jones later argues that what matters for trustworthiness is not the domain, but the truster's expectations:

> "...if a physician refuses to allow the expectations of her patients to shape her understanding of what, here and now, good medical practice consists in, her patients would not be justified in trusting her. (This explains why a physician might have reservations about having someone as her patient:

> if she feels that she will have objections to living up to her patient's expectations, she will think it difficult to maintain the proper relationship of trust.) For this reason, it would be a mistake to think that the ideally moral are always properly trusted. While it might be true that the ideally moral are properly trusted by those who are themselves ideally moral, it doesn't follow that they are properly trusted by those who are not." (Jones, 1996, p. 10-11)

Thus, in cases where professional obligations (as determined by the domain of interested) are in conflict with the obligations associated with the truster's expectations, it is the latter that take priority.

These two statements appear at odds with each other. On the one hand, trustworthiness requires competency, which includes understanding what actions or attitudes are appropriate to a given domain and thus what obligations arise out of the domain. On the other hand, trustworthiness requires allowing the truster's expectations to shape what actions or attitudes the trustee is obligated to do or hold. We expect our physicians to be technically competent, but also to shape their understanding of good medical practice to our expectations. But what happens if trusters do not have the right kind of expectations for a given domain? What if I expect my physician to do something medically inappropriate? On Jones' account, I should count the physician who refuses inappropriate treatment as untrustworthy, despite their competence and willingness to provide good medical care. Instead, I should count the physician willing to provide the shoddy medical care that meets my ill-informed expectations as trustworthy.

Annette Baier's work also demonstrates the tension between prioritizing obligations that arise out of competency with respect to a domain and the obligation to meet the truster's expectations.

Consider one of Baiers' examples:

> "One way in which trusted persons can fail to act as they were trusted to is by taking on the care of more than they were entrusted with-the babysitter who decides that the nursery would be improved if painted purple and sets to work to transform it, will have acted, as a babysitter, in an untrustworthy way, however great his good will." (Baier, 1986, p. 236)

Part of what it is to be trustworthy, on Baier's account, is to exercise good judgement in using one's discretionary powers. The babysitter is thus untrustworthy because he did not exercise good judgement. At this point, it is unclear whether Baier thinks that the failure to exercise good judgement pertains to ascertaining what obligations the domain of babysitting properly entails or pertains to obligations pertaining to the expectations of the parents. In this case, the failure to ascertain the obligations associated with the domain of babysitting overlaps with the failure to ascertain the obligations associated with the parent's expectations.

Later, however, she provides an example where she argues that the trustworthy person is obligated not to exercise discretion in how to meet the truster's expectations, but to exercise discretion in whether and when to meet those expectations. In that example, she suggests that trustworthiness has more to do with exercising good judgement with respect to the domain than in meeting the expectations of the truster:

> "To be someone to be trusted with a promise, as well as to be trusted as a promisor, one must be able to use discretion not as to when the promise has been kept but, rather, as to when to insist that the promise be kept, or to instigate penalty for breach of promise, when to keep and when not to keep one's promise. I would feel morally let down if someone who had

> promised to help me move house arrived announcing, "I had to leave my mother, suddenly taken ill, to look after herself in order to be here, but I couldn't break my promise to you." From such persons I would accept no further promises, since they would have shown themselves untrustworthy in the always crucial respect of judgment and willingness to use their discretionary powers." (Baier, 1986, p. 251-252)

This example indicates that competency with respect to a domain requires trustees to understand the relative importance of that domain in relation to other domains. Thus, when a truster's expectations conflict with the appropriate exercise of discretion in whether to meet those expectations, the trustworthy person prioritizes what they determine is more valuable. Suppose the example were reversed and the truster insisted that the promise should have been kept, even if that meant leaving the ill mother to fend for herself. The trustee has appropriately exercised her discretionary powers and thus should, according to Baier, be counted trustworthy, even though she failed to meet the truster's expectations.

However, later, Baier argues that when something is wrong with the trust relationship, we may have a moral obligation *to be untrustworthy*:

> "When the trust relationship itself is corrupt and perpetuates brutality, tyranny, or injustice, trusting may be silly self-exposure, and disappointing and betraying trust, including encouraged trust, may be not merely morally permissible but morally praiseworthy. Women, proletarians, and ex-slaves cannot ignore the virtues of watchful distrust, and of judicious untrustworthiness." (p. 253)(Baier, 1986).

This quote seems to endorse the view that trustworthiness has to do with meeting expectations, rather than exercising appropriate judgement in whether and when to meet expectations. If this is correct, then sometimes it is right to be trustworthy,

but other times it is wrong to be trustworthy. Trustworthiness, in this quote seems to track obligations that arise from truster's expectations, rather than the obligation to exercise appropriate discretion with respect to the relevant domain.

Both Jones and Baier's accounts illustrate this tension, I think, because it is unclear what obligations are associated with competency regarding a domain. Both view competency as key to trustworthiness, but the requirement of competency sneaks normativity into the concept of trustworthiness in a way that appears, in some cases, incompatible with the notion that trustworthiness is about meeting truster's expectations - that is, in doing what they have been trusted to do. Paul Faulkner takes the conflict between a truster's misplaced expectations and a trustee's commitment to doing what is right as evidence that trust should not be viewed primarily as a three-place relation (Faulkner, 2015). If trustworthiness is tied to a truster's expectations, rather than the fact of their dependence, then the trustworthy person is liable to engaging in whatever unethical behavior the trustee expects of them. Thus, he argues that in order to preserve the analytical connection between trust and trustworthiness, understood as doing the right thing, we must reject a three-place view of trust. Instead, he argues that we should understand trust as a one or two-place relation because this allows trustworthiness to be tied to doing the right thing, rather than tied to a truster's (potentially misguided) expectations. However, I think that it is not necessary to reject a three-place account of trust in order to resolve this tension. In what follows, I present a way to understand trustworthiness that corresponds with three-place trust.

2.3 Naive Trustworthiness

In order to resolve the tension between expectations and domain competency, I propose distinguishing two kinds of trustworthiness. In the first type, the trustworthy agent must be likely to meet the truster's expectations, regardless of how misguided

or immoral those expectations might be. I call this type of trustworthiness *naive trustworthiness*:

> **A is *naively trustworthy* to the extent that, were B to trust them with respect to *x*, A would be likely to meet B's expectations regarding *x*.**

I call this view "naive trustworthiness" because it is naive to the vulnerabilities that precede trust and whether such vulnerabilities are pathogenic. The quality of the truster's expectations determine the quality of the trustee's actions. If the truster's trust is preceded by pathogenic vulnerabilities that warp the truster's expectations, then the trustee's actions may reflect that warping. Thus, the naively trustworthy agent may create or reinforce pathogenic vulnerabilities with respect to the domain of trust. I define trustworthiness in terms of "to the extent that", rather than "if and only if", to indicate that trustworthiness comes in degrees (Alfano and Hujits, 2020). An agent can be more or less trustworthy in a particular domain and in relation to particular people.

Competency only enters the picture insofar as it is necessary for meeting the truster's expectations - it does not serve a corrective function in identifying whether expectations are misguided. This means that a naively trustworthy agent might actually display incompetence or gross negligence with respect to professional obligations or technical skill. A naively trustworthy physician, trusted to provide unnecessary and addictive medications, will do so, despite the harm their actions may inflict on the truster. The relevant competence needed in this case might include the ability to skirt the safeguards that are in place to prevent this kind of behavior. This competence, however, demonstrates deep negligence with respect to the domain of medicine.

If the truster has well-informed expectations with respect to the domain of trust, then the competency needed to meet the truster's expectations will likely align with

professional, moral or technical competencies. For example, a truster who has appropriate expectations of friendship, such as the expectation that friends will demonstrate loyalty, kindness, generosity, and the like, will count those who are morally competent as trustworthy.

As Jones points out, the ideally moral will properly trust those who are ideally moral and those that are not, will trust those who are not. Similarly, those who are knowledgable with respect to a particular domain are those most capable of developing appropriate expectations in that domain. Thus, the ideally competent will properly trust those who are also ideally competent. Those whose trust-preceding vulnerabilities include ignorance about the domain are more liable to develop quite poor expectations regarding that domain, and thus are more liable to count incompetent or negligent agents as trustworthy. Consider the way that a child is vulnerable to their parent. A child is completely dependent on their parent or guardian for care. Additionally, young children don't know what the domain of proper care entails. Thus, children trust their caretakers not only with care, but also with determining what constitutes that care. It is not uncommon for people to fail to recognize their caretakers' actions as abusive or unhealthy until adulthood, if ever, because those actions shaped their expectations for what care should look like. Consequently, children are vulnerable not only to receiving poor care, but to developing the expectation that poor care is normal or appropriate.

The kind of vulnerabilities that render a truster unable to determine the proper terms of trust are not unique to children. Indeed, most rational adults occasionally find themselves in situations where they trust others not only to act in accordance with their trust but to determine in important ways the terms of that trust. Whenever ignorance is the vulnerability that precedes a trust relationship, the trustee must do more than simply do what is trusted to them. They must, in addition, fill in the terms of the trust, so to speak. Naive trustworthiness requires that when A trusts B

to x, B will x. However, in the these cases, A also trusts B to determine the terms of x. We could say that in such cases A trusts B to y, where y equals "determine and do x". Naive trustworthiness would simply require that B determines what A should expect them to do and then does it. However, when a trustee does y, they create the terms by which their own trustworthiness is assessed. A trustee in such cases could harm the truster and still count as naively trustworthy.

Consider the way that a patient is vulnerable to their doctor. The lack of familiarity with a medical condition which precedes a visit to the doctor is the same vulnerability that often renders them unable to determine whether the prescribed treatment is appropriate. The patient thus trusts the doctor for care and also trusts them to determine what constitutes that care. Ideally, the patient can confirm that the doctor is trustworthy via the effects of the treatment, however, this is not always the case. For example, suppose a patient is receiving treatment, but their condition is not improving. They are told that the treatment is preventing their condition from worsening, but that no treatment will improve their condition. It might be the case that the physician is providing the best care possible and the condition is unfortunately truly treatment resistant. However, it might also be the case that there is an effective treatment that the physician could implement, but doesn't, perhaps because the treatment is new and the physician is reluctant to change their treatment strategies. The patient, whose trust is preceded by ignorance of the information that would enable them to distinguish between the two cases, may have no way to determine whether they are receiving adequate care. Indeed, the care of the physician reluctant to explore new treatment options may lead the patient to distrust a doctor who could actually help. Once poor care is accepted as trustworthy care, it is difficult to undo the damage because trust tends to reinforce itself. Jones describes the mechanism of reinforcement as "affective looping", wherein trust provides the grounds for its own continuance (Jones, 2019). Thus, a naively trustworthy physician

may reinforce the pathogenic vulnerabilities - preventable illness and lack of familiarity with that illness - that motivated the trust in the first place through setting sub-par expectations.

This problem may be exacerbated by what Thi Nguyen terms "agential gullibility", which occurs when a truster is too willing to incorporate other people or objects into their practical agency (Nguyen, 2019). Typically, gullibility is understood as a willingness to believe anything that others say. However, Nguyen argues that there is a sense in which, when we trust, we outsource our agency to other entities. We allow others to act on our behalf, and in doing so, allow others to shape our own actions, decisions, and expectations. The agentially gullible person does this without appropriate reflection or care. A patient might demonstrate agential gullibility in accepting whatever treatment plan a physician gives them without seeking other professional opinions.

In addition to cases where trusters are ignorant of the appropriate conditions of trust or are agentially gullible, there are cases where a truster may have a general idea of what they should entrust to the trustee and even have their trust met, but still be harmed. Continuing with medical example, the effects of a treatment may meet the patient's expectations, but not constitute trustworthy care. If a patient struggles with chronic pain, they may trust their doctor to help ease their pain, an altogether reasonable expectation. However, if they are not familiar with their condition or appropriate treatment methods, they are vulnerable to an unscrupulous physician prescribing a highly addictive pain-killer. This action eases the pain, and thus meets the patient's trust, but generates a pathogenic vulnerability for addiction. The unscrupulous physician thus meets the requirements for naive trustworthiness, but harms the patient in the process.

If we take naive trustworthiness as the standard for who to trust, then many people who perpetrate harm on others, either intentionally or unintentionally, meet that standard. When trying to determine who to trust, naive trustworthiness is only useful insofar as the truster's existing expectations are appropriate to the domain. Unfortunately, in many cases, trust is preceded by vulnerabilities that render the truster disadvantaged at forming expectations. For example, I trust my physician precisely because I often don't know what expectations should shape good medical practice regarding my medical condition. Those who are not similarly vulnerable, i.e. those who are medically competent, are in a better position to determine who is ideally competent and thus who they should trust but they also have less need to trust others with respect to that domain in the first place.

Someone might object at this point that while the disadvantaged truster may not have particularly well-informed expectations, they can still have general expectations that are well-placed. I may not know what appropriate medical practice looks like, but I may simply expect my physician to "foster my health". I may have no idea what exactly that looks like, but the expectation is such that negligent or incompetent physicians are ruled out as candidates for trustworthiness. This is correct. However, ignorance leaves a person much more vulnerable to developing bad expectations. People are much more likely to face exploitation or be otherwise harmed when they don't know what they should expect of others in a given domain. Thus, those whose trust is preceded by deep vulnerabilities, such as those born out of ignorance, abuse, or trauma, are also those whose trust in others is most likely to harm them even as their expectations are met. Their trust is most likely to create and reinforce pathogenic vulnerabilities, whereas the trust of those who are not so vulnerable is more likely to mitigate pathogenic vulnerabilities.

When we ask the question, "Is my trust well-placed?", we might mean one of several things. We might wish to know whether the agent will act we expect them to

act. In this, naive trustworthiness offers an adequate answer. However, we might also wish to know whether the agent will act appropriately toward us in a given domain, regardless of, or perhaps in spite of, any vulnerabilities that inhibit our ability to form appropriate expectations about that domain. In the latter, naive trustworthiness fails us. Naive trustworthiness is only as good as the truster's expectations, so if a truster is prone to holding inappropriate expectations, naively trustworthy others are likely to bring them harm. If we want an account of trustworthiness that answer the latter question, we need one that attends to both preceding and trust-situational vulnerabilities.

2.4 Robust Trustworthiness

If the obligations associated with trustworthiness do not arise from the expectations of the truster, where do they come from? Given that vulnerabilities are central to why we trust and why we often trust poorly, it may be helpful to take a step back and return to the literature on vulnerability. Within that literature, there is a question of why vulnerability generates obligation. One answer is that the vulnerabilities are the source of obligation (Kittay, 1999). Another is that vulnerabilities are a signal that alert us to the presence of salient moral claims, in particular those of need or harm (Mackenzie et al., 2014). This debate mirrors that of trust and obligation. Naive trustworthiness provides the conditions for trustworthiness when we view trust as the source of obligation. How should we understand trustworthiness when trust is not the source of obligation but a signal of some other source?

Trust, as described in section one, relates to vulnerability in two ways: vulnerability precedes it and follows from it. Trust, then, signals the presence of both preceding and trust-situational vulnerabilities. These, in turn, signal the presence of certain claims. These claims, of course, are not all equally important or legitimate. While I do not wish to delve deeply into what makes a claim legitimate as it is beyond

the scope of this project, I shall assume such claims include at least those arising from need or harm. If the claims are legitimate, then trust, by way of preceding and trust-situational vulnerabilities, signals the presence of needs or harms. The trustworthy agent, then, is the agent that meets those needs or mitigates those harms.

Thus, as an alternative to naive trustworthiness, which focuses only on mitigating trust-situational vulnerabilities through meeting the truster's expectations, I propose *robust trustworthiness*:

> **A is *robustly trustworthy* to the extent that, were B to trust them with respect to *x*, A would be likely to meet B's needs regarding *x* and/or mitigate harms to B associated with *x*.**

I call this kind of trustworthiness "robust trustworthiness", because it is able to withstand pressure from corrupt or wrong-headed trust relationships. It does so because the constraints on trustworthy action are determined not by the truster's potentially misguided expectations, but by the needs or harms that motivate the trust in the first place.

One point that is muddied in relying on the literature regarding vulnerability is the nature of the claims and obligations involved in trust relationships. Much of the vulnerability literature comes from care ethics and is concerned primarily with *moral* claims and obligations. But trust need not always primarily concern the moral. Trust can also involve epistemic vulnerabilities. A scientist, for example, might need some bit of information in order to continue their research project. They could expend the time and energy to do the experiments that would provide evidence of that information themselves, but they could also choose to trust the results of their fellow researcher's study without replicating it or independently verifying the results. In doing so, they become epistemically vulnerable. These vulnerabilities indicate certain needs that generate epistemic obligations, given that the epistemic

vulnerabilities signal legitimate epistemic needs (Johnson, 2020). Trust, then, can involve either moral and epistemic vulnerabilities and trustworthiness can involve meeting either moral or epistemic obligations.

Robust trustworthiness, unlike naive trustworthiness, cannot be determined through appealing to a truster's expectations, which may be under-determined or misdirected. Instead robust trustworthiness depends on whether a trustee meets the relevant needs or mitigates the relevant harms. The relevancy of a given need or harm depends on the domain of trust.[1] If the domain is child-care, as in the case of a child trusting their parent, then the relevant needs include physical needs such as food and shelter, as well social and emotional needs. Meeting these needs certainly requires both practical and moral competency, but the competency is with respect to the domain - not the child's expectations. After all, the child is ill-positioned to know what they ought to expect regarding child-care.

Of course, in grounding robust trustworthiness in needs and harms, robust trustworthiness cannot stand alone as a moral or epistemic concept. It depends on background frameworks for determining what counts as a legitimate need or relevant harm and thus what counts as a pathogenic or non-pathogenic vulnerability. As I noted earlier in this chapter, it is not my aim to here provide such a framework. However, this dependency has several important consequences worth noting. First, the dependency on an external evaluation framework means that trusters are not always in an epistemic position to assess the trustworthiness of a trustee. As noted above, trusters are frequently vulnerable to manipulation and deception. A truster's vulnerabilities often render them especially ill-positioned to identify trustworthiness or the lack thereof. Secondly, trustees may also be epistemically ill-positioned to determine their own trustworthiness. A trustee, for example, may be able and willing to

[1] I will present a formal definition of domains of trust in the following chapter and discuss how trust domains are negotiated in particular trust relationships.

meet a truster's expectations, but fail to recognize how meeting such expectations may reinforce a pathogenic vulnerability. This failure may arise from an ignorance of the needs or harms relevant to the domain or trust. Alternatively it may arise from ignorance of a truster's particular circumstances. For example, a trustee may unwittingly enable a trusters' harmful behavior if they are unaware that the truster has a particular harmful pattern of behavior. Finally, an external evaluation framework prevents us from counting as trustworthy a problematically paternalistic trustee who simply does whatever they think is best for the truster, regardless of what the truster wants or needs.

Robust trustworthiness avoids several of the challenges that limit the useful application of naive trustworthiness, while retaining many of its beneficial features. Many of the cases in which naive trustworthiness is usefully applied, such as in making sense of contractual trust or economic dynamics, robust trustworthiness may be similarly useful. For example, in the simple case of online shopping, what we expect of the naively trustworthy agent is that they will deliver what has been ordered. In the majority of cases, this is also what the robustly trustworthy agent does, with some important exceptions. The robustly trustworthy online marketer, unlike the naively trustworthy marketer, must give some care as to the content of their sales. The seller of illegal items or otherwise harmful wares does not meet the standard for robust trustworthiness, regardless of their competence or ability to meet their customer's expectations. The robustly trustworthy online marketer must also give some care as to how they meet their customers' expectations. It is not enough *that* they deliver what is expected, they must also demonstrate respect toward their customer's privacy and autonomy in how they treat their customers throughout the process. Robustly trustworthy online sellers would not, for example, sell their customers' data without permission, even if data privacy is not expected, because data privacy is important to the domain of online interactions. In cases where the content of the trustee's actions

in meeting the truster's expectations do not violate norms that are important to the domain of trust, then individuals who are naively trustworthy will most likely also count as robustly trustworthy.

This is important because there are many instances of trust in which the preceding vulnerabilities are not laden with grave moral or epistemic considerations and, as such, do not signal the presence of important needs or harms. If robust trustworthiness were focused only on meeting important needs or mitigating serious harms, then robust trustworthiness would have little use in many day-to-day instances of trust. While many day-to-day cases of trust do not involve important life-or-death vulnerabilities, they instead signal the presence of mundane, and often inherent, vulnerabilities of embodied existence. For example, I might trust my partner to pick up some groceries on his way home from work because I am located further away from the store than he is, or am tired, or just don't feel like stopping myself. My finitude and my body's propensity to get tired at the end of the day are vulnerabilities that do not signal the presence of serious needs or harms. The legitimacy of my claim does not lie in morally or epistemically salient needs or harms, but in the give-and-take of my particular relationship with my partner, in the arbitrary expectations that we have of the other and the routines that we follow. If my partner forgets to stop for groceries, I won't be significantly harmed, just slightly inconvenienced. In such cases, the preceding vulnerabilities, finitude and weariness, tend to fade into the background because they are simply part and parcel of human life.

When the preceding vulnerabilities are not pathogenic, then they do not signal the presence of immediate needs or serious harms which ought to be eliminated. Instead, they signal the presence of needs that must be lived with and managed, i.e. those associated with inherent vulnerabilities or non-pathogenic situational vulnerabilities. The relevant needs in cases like are those associated with the truster's expectations regarding the trust-situational vulnerabilities. Thus, when the preceding vulnerabilities

are not pathogenic, the robustly trustworthy person attends to the trust-situational vulnerabilities through meeting the truster's expectations. It is in these cases that robust trustworthiness and naive trustworthiness deem the same people trustworthy.

There are other cases, however, in which robust trustworthiness draws helpful distinctions where naive trustworthiness does not. Robust trustworthiness helps us to explain why some cases of unwanted trust still generate obligations, even when the trustee signals that they do not wish to be trusted. We can, for example, distinguish the neglectful parent case from the philosophy chair case. In both cases, the trust was rejected. In applying the standards of naive trustworthiness, both are deemed trustworthy. However, in applying the standards of robust trustworthiness, we can distinguish between these cases and explain why the neglectful parents ought not count as trustworthy. In the philosophy chair case, I trust the chair to give me one million dollars, and am trust-situationally vulnerable to him not doing so. However, that trust does not signal the presence of a legitimate claim. Instead, it signals the presence of morally questionable trust-preceding vulnerabilities: my own greed and self-entitlement, neither of which generates an obligation to respond. I do not need that money and have no good reason to lay claim to it. In the neglectful parent case, however, the child's trust-preceding vulnerability *does* signal the presence of certain legitimate claims. A child *needs* care and is dependent on their parents to guarantee that care.

Additionally, robust trustworthiness is a more appropriate standard than naive trustworthiness in cases where the trustee helps shape the truster's expectations, which will then be used to evaluate the trustee. Naive trustworthiness is not particularly helpful in identifying those who manipulate, exploit or otherwise harm those whose expectations in a given domain are particularly malleable.

2.5 Conclusion

The poet David Whyte writes of vulnerability:

> "Vulnerability is not a weakness, a passing indisposition, or something we can arrange to do without. Vulnerability is not a choice. Vulnerability is the underlying, ever-present and abiding undercurrent of our natural state." (Whyte, 2020, p. 261)

In this chapter I have explored several varieties of vulnerability and have argued that trust is an important tool for mitigating the harmful effects of this "abiding undercurrent of our natural state". However, even as we use trust in this way, we become further vulnerable to those in whom we trust. Navigating trust relationships is difficult because when the need for trust arises, it is not always obvious whom we should trust. I have argued that there are two different ways we can resolve this problem. One way is to seek to trust those who are likely to do what we expect or want them to do. Another is to seek those who are likely to meet our needs with respect to the domain that we trust them with. These often coincide, but do not always. If we adopt the latter view of trustworthiness, as I have suggested we ought to, then, in order to identify the relevant needs or harms to which the trustworthy person ought to attend, it is important to know where lie the limits of a given domain of trust. If I desire to be trustworthy in the robust sense, what needs should I be aware of when trusted? How is the relevance of a need or harm to a given domain determined? I turn to these questions in the following chapter.

Chapter 3

Negotiating Domains of Trust

Introduction

When we tell someone that we trust them, we generally do not mean that we trust them *with everything*. Rather, we mean that we trust them regarding some particular domain of interest. Formally, we can express the trust relationship as a three-place relation: A trusts B with respect to domain x (Baier, 1986; Jones, 1996; D'Cruz, 2018; Hawley, 2014; D'Cruz, 2020) [1]. When I tell the hairdresser that I trust them, I mean that I trust them to make my hair presentable, not that I trust them to perform dental surgery on me or commit themselves to spend the rest of their lives with me. Likewise, I trust the dentist with my teeth and my partner with sharing in my life journey. The domain specificity of trust is nothing new in the literature (Baier, 1986; Jones, 1996; Hardin, 2002; Hardin, 2006; Nguyen, 2019), yet little is said about what a domain is or what determines the boundaries of a given domain. Are the boundaries determined by societal norms? Individual ideals? Platonic forms? Additionally, it is often taken for granted that these domains, however determined, are obvious to those involved. Yet, in this paper, I argue that, in many cases, the boundaries of trust domains are not always obvious. Indeed, I believe that people regularly disagree about or misunderstand the boundaries of these domains. These disagreements and misunderstandings can lead to broken trust, sometimes despite the best efforts of both

[1] Trust is also often understood as a two-place relation wherein A trusts B. It is an open question how two-place trust relates to three-place trust (Domenicucci and Holton, 2017; Faulkner, 2015). In this chapter, I bracket this issue and focus on trust as a three-place relation.

truster and trustee. When this occurs, it is not always obvious who is to blame for the broken trust, if anyone. This chapter has several aims. I first wish to define trust domains and distinguish them from related concepts. Secondly, I differentiate three kinds of broken trust that arise from disagreements and misunderstandings regarding trust domains. I then describe three features of trust domains that can generate these disagreements or misunderstandings and discuss how these features complicate the process of attributing blame when trust is broken. Finally, I distinguish blameless trust-breaking from faultless trust-breaking and argue that both are possible.

3.1 Trust Domains

As noted above, the domain specificity of trust is frequently mentioned in accounts of trust as a three-place relation. For example, Karen Jones's (1996) account of trust states that trust is "an attitude of optimism that the goodwill and competence of another will extend to cover the domain of our interaction with her, together with the expectation that the one trusted will be directly and favorably moved by the thought that we are counting on her" (Jones, 1996, 4). Jones elaborates on the domain specificity of trust in the following:

> "This is not to say that the optimism itself is qualified and instead of being unreserved optimism is a qualified or restricted optimism. What is qualified is not the optimism itself, but the domain over which it extends. So, for example, the optimism we have about the goodwill and competence of strangers does not extend very far. We expect their goodwill to extend to not harming us as we go about our business and their competence to consist in an understanding of the norms for interaction between strangers. For a man to run up at full speed behind a woman on an ill-lit street is to display a lack of such competence, and, even if he was simply out for

> a late night run and meant no harm, he has given the woman reason to distrust him." (Jones, 1996, 7)

Jones here seems to suggest that trust domains include norms of interacting, although it is unclear whether she thinks that domains always include these. If we do think of domains as determined by social norms, then a domain is a kind of social construction that, at least in part, is determined by factors external to both truster and trustee. Viewing domains as social constructions has some advantages. For example, they could function as guidelines for both truster and trustee, indicating when trust is placed poorly and when trust has been betrayed. I have trusted poorly if I trust someone with something that lies outside of the social norms associated with that relationship. I trust my dentist poorly when I trust them to cut my hair instead of check my teeth for cavities. Similarly, I am an untrustworthy dentist if I can cut your hair, but not check your teeth for cavities. Understanding domains as social constructions, however, obscures some important dynamics of trust relationships. We feel betrayed, for example, not when someone acts in a way that is incompatible with social norms but when someone acts in a way incompatible with the expectations that we, as individuals, have of them. The expectations that matter to a trust relationship, I believe, are those held by the person doing the trusting, rather than the society to which they belong.

Thus, I take trust domains to be sets of expected behaviors or attitudes held by individual trusters. Expectations come in many varieties. Some expectations are normative, expressing what we think should be done, but not necessarily what we think will actually be done. Others are just the opposite, describing what we think will happen, but not necessarily what ought to happen. It is an open question which kind of expectation is relevant to trust (Jones, 2012; Nguyen, 2019). I think, however, that what follows applies regardless of which variety of expectation one focuses on and so I will set aside the question of normative versus descriptive expectations.

While the objects of our trust are not necessarily expectations, our expectations determine the boundaries of what behaviors or attitudes are included in the trust domain. For example, when I consider the domain of dentistry, I expect behaviors like diagnosing dental problems and filling cavities and I expect attitudes like general goodwill towards myself. We can represent such domains formally as sets:

$$Dentistry = \{diagnosing\ dental\ problems,\ filling\ cavities,\ general\ goodwill\}$$

Note, however, that each item in the set can be broken down into another set of expectations. For example, the expected behavior of filling cavities includes expected behaviors like administering anesthesia, drilling out rotten areas, and replacing the drilled out area with the appropriate substance. Thus, trust domains can consist of *nested sets* of expectations:

$$\begin{aligned} Dentistry = &\{\{diagnosing\ dental\ problems\},\ \{administering\ anesthesia,\\ &drilling\ out\ rotten\ areas, filling\ area\ with\ appropriate\ substance\},\\ &\{general\ goodwill\}\} \end{aligned}$$

These domains are determined by the expectations of individual trusters. Thus, trust domains can differ among different people. A and B might hold different expectations regarding the domain of dentistry, in which case their trust domains diverge accordingly. In the above examples, I haven't qualified this. Trust domains would be better characterized as *domains given the individuals whose expectations comprise the domain*:

$$\begin{aligned} Dentistry|A = &\{\{diagnosing\ dental\ problems\},\ \{administering\\ &anesthesia,\ drilling\ out\ rotten\ areas,\\ &filling\ area\ with\ appropriate\ substance\},\\ &\{general\ goodwill\}\} \end{aligned}$$

Importantly, what a truster expects may differ from what a trustee *believes* that a truster expects. It is important to recognize that a trustee's perceptions of a trust domain can, and often do, come apart from the trust domain itself. Thus I will call a trustee's perception of a trust domain a "perceived domain". A perceived domain is the set of behaviors and attitudes that a trustee believes is expected of them in a trust relationship. Ideally, the trustee holds accurate beliefs about what the truster expects of them and thus the perceived domain should be identical to, or at least similar to, the trust domain. We can represent the ideal case where a dentist, B, accurately perceives a patient, A's, expectations as follows:

$$\begin{aligned} B(Dentistry|A) = \{&\{diagnosing\ dental\ problems\},\ \{administering \\ &anesthesia,\ drilling\ out\ rotten\ areas, \\ &filling\ area\ with\ appropriate\ substance\}, \\ &\{general\ goodwill\}\} \end{aligned}$$

A perceived domain is different still from the set of behaviors and actions that a trustee commits to doing when accepting another person's trust, which I call the "commitment domain". When a patient goes to the dentist, they may simply expect the dentist to diagnose any dental problems. The average patient, however, probably isn't aware of the domains that are nested within that fairly coarse-grained expectation of diagnosing dental problems. That is, they are most likely unaware of what dental problems to look for or how to do so effectively. The average dentist, however, hopefully is aware of what expectations should be nested within this coarse-grained expectation and is also probably aware that the patient is ignorant of these. Thus, what the patient expects, what the dentist thinks the patient expects and what the dentist commits to doing can all come apart. While a commitment domain may be different from a trust domain, a positive feature in cases like dentistry, it should still ideally be compatible with the trust domain. As I am not a dentist, I will refrain

from attempting to spell out the range of commitments that dentists take on when accepting the trust of patients. It is enough to recognize that the set of commitments will be far more fine-grained than the average non-dentist's set of expectations. We can represent commitment domains generally as follows:

$$TrustRelationship * | B = \{B's\ commitments\ to\ A\ with\ respect\ to\ the\ trust\ relationship\}$$

I believe this account of trust domains is compatible with domains as they are described in most accounts of three-place trust. It is, for example, compatible with Jones' stating that we trust others to follow social norms, because our individual expectations are often shaped by the society and culture to which we belong. Indeed, this shared understanding is important for accurately perceiving one another's trust domains. However, the expectations that bound trust domains also arise from a variety of sources beyond shared norms, including our knowledge, our personalities, prior experiences, and the like (Potter, 2002; Scheman, 2020). Additionally, these expectations may be distorted due to prejudice or bias, leading us to distrust when we ought to trust and vice versa (Medina, 2020). In practice, then, I think it is actually often quite difficult to pick up on what others expect of us and even when we do, we may find their expectations inappropriate. When this happens, trust is broken. In the following section I will elaborate on how trust-breaking occurs.

3.2 Types of Trust-Breaking

A trust relationship involves a truster forming some set of expectations of another, a trustee who attempts to understand these expectations and who commits to a set of actions in response. A trustee breaks a truster's trust when their actions or attitudes violate the expectations in the trust domain. There are three different ways that this can occur: 1) they may inaccurately perceive the trust domain, 2) they may accurately

perceive the trust domain, but commit to actions that are outside its limits, or 3) they may accurately perceive the trust domain, commit to fulfilling those expectations, but violate a hidden expectation. I will use an example from Annette Baier's (1986) paper to demonstrate the differences between these types of trust-breaking. In the example, a babysitter is entrusted with the care of a child. The babysitter thinks the nursery would be improved with a coat of purple paint and sets to work accordingly. In taking on this task, Baier says that the babysitter has not acted in a trustworthy manner. However, it isn't immediately clear what has gone wrong in this trust relationship or why we should consider the babysitter untrustworthy. I am proposing that there are three distinct ways in which the babysitter might have violated the parents' trust. Unlike Baier, however, I argue that in each type of trust-breaking, it is not obvious that the babysitter is at fault for the broken trust.

3.2.1 Direct Misunderstanding Mismatch

Firstly, it is possible that the babysitter simply did not accurately perceive what the parents expected of him. Perhaps he misheard their instructions. Or, perhaps every previous babysitting gig involved repainting nurseries and so he assumed this one did as well. Either way, the trust domain differed from the perceived domain:

$$Childcare|Parents = \{feed\ child,\ change\ diapers,\ put\ child\ down\ for\ nap,\ age\ appropriate\ play\}$$

$$Babysitter(Childcare|Parents) = \{feed\ child,\ change\ diapers,\ put\ child\ down\ for\ nap, age\ appropriate\ play,\ upgrade\ bedroom\}$$

In this case, the babysitter simply misunderstood what expectations were included in the trust domain. That is, there is a mismatch between the perceived domain and

the trust domain. I call this first kind of trust-breaking a *direct misunderstanding mismatch*. When these mismatches occur, trust is broken as trustees act on their misperceptions of what is entrusted.

Broken trust from a direct misunderstanding mismatch may be the fault of either the trustee or truster. We might attribute fault to the trustee when they commit to doing something that they actually don't know how to do and thus fail to meet the trust. If someone commits to babysitting, but they don't actually know anything about children, then it is unlikely that they will accurately perceive what the parents expect of them. Even in cases where the trustee is familiar with the entrusted task, they may still be at fault if they fail to take the appropriate measures to listen to and understand the particular truster's expectations. If the parents left written instructions asking the babysitter not to paint the nursery, but the babysitter failed to read the instructions because he felt confident that he already knew what was expected, then the parents would have good reason to blame the babysitter for their broken trust.

More problematically, epistemic injustice can cause direct misunderstanding mismatches. Trustees can demonstrate what Miranda Fricker terms "testimonial injustice" when they do not assign an appropriate level of credibility to the words of the truster (Fricker, 2007). This may happen if they wrongly make assumptions about the trust domain based on unfair biases or stereotypes, rather than on what the truster has actually said to them.

Alternatively, direct misunderstanding mismatches may be the fault of the truster, rather than the trustee. We might blame the truster for broken trust due to a direct misunderstanding when they fail to communicate their expectations clearly. Suppose the babysitter thinks he is supposed to arrive at 7:30pm, but the parents expect him to arrive at 5:00pm. If the parents never communicated the expected arrival time to the babysitter, the babysitter can hardly be blamed for arriving late.

However, it is not always obvious who is to blame when a direct misunderstanding mismatch occurs. The trustee may argue that the limits of the trust domain were not communicated clearly. Likewise, the truster may argue that the trustee failed to take appropriate measures to listen and understand what was communicated about the limits of the trust domain.

3.2.2 Direct Conflict Mismatch

Secondly, it is possible in the babysitting case that the perceived domain is identical to the trust domain, but the commitment domain is incompatible with the trust domain. Suppose that the babysitter recognizes that the parents do not expect or want him to renovate the bedroom, but he decides to paint the nursery anyway because he thinks that it is in everyone's best interest to do so:

$$Childcare|Parents = \{feed\ child,\ change\ diapers,\ put\ child\ down\ for\ nap,\ age\ appropriate\ play\}$$

$$Babysitter(Childcare|Parents) = \{feed\ child,\ change\ diapers,\ put\ child\ down\ for\ nap,\ age\ appropriate\ play\}$$

$$Childcare*|Babysitter = \{feed\ child,\ change\ diapers,\ put\ child\ down\ for\ nap,\ age\ appropriate\ play,\ upgrade\ bedroom\}$$

I call this kind of problem a *direct conflict mismatch* because trust is broken due to a direct conflict between the expectations in the trust domain and the trustee's commitments. Direct conflict mismatches are, I believe, the most commonly discussed type of broken trust. The standard case of broken trust is one in which the trustee accepts another's trust, but does not follow through. That is, the trustee fails to commit to doing what is necessary to meet the expectations of the truster. However,

not all cases of direct conflict mismatches are due to a moral failing on the part of the trustee. In some direct conflict mismatches, the trustee may refuse to meet trusters' expectations because they believe those expectations are misplaced. For example, the babysitter may think that while the parents don't expect a purple nursery, they *should* expect it.

Here again, trustees can break trust unfairly when they presume to know better than the truster. Yet sometimes trustees have legitimate reasons to commit to actions that do not align with the expectations in the trust domain. Trust can, for example, be exploitative. If an employer entrusts an employee with a task over and beyond what they are paid to do, the employee has good reason not to commit to doing it. Trust can also be genuinely misguided. Suppose a patient has chronic pain and trusts their doctor to make it stop. If this is not medically possible, however, the doctor should not commit to meeting that expectation, but should commit to some alternative course of action.

3.2.3 Indirect Misunderstanding Mismatch

Direct conflicts are not the only source of mismatches between the commitment domain and the trust domain. Recall that commitment domains can be, and often are rightly, finer-grained than the expectations that bound trust domains. Suppose that the parents entrust the babysitter with the general well-being of their child. The babysitter takes this task very seriously. Having read that purple rooms help foster a child's cognitive development and noting that the child's room is plain white, he sets about to remedy the situation with a few coats of purple paint. After all, the parents have entrusted this child's well-being to him and surely fostering a child's cognitive development is part of promoting general well-being.

In this case, the following has occured:

$$Childcare|Parents = \{feed\ child,\ change\ diapers,\ put\ child\ down\ for\ nap,\ promote\ general\ well-being\}$$

$$Babysitter(Childcare|Parents) = \{feed\ child,\ change\ diapers,\ put\ child\ down\ for\ nap,\ promote\ general\ well-being\}$$

$$Childcare*|Babysitter = \{feed\ child,\ change\ diapers,\ put\ child\ down\ for\ nap,\ \{age\ appropriate\ play,\ bedroom\ upgrades\}\}$$

It is less clear how trust was broken in this example. The babysitter did, after all, understand that they were being entrusted with promoting the child's general well-being and did, indeed, promote the child's well-being (assuming that purple nurseries do promote cognitive development). However, the babysitter's commitments with respect to promoting the child's general well-being were still incompatible, but incompatible with what? The commitments were compatible with every expectation in the trust domain. Thus, the mismatch is not between the commitment domain and the trust domain. Instead it is between the commitment domain and some other expectation that was not previously apparent. Indeed, the parents had probably never even considered whether their trust domain should include or exclude renovations. Yet, once faced with a purple bedroom, it was clear to them that this was a violation of an expectation implicit in the trust domain but perhaps not recognized by either party. I call these *hidden expectations* because neither party is aware of them when the trust relationship begins. Thus, despite doing his best to be trustworthy, the babysitter still broke the parent's trust because he violated a hidden expectation. When a commitment domain contains more fine-grained expectations than the trust domain, trusters can find themselves faced with unexpected responses their trust.

In this case, the response was incompatible with hidden expectations. I call such problems *indirect misunderstanding mismatches*.

However, not all unexpected behavior is incompatible with hidden expectations. If the babysitter washes all the dirty dishes, this may not have been expected, but the parent's may find it acceptable as it does not violate any hidden expectations. The actions may actually set a precedent for expectations of future babysitters. Baier argues that

> "When we are trusted, we are relied upon to realize *what* it is for whose care we have some discretionary responsibility, and normal people can pick up the cues that indicate the limits of what is entrusted." (Baier, 1986, 236)

This is similar to Jones' statement that competency includes an "understanding of the norms for interaction". If Baier and Jones are correct, normal people should recognize that washing the dishes, while not explicitly within the trust domain, is not incompatible with its boundaries or with any hidden expectations, while redecorating the nursery is obviously incompatible. However, picking up these cues is not nearly as simple as Baier and Jones seem to assume.

It is difficult, when trusted, to know exactly where the boundaries of the trust domain lie or what expectations may be hidden therein. Trusters cannot account for all possible scenarios that the people they trust may face and thus cannot make explicit all the relevant expectations. It is quite reasonable, for example, for parents to have relatively general expectations of babysitters, like "promote my child's well-being". This provides the requisite leeway for babysitters to handle unexpected situations. Trustees thus need to exercise discretion when determining how to carry out such coarse-grained expectations. This leaves trusters vulnerable to trustees' discretionary powers (Baier, 1986; Mackenzie et al., 2014). Conversely, trustees cannot account for all possible expectations, hidden or not, that might be placed on them.

This puts the trusted person in an awkward position. They are trusted with respect to a certain trust domain, but may not have a clear idea where the boundaries of that domain lie or what commitments may violate a hidden expectation, leading to a direct or indirect misunderstanding mismatch. One person may appreciate having their dishes washed. Another may find it offensively patronizing to have someone else clean up their mess without being asked to do so.

If a trustee wishes to be found trustworthy, it is important that they navigate well the boundaries of others' trust domains when exercising their discretionary powers. That is, it is important that they try to align their commitment domain with others' trust domains, unless inappropriate to do so. Similarly, trusters who do not wish to be disappointed must fix the boundaries of their trust such that trustees can align their commitment domain accordingly. Both tasks, however, are challenging. Thus, when trust is broken, it may not be obvious who is at fault or what must change in order to restore broken trust. The process of keeping trust and restoring broken trust requires that both parties negotiate the boundaries of their respective domains in order to avoid future mismatches. There are three features of domains that take center stage in these negotiations: the scope of a domain, the ordering of a domain, and the rigidity of a domain. I will focus here on trust domains, rather than perceived domains or commitment domains. However, perceived and commitment domains also have these features.

3.3 Features of Domains

3.3.1 Domain Scope

Domains of trust vary in scope. As expectations move from general to specific, the trust domain narrows as more fine grained expectations are added. For example, I

might go to the hairdresser and ask them to cut my hair so that it is shorter:

$$Hairdresser|A = \{cut\ hair\ shorter\}$$

This is a fairly *wide* trust domain. It is wide in the sense that there are a wide range of actions that a hairdresser can do that will be compatible with the trust domain. They could cut my hair into a bob, give me a pixie cut or shave my hair off completely. Any of these actions would do.

There are two different ways that a domain can narrow. One way is that an expectation can become more fine-grained. This is where the idea of nested sets becomes important. Suppose that I do want shorter hair, but I bring in a picture expecting to leave looking exactly like that picture. The resulting domain might look like the following:

$$Hairdresser|A = \{\{trim\ hair\ into\ bob,\ add\ highlights,\ add\ lowlights, ...\}\}$$

In this case, the domain narrows because the expectation "cut hair shorter" has been made more fine-grained and thus the set of actions compatible with the domain is more narrow. Domains can also narrow by adding additional expectations to the domain. Suppose, for example, that I trust hairdressers to not only cut my hair shorter, but to provide a certain kind of experience while doing so:

$$Hairdresser|A = \{\{cut\ hair\ shorter\},\ \{provide\ advice\ on\ haircare\},$$
$$\{provide\ positive\ conversation\}...\}$$

In this case, the additional expectations unrelated to the actual haircut restrict the range of compatible actions.

Other things being equal, trust is more easily broken when the domain of trust is narrow because there are fewer compatible courses of action. Whenever we trust another, we risk the chance that they will violate our trust. However, we are not only vulnerable to the actions of those we trust, we are also vulnerable to our own

expectations. If someone's expectations are too specific, i.e. their trust domain is too narrow, it may be impossible to satisfy them. When this happens, individuals find themselves continuously disappointed in others and eventually isolated as successive trustees fail to act as expected. Even the best hairstylist may be unable to meet my expectations if the hairstyle I desire is not suited to my hair type. Thus, trust domains ought not be so narrow that it becomes impossible for others to act within their boundaries.[2]

Yet, trust domains ought not be so wide that all actions are acceptable. When someone has very few expectations or very coarse-grained expectations, trustees may easily take advantage of them. If I don't have precise expectations regarding some domain, I am vulnerable to other people meeting those expectations in ways that I may not understand or approve of (Jones, 2015). In such cases, my expectations have been met and yet I have still been harmed. For example, someone seeking help for chronic pain may simply expect a doctor to help ease their discomfort. However, an unscrupulous physician might prescribe them a highly addictive painkiller, thus meeting the patient's expectations of decreased pain, but, in doing so, also risking their long-term well-being.

In many cases, as a person gains familiarity with a given type of trust relationship, the more balanced their associated trust domain becomes. With respect to medical contexts, Wendy Rogers (2014) notes

> "...most patients have little understanding of the etiology, pathophysiology, or prognosis of the illnesses from which they suffer, nor do they know which treatments to consider. If patients had this information, they would remain vulnerable to the vicissitudes of embodied existence but would be

[2]Catherine Elgin describes an analogous phenomenon regarding knowledge. She argues that a belief is "shaky" if a belief's truth conducers and its relation to them must be almost precisely as they are for the belief to be justified (Elgin, 2008). Any small perturbation undermines such beliefs. Similarly, narrow trust domains require such precise courses of action that the trust is easily undermined if a trustee deviates even slightly.

> less vulnerable overall as they would, to a greater or lesser extent, be able to look after their own health interests." (Rogers, 2014, pg.72)

If the patient in the previous example had no medical training, they were doubly vulnerable. Without a clear understanding of what is medically possible, the patient is vulnerable to setting expectations that no doctor is able to meet. Additionally, without a clear understanding of what is medically appropriate, the patient is vulnerable to others meeting their expectations in harmful ways. However, if the patient gains familiarity with their medical condition and associated treatment options, the patient is in a better position to determine the appropriate boundaries of the trust domain through negotiating between what is desired and what is possible.

Even with familiarity, however, individuals can disagree about the proper scope of trust. Suppose Ann and Dylan are managers working for the same company. Ann has a wide trust domain regarding the employees, expecting employees to get work done, but trusting them to use their discretion in how and when that work happens. Dylan has a relatively narrow trust domain, expecting employees to follow company rules, procedures and norms. In a disciplinary meeting regarding an employee's repeated late arrival to work, Dylan deems the employee untrustworthy. Ann, however disagrees and argues that the problem does not lie with the employee, but with the Dylan's unreasonable expectations. Dylan, however, responds that it is irresponsible not to expect employees to arrive on time. In this situation, Ann and Dylan are faced with whether or not to renegotiate the boundaries of their respective domains of trust regarding employees. Whether and how they renegotiate these boundaries determines how they view the trustworthiness of the tardy employee. If Ann realizes that her trust domain is too wide, she may conclude that the employee demonstrated untrustworthiness through taking advantage of her. Alternatively, Dylan may realize his trust domain is too narrow and conclude that the employee is worthy of trust after all.

3.3.2 Domain Ordering

Trust domains are not just sets of expected behaviors or attitudes. These sets are ordered according to importance. Some types of trust relationship, such as that between a dentist and patient, have obviously important expected behaviors which cannot be reasonably removed from the domain. If A's dentist cannot perform routine dental procedures, then it would be reasonable for A to end the trust relationship. When trust relationships involve professional standards, norms and systems of accountability, as in the case of dentistry, trustees are provided with guidance on what they ought to expect. Trustees can also simply expect their dentist to hold the legally required certifications and thus offload the work of determining what is important onto the relevant governing bodies.

Other domains, however, may not have an obvious ordering of expected behaviors. Individuals' trust domains with respect to friendship, for example, often vary widely. A might place greater importance on community ties than on honesty. Thus, B might be dishonest, but if he is deeply embedded in A's community, she may be willing to renegotiate her trust domain to exclude honesty rather than end the trust relationship altogether. If someone else, C, places a very high value on honesty, they may end the friendship even if B is part of their inner circle.

Sometimes, the importance of a given expectation isn't clear until that expectation is made salient. The parents in the babysitting case, for example, had not considered whether refraining from redecorating was an important characteristic of babysitters until faced with a babysitter who did not refrain from redecorating. In college, I played a game called "Deal Breaker". In the game, players were asked whether they would break up with their significant other if that other had some unpleasant feature. Most of the time, the unpleasant feature was funny, but harmless. For example, would you break up with your partner if they walked everywhere backwards? Would you break up with your partner if they preceded every sentence with a nasally giggle?

In retrospect, the game was weird. However, it points to an important phenomenon. When thinking about what to expect in a partner, I am guessing that most people wouldn't think about whether to include walking backwards or constant nasally giggles. However, after these are brought to our attention, would we be willing to include these features in our expectations? Answers to these questions reveal how important the presence or absence of these features are to the domain of "partner". This is part of the process by which "hidden expectations" become revealed.

Deal breakers, then, are expected behaviors or attitudes that we would be willing to end the trust relationship over rather than alter our domain of trust to accommodate them. Regarding dentistry, the inability to diagnose or treat dental problems is a deal breaker for me. When people order their expectations differently, they may have different deal breakers for the same domain. For example, the inability of a hairdresser to provide pleasant conversation while cutting my hair is not a deal breaker for me, but it might be for someone else. Similarly, dishonesty may be a dealbreaker with respect to friendship for one person, but not another.

We can disagree, however, over whether a particular expectation should be a deal breaker with respect to a certain kind of trust relationship. Should we trust a dishonest friend? Should we trust dentists who cannot treat dental problems? Should we trust a grumpy hairdresser? Disagreement regarding deal breakers complicates the process of assessing the trustworthiness of another and attempting to be trustworthy. When trust is broken, it may be due to the presence of unfair or morally problematic deal breakers in the trust domain. An employer, for example, should not have as a deal breaker the expectation that their employees will be readily available twenty-four hours a day. This is an exploitative expectation and thus it is unfair to count employees who fail to meet it as untrustworthy. However, other times when trust is broken, it is because the trustee failed to meet a morally important expectation that ought to be a deal breaker. The expectation of not being physically harmed in an

intimate relationship, for example, should be a deal breaker. The abusive partner should be viewed as untrustworthy.

When it is unclear whether a given expectation should be a deal breaker, trustees may dispute accusations of untrustworthiness. In the example of the lying friend, B may argue that his dishonesty wasn't *that* big a deal and C has trusted poorly in holding honesty in such high regard. The issue here is not whether B is capable of being a trustworthy friend, but whether being a trustworthy friend should require honesty. If it does, then B is counted untrustworthy, otherwise he is not.

3.3.3 Domain Rigidity

In navigating the boundaries of trust domains, then, trusters and trustees must negotiate about the relative importance of different expectations. If they can reach agreement, then they can alter the boundaries of their trust domain, or commitment domain, accordingly. The willingness to alter a domain, however, is a further variable in negotiations. I call the ease with which domain boundaries change "rigidity". [3] The scope of a domain has to do with the existing boundaries of a trust domain, while rigidity has to do with how easily those boundaries can change. It is worth noting that the scope, the presence of deal breakers and rigidity are not independent features. Rather, they are conceptually linked to one another. In particular, the presence of deal breakers is tied to the rigidity of a domain. To the extent that a domain has deal breakers, that domain is rigid. If a domain does not have any deal breakers, then it cannot be rigid because there are no boundaries to be rigid. Additionally, domains with no deal breakers cannot be narrow in scope. Trust domains narrow in scope as more deal breakers are added.

[3] I am not using the term "rigidity" in the Kripkean sense of rigid designators.

Suppose someone's trust domain regarding hairdressers looks like the following:

$$Hairdresser|A = \{\{cut\ hair\ shorter\},\ \{provide\ advice\ on\ haircare\},\ \{provide\ positive\ conversation\}...\}$$

However, they recently went to a hairdresser who gave them the best haircut they've ever received, but were rather grouchy while giving it. If A is willing to remove the expectation of positive conversation, then hairdressing is a *flexible* domain. However, let's imagine that after their haircut, A heads to the dentist because of a toothache with the following trust domain:

$$Dentistry|A = \{\{diagnosing\ dental\ problems\},\ \{administering\ anesthesia,\ drilling\ out\ rotten\ areas,\ filling\ area\ with\ appropriate\ substance\},\ \{general\ goodwill\}\}$$

The dentist, however, is a terrible dentist. Indeed, the dentist doesn't know anything about teeth. If A is unwilling to remove the expectation of diagnosing or treating dental problems, then the domain is a *rigid* domain.

Negotiations regarding rigidity focus on how willing we ought to be to change our expectations. This is, as noted, related to negotiations regarding deal breakers, but is slightly different. Negotiations about deal breakers often focus on particular expectations. Negotiations about rigidity often focus on domains more generally. For example, if someone is willing to continue seeing the inept dentist despite their failure to provide adequate care, we might encourage them to develop a more rigid domain of trust with respect to dentistry. We might argue that they have set their standards too low, without necessarily spelling out what expectations, specifically, should be added as deal breakers. Similarly, if a domain is too rigid, we may argue that the truster should be more flexible without specifying which deal breakers are unreasonable.

3.4 Faultless and Blameless Trust-Breaking

I have outlined several ways in which trust-breaking occurs and have argued that it is not always immediately obvious who is at fault when it happens. Is it, however, possible for trust to break without either the trustee or truster being at fault? I believe that there are such cases of faultless trust-breaking.

There are at least three ways in which we might understand fault. One way we might understand fault is in a causal sense. If we understand it so, then someone or something is at fault insofar as they causally contribute to an event. This is not the sense that I am interested in exploring. There are at least two other ways in which we might understand fault. One way is that neither person has done anything wrong. Another is that someone has done something wrong, but is not culpable for their actions. When trust-breaking occurs despite neither person having done anything wrong, I call this an instance of *faultless trust-breaking*:

> **A *faultlessly breaks trust* with B if and only if, neither A nor B has done wrong in trusting or meeting trust.**

When trust-breaking occurs and someone has done wrong, but is not culpable for their actions, I call this an instance of *blameless trust-breaking*:

> **A *blamelessly breaks trust* with B, if and only if, either A or B has done wrong in trusting or meeting trust, but is not culpable.**

Let's suppose that someone invites two friends for dinner. One of the guests is a practicing Jew and the other a vegan, both of whom do not eat pork because it violates their respective religious and moral commitments. The host, however, is totally unaware of this and prepares a soup that contains pork and does not realize his mistake until halfway through the meal. The guests are horrified and the host feels terrible. This situation raises several pertinent questions. The first question is

whether anyone in this situation has done anything wrong. Was the host wrong for serving pork? Are the guests wrong for being upset about what was served to them? Or is it possible that neither party has done anything wrong?

The second question is whether anyone in this situation is blameworthy. Let us suppose that the host was wrong for serving pork, but should he be blamed for his wrongdoing? He certainly didn't intend to do wrong in serving pork. However, we might argue that he should known. Those who believe that eating any meat is immoral might argue that the host should know this. Alternatively, someone might argue that the host should have asked in advance whether their guests had any dietary restrictions. Indeed, asking about dietary restrictions has become expected over the last couple decades and we might hold the host responsible for failing to act upon this expectation. Conversely, we might argue that the guests have done wrong because they should have notified the host in advance of their dietary restrictions. Should they be blamed for their actions? Or perhaps both parties are to blame for the unfortunate event? Or neither?

As Miranda Fricker states in her 2007 book on epistemic injustice, "The line between what we should and should not blame people for, what we may and may not properly expect of them (and of ourselves), is surely fuzzy, and especially so across historical distance." (Fricker, 2007, p. 103). Her discussion of culpability in the context of epistemic injustice provides helpful insights into the possibility of faultless trust-breaking due to non-culpable wrongdoing. I will briefly outline her argument for how someone might wrong another without being culpable and what the appropriate response to such cases might be.

3.4.1 Blameless trust-breaking

Fricker argues that someone might do something wrong, but fail to be culpable because they are not in a position to know better. She uses an example from the

screenplay *The Talented Mr Ripley – Based on Patricia Highsmith's Novel*, in which a man, Herbert Greenleaf, fails to grant a woman, Marge Sherwood, the credibility she is due because of his prejudice toward women. Fricker argues that Greenleaf wrongs Marge, but isn't culpable for his actions because in the historical setting, 1950's Venice, women are viewed as unreliable and prone to hysteria. This historical context has trained Greenleaf's sensibilities to view women as less credible on a "spontaneous, unreflective level". Fricker argues that Greenleaf is the subject of a particular kind of moral bad luck:

> "More precisely, his case exemplifies a compound of circumstantial and constitutive bad luck, for it is specifically historical circumstance that has constituted him as someone who is unable to have a reason to doubt his lack of trust in Marge." (Fricker, 2007, p. 103)

Thus, while Greenleaf's treatment of Marge is morally wrong due to the way it undermines her sense of self, Fricker argues that he is not culpable for this wrong. We might, then, consider such cases blameless, insofar as blame is inappropriate.

While blame is inappropriate, she does argue that we are justified in feeling disappointment in cases such as Greenleaf's:

> "In short, we want to complain that he could have done better, where this registers not just an epistemic disappointment but also an ethical one. We may well feel, then, a certain *resentment of disappointment* – an attitude which is closely related to the resentment of blame, but falls short of it. The resentment of disappointment is still focused on the individual, but the individual conceived in a historically situated manner." (Fricker, 2007, p. 104)

If people are able to do wrong without being culpable for it in the way that Fricker describes, then it is possible for someone to wrongly break another's trust

and not be culpably blameworthy for it. That is, they may blamelessly break trust. Additionally, Fricker's discussion of the concept of "resentment of disappointment" highlights a limitation in previous discussions of broken trust. Often, trust has been identified by the reaction that people have when that trust is broken. People are said to trust when the breaking of that trust results in feelings of betrayal, not mere disappointment. In such accounts, trust, by definition, involves feelings of betrayal when trust is broken. For example, Annette Baier states that "We all depend on one anothers' psychology in countless ways, but this is not yet to trust them. The trusting can be betrayed, or at least let down, and not just disappointed." (Baier, 1986, p. 235). Richard Holton makes the connection between trusting and betrayal explicit: "I think that the difference between trust and reliance is that trust involves something like a participant stance towards the person you are trusting. When you trust someone to do something, you rely on them to do it, and you regard that reliance in a certain way: you have a readiness to feel betrayal should it be disappointed..." (Holton, 1994, p. 67).

However, if Fricker is right, then there is room for trust to be broken, but for disappointment to be the right response, rather than betrayal. Had the two guests entrusted the task of providing dinner to the host, and the host failed to provide them with an appropriate meal, betrayal doesn't seem the correct response. Betrayal indicates that the trustee is culpable for their actions. However, the guests would be justified in feeling like their host should have known better; that he should have been sensitive to his moral obligations toward animals or to his moral obligations toward others' religious commitments.

This discussion of Fricker's work highlights two key points about trust-breaking. The first is that someone can break another's trust, but not be worthy of blame for that trust-breaking. The second is that when this occurs, the response need not be one of betrayal, but may instead consist of the resentment of disappointment.

I take it as relatively uncontroversial that blameless trust-breaking, as I have described it, occurs with some regularity. Many of the examples in this chapter may be described as blameless trust-breaking. The infamous babysitter, for one, might be considered blameless in the cases where he simply misunderstood what was expected. Faultless trust-breaking, however, is not so obviously possible. Faultless trust-breaking, recall, is trust-breaking in which no one is culpable for wrongdoing because no wrongdoing has been done. In order for trust-breaking to be faultless in this stronger sense, we must commit to the idea that a trustee can act in a way that is incompatible with the expectations in a given trust domain, but the trustee's actions and the truster's expectations are both appropriate. The issue, then, is whether it is possible for a truster to place their trust well and for a trustee to meet that trust appropriately, but for the resulting trust relationship to end in broken trust. In the next section, I will draw upon the existing literature regarding faultless disagreement to argue that trust-breaking in the more robust sense of faultless trust-breaking is possible.

3.4.2 Faultless trust-breaking

The question of whether it is possible to faultlessly break trust is related to the question of whether it is possible to faultlessly disagree (Kölbel, 2004). In this section I argue that faultless trust breaking involves faultlessly disagreeing about whether a set of commitments is compatible with a set of expectations. I begin with an overview of the argument for the existence of faultless disagreements. I then describe in what sense trust-breaking can be faultless. Finally, I outline what this means for a theory of trustworthiness.

Max Kölbel defines a faultless disagreement as

> "...a situation where there is a thinker A, a thinker B, and proposition (content of judgement) p, such that:
> (a) A believes (judges) that p and B believes (judges) that not-p
> (b) Neither A nor B has made a mistake (is at fault)."

"Making a mistake", in this context, involves believing something that is not true. If one person believes p and another believes not-p, how is it possible for someone to not have made a mistake? Accepting the view that disagreements can be faultless requires rejecting realism in favor of a version of relativism. Kölbel's relativism is constrained to those propositions whose truth can only be evaluated relative to a person's perspective, such as propositions about matters of taste. It does not, however, allow that faultless disagreement is possible in non-discretionary, objective areas. He distinguishes this relativism from indexical relativism, where the content of the proposition is relative to the speaker. For example, in indexical relativism, the content of the utterance "I am late" depends on the speaker. In Kölbel's relativism, it is not the content, but the truth-value of the content that is relative.

While not all disagreements can be faultless, disagreements regarding beliefs containing discretionary contents may be. Kölbel argues that "there are *a priori* constraints that tie a belief to certain features of its possessor. If these features then differ from thinker to thinker, then the content better be of the discretionary sort" (p. 69). Thus, beliefs about the quality of taste, for example, are tied to an individual's particular constitution, including their physical features (their physiological mechanisms responsible for taste) and perhaps their cultural features (which foods they are familiar with). When individuals assess the truth of such propositions, they do so relative to their perspective. Thus, speaker A can assert that sauerkraut is delicious, speaker B can assert that sauerkraut is disgusting, and neither has made a mistake.

This is because from speaker A's perspective, it is true that sauerkraut is delicious, while from speaker B's perspective the opposite is true. Kölbel's relativism is thus restricted to things that can reasonably differ in truth value across perspectives.

In viewing trust-breaking as involving faultless disagreement, it must be possible for the appropriateness of a set of commitments in meeting a set of expectations to differ across perspectives. That is, it must be possible for truster A to hold a certain set of expectations and for trustee B to make a certain set of commitments in response and for the expectations and commitments to be compatible when viewed from one perspective, but incompatible when viewed from the other perspective. I define faultless trust-breaking as:

> A situation where there is a truster A, a trustee B, a trust domain x, and a commitment domain y, such that:
> (a) A trusts B with respect to trust domain x,
> (b) B meets x with commitment domain y,
> (c) x and y are compatible from B's perspective, but incompatible from A's perspective,
> (d) neither A nor B has made a mistake in forming the respective trust and commitment domains.

There are several important dissimilarities between faultless disagreement and faultless trust-breaking. First, what is meant by "making a mistake" is different. In faultless disagreement, "making a mistake" involves believing something that is not trust. In faultless trust-breaking, however, "making a mistake" involves expecting something that is not appropriate (if the truster) or committing to doing something that is not appropriate (if the trustee). Making a mistake in a trust relationship might involve holding untrue beliefs, but not necessarily. Someone might trust, for example, that another is competent to perform some action (when they are not). Alternatively,

someone might believe themselves competent to meet another's trust (when they are not). However, making a mistake need not involve believing something false. For example, I expect my friends to help me feel accepted in their presence. While I can frame this as a belief (I believe that trustworthy friends will accept me), doing so glosses over the affective elements of this expectation. It is not so much that I believe that my friends accept me, as that I feel accepted when I am in their presence. Making a mistake, then, might involve feeling something inappropriate to the trust domain. For example, this could involve receiving verbal abuse as an expression of love, rather than as abuse.

Like Kölbel, I am not arguing that just any case of trust-breaking can be faultless. The quack dentist, who knows nothing about teeth, but commits to performing a risky dental procedure on a naive patient is blameworthy for his failing, regardless of how competent he believes himself. Instead, faultless trust-breaking is possible in cases where the appropriateness of a commitment in response to an expectation can vary across perspectives. This variance may be due to individual preference, as Kölbel, describes. Suppose I trust a friend to bring a tasty dessert to the potluck I am hosting. I have good reasons to believe that my friend is a capable cook and willing to do what I've trusted them to do. My friend, however, has recently become convinced that cooking with sugar is immoral for health and environmental reasons. She has purged sugar from her diet and has gotten used to a sugar-free diet. She brings a sugar-free dessert that she truly finds tasty. Everyone else, however, finds it disgusting. In this case, I believe that no mistakes have been made. I had a reasonable domain of trust that my friend committed to meeting, and, from her perspective, actually did meet. She provided a tasty dessert as she was trusted to do, when the state of affairs is evaluated from her perspective. When evaluated from my perspective, however, she failed to provide a tasty dessert.

The appropriateness of a commitment in meeting an expectation can vary not only across individuals, but also across cultures. For example, in some Nordic cultures, parents let their children nap outside in the cold of winter. This cultural practice, however, is seen as inappropriate in my own American culture. Suppose an American parent trusts their Nordic friend to watch their infant on a cold winter afternoon. The friend bundles the baby up and puts them outdoors for a nap. When the parents learn of this, they are outraged because the commitment to care for the child by leaving them outside is inappropriate, when evaluated from their American perspective. From their friend's perspective, however, this action was compatible with what was expected.

In both cases, what has happened is that the trustee's commitments violated a hidden expectation in the trust domain. Doing so, however, is not a mistake. The commitments made appropriately fulfill the known expectations when evaluated from a given perspective. Once hidden expectations are revealed, however, it might be mistake to continue trusting without first renegotiating the domain of trust. I should not continue to trust my sugar-free friend to make tasty desserts if she is not willing to make desserts that I will find tasty. Similarly, I should not commit to babysitting if I am unwilling to engage in culturally appropriate care. Faultless trust-breaking, then, is limited to cases where expectations remain hidden. Once these expectations are revealed, if trust is extended without renegotiations and then subsequently broken, then someone involved has done something wrong. In some cases, renegotiation may not end in future trust relationships. If my friend is unwilling to make desserts that I will find tasty and has good reasons for doing so, and I have good reasons to not accept what she is willing to make, then I shouldn't trust her in the future with bringing dessert to my potlucks.

3.5 Conclusion

In this chapter, I have argued that trust breaking is not always due to negligence, incompetence or ill will. Instead, I believe that trust breaking can, and often does, occur despite the genuine efforts of both parties. This can be an especially painful and confusing experience for both the truster and trustee because it isn't always immediately apparent who, if anyone, is to be blamed for the broken trust. Sometimes it may not be important to know who is to blame. Instead, it may be more important to figure out how to avoid future trust breaking. Identifying the type of trust-breaking that occurs and assessing the associated domains can help clarify who is at fault, if appropriate to determine fault, as well as facilitate negotiations for future trust relationships.

If the trust-breaking was a direct misunderstanding mismatch, we can ask whether trusters communicated their expectations clearly and whether trustees took the appropriate measures to listen well. Did the truster give reasonable guidance as to what they were entrusting? Did trustees seek to understand what was being entrusted to them or did they jump to unjustified conclusions?

If the trust-breaking was a direct conflict mismatch, we must look carefully at the trust domain. In particular, did the trust domain have an appropriate scope and were the deal breakers reasonable? Were the boundaries of their trust domain so narrow that it was impossible for the trustee to act within them? Were the boundaries so wide that the truster was naively susceptible to exploitation? Should they have thought more carefully or done more research about what they ought to expect? Did they place too high a value on something that shouldn't be considered that important?

Conversely, trustees must ask themselves whether they should have put forth more effort to care for what was entrusted. Could they have done what was entrusted to them? If they could not because the domain was too narrow, did they do their best to communicate their reservations about this clearly and alert the truster as to what

they were willing to commit to? Did they appropriately signal the limits of their commitments?

Finally, if the trust-breaking was an indirect misunderstanding mismatch, we may ask whether there were hidden expectations that trusters should have made explicit. Did they do the work of self-searching to know in advance what they expected? Did they communicate this effort clearly to the trustee? Should trustees have known about these hidden expectations even though they weren't made explicit? Should they have asked more questions about the truster's expectations before accepting the trust? In the absence of this greater knowledge of the truster, were their actions reasonable and appropriate?

These questions can help both parties assess whether the broken trust is, in part or in whole, their own fault and thus whether the truster is justified in counting the trustee untrustworthy. Sometimes, however, even after carefully evaluating what happened, who is at fault may still be unclear and both parties left wondering what, if anything, they should have done differently. Sometimes the resulting confusion is enough to end trust relationships entirely. Other times, negotiating about the domains may repair the relationship enough for the truster to trust the trustee again in the future.

Chapter 4

Trustworthy AI

Introduction

At this point, I'll recount the argument made thus far. Ultimately, I aim to provide an account of trustworthiness such that it is applicable to AI. In Chapter One, I argued that people can, indeed, trust AI in important ways, despite the common portrayal of trust as relegated to the interpersonal domain. I argued that even on more stringent accounts of trust where trust depends on viewing the trustee as a person, trusters may come to trust AI through misperception or the exercise of imagination. On less stringent accounts, such as Nguyen's account of trust as an unquestioning attitude, trusting AI does not require people to misperceive or imagine the world as other than it is. Instead, it simply involves holding an unquestioning attitude about whether the AI will meet expectations. However, when those expectations include features like goodwill, intentionality, etc., then the interpersonal and unquestioning accounts have similar consequences. This highlights the importance of the trust domain - the kinds of expectations included have important implications for whether a truster needs to misperceive reality in order to extend trust.

While one might concede that people do trust AI, it does not follow that AI is always worthy of this trust. As an aside, while I speak of AI as being "trustworthy" or "not trustworthy", this is really a simplification. Just as trust comes in degrees and is sensitive to context, trustworthiness also has these characteristics. Thus, it is more accurate to talk about the extent to which a given application of AI is worthy of trust

(with respect to a particular domain, trustee, and environmental context). However, for the sake of simplicity, when I am speaking generally about the trustworthiness of AI, I will typically refer to it as either trustworthy or not.

In Chapter Two, I distinguished two kinds of trustworthiness, but primarily focused on how these forms of trustworthiness manifest in human-to-human relationships. In this chapter, I argue that these forms of trustworthiness can be sensibly applied in human-to-AI relationships, in contrast to the view that trust and trustworthiness are fundamentally interpersonal concepts. I believe that the question we should be asking is not whether it is sensible to consider AI trustworthy, but when and under what conditions should we count it naively or robustly trustworthy? In this chapter, I discuss what it would mean for AI to be both naively trustworthy and robustly trustworthy. In section one, I describe the conditions under which AI counts as naively trustworthy. Additionally, I highlight several ways in which the process of assessing the naive trustworthiness of AI differs from that of humans.

4.1 Naively Trustworthy Technology

Suppose someone acquires a companion robot. In order for that robot to be naively trustworthy, it must be likely to meet that truster's expectations regarding companionship. However, AI cannot fulfill the expectations that are typically associated with companionship, such as reciprocal understanding or mutual goodwill. Currently, AI does not have the capacity for understanding our human experiences or reciprocating emotions (and potentially never will in the relevant sense). While we might feel or act as though AI has such capacities, they are not actually meeting our expectations. Thus, it seems obvious that the companion robot is naively untrustworthy because it cannot actually do what is entrusted.

However, this conclusion is not as obvious as it may seem at first. We could, after all, renegotiate the boundaries of our trust domain to accommodate the capabilities

of the companion robot. We cannot renegotiate trust domains with AI as we do with humans—through ongoing communication and compromise—because AI isn't responsive in this way. It simply will do what it was designed to do. Any negotiations about the aims of the AI and the values we wish it to express would have to happen with the people who developed or deployed the AI, rather than the AI itself. Thus, when we think about the commitment domain of an AI, that domain consists of the potential abilities of the AI. For example, consider an AI system that is trained to classify animals in photographs. When deployed to do this, its outputs can be considered as the set of its commitments. Given an input photograph, it "commits" to mapping this to its output. Speaking of commitments in this context is unfortunately anthropomorphic given that commitments generally imply intentionality, however, in this context, a commitment is just what the system will do given a task.

While we can't renegotiate the commitment domain as we do in human-human trust relationships, we can engage in a one-sided process of negotiation with the AI itself through reflecting on the expectations within our trust domain and altering them to align with how the AI works. For example, if we alter our expectations of companionship so that it excludes any expectations that require personhood, emotions, sentience, etc., then we could count the robot naively trustworthy. There are several ways we might do this. First, we could reimagine companionship entirely, arguing that it isn't an interpersonal relation. For example, we could exchange the expectation associated with companionship that companions are agents that we can talk *with* for the expectation that companions are things we can talk *at*. And we could likewise exchange all person-dependent expectations for expectations that machines can meet.

However, a worry here is that the domain thus called "companionship" is so divorced from existing conceptions of companionship that it doesn't merit the same label. There is a related worry in the literature on conceptual engineering. Conceptual

engineering aims to improve our existing concepts for various scientific or social aims. This method has recently become common in feminist philosophical literature, where targets of conceptual engineering include "woman", "gender", etc., and the aim is to rectify existing harms related to how people have understood gender (Haslanger, 2000). For example, we might re-engineer the concept "women" so as to include transwomen, who have not counted as women in historical conceptions of womanhood. An objection to this project, however, is that in re-engineering concepts, we are simply changing the subject (Strawson, 1963; Prinzing, 2018). Similarly, the worry here is that in swapping out the typical expectations in the domain of companionship for AI-compatible expectations, then the domain ceases to be the domain of companionship. The force of this objection is minimized, however, if we recall that trust domains are determined by the truster. An individual's trust domain does not have to align with social norms or shared understandings of the domain's boundaries. Trusters can have idiosyncratic trust domains that may make no sense to those around them. In such cases, however, trusters are likely to find their trust continually disappointed due to ongoing miscommunication. Thus, it is quite possible for trusters to rework their trust domains regarding companionship so as to find a robot a suitable companion.

Secondly, instead of reworking the domain of companionship entirely, we could simply argue that the illusion of companionship is good enough. On this approach, as long as the human truster perceives or imagines that the AI meets their existing expectations of companionship, then this is sufficient. The expectations regarding companionship stay largely the same, with the exception that instead of expecting *actual* goodwill, trusters expect to feel as though the AI has goodwill. Thus, what matters for trustworthy companionship is not having actual reciprocal understanding or goodwill, but the ability to act as though one does. As long as we *feel* like the AI cares, then we can count it as naively trustworthy.

For naive trustworthiness, both approaches are sufficient. We can easily count AI agents as trustworthy companions simply by renegotiating the domain of trust, either through drastically altering our expectations or by settling for the illusion that existing expectations are being met. Yet, not all cases of trust in AI require such responses. Whether imagination, misperception, or re-engineering is required in order to find an agent naively trustworthy depends on the contents of the trust domain. When people trust AI with respect to things that are within its actual capabilities, then these responses are not needed. Thus, people can trust AI to sort information, make calculations and predictions, generate models, and perform certain motor tasks without using any imagination or altering their expectations. Trusting an AI to predict which environments will be suitable for certain chemical reactions, for example, does not require these responses. I do not, however, want to give the impression that it is straightforward to distinguish which trust domains require trusters to imagine or misperceive technology or re-engineer their trust domains. This task requires understanding what capacities are needed to meet the expectations in a given existing trust domain and whether the AI in question has those capacities.

Even if we assume that trust necessarily requires features of persons, it does not straightforwardly follow that AI must be untrustworthy. Some philosophers and computer scientists, for example, are optimistic that eventually AI will be able to do everything that people can do, including feeling emotions, exercising moral judgement, and perceiving sensations such as pain and pleasure (Asada, 2019; Bronfman et al., 2021; Shevlin, 2018). However, there is considerable debate over just what it is that people do when they feel emotions, exercise moral judgement, etc. Thus, it is unclear what conditions must be met for AI to count as having these capabilities. In the case of exercising moral judgement, for example, some philosophers think that AI must be conscious (Himma, 2009) or have a sense of self (Parthemore and Whitby, 2013). Others deny that sentience is needed, and instead argue that the AI must be

autonomous in certain respects (Sullins, 2006). Of course, these disagreements can also extend to the conditions of consciousness, selfhood, or autonomy. This means that how people conceptualize moral judgement, or other characteristics of persons, will influence whether they take imagination or misperception as necessary in order to count the AI naively trustworthy.

Naive trustworthiness, recall, has to do with only trust situational vulnerabilities. When a truster trusts someone, they become vulnerable to the trustee not following through. The key question, here, is will the trustee do what the truster expects them to do? In the context of AI, does the AI do what it was trusted to do? In order to answer this question, we need to know a) what it was trusted to do, and b) how to determine whether or not those expectations are being met.

The first issue, knowing what an AI was trusted to do, is deeply dependent on the particular aims of individual developers or users. These aims are as diverse as the people who develop and use AI. Some aims are admirable, others less so. I may develop an AI that helps me hack into banks so that I can steal their money. I could also develop an AI that identifies which children are at risk of falling behind academically, so that they can receive additional support. The issue of understanding what an AI was trusted to do simply depends on the truster's trust domain. These domains can be sensible, weird, morally corrupt, admirable, etc. In this, trusting AI is not really different from trusting other people.

Someone might object that what matters for trustworthy technology is not what the end users trust the AI to do, but whether the AI does what it was *designed* to. If this is the case, then we need only pay heed to whether the AI does what the engineers who made it wanted it to do. However, recall that naive trustworthiness has to do with whether the trustee fulfills the expectations *in the trust domain*, which differs from person to person. What matters, then, for naive trustworthiness is the individual truster's trust domain. It does not matter whether the truster is the person

who designed the AI or an end user who may trust it to do something that it was never intended to do. In the latter case, it is just more likely to be found untrustworthy.

While trusting AI is often similar to trusting humans, there is at least one interesting and important difference. The ability to assess the naive trustworthiness of a human frequently requires different skills or strategies than assessing the naive trustworthiness of an AI. While the conditions for trustworthiness remain the same (meeting the expectations in the trust domain), there are interesting issues related to our ability to assess the success of an AI that do not arise in human-human trust relationships, or at least do not arise for the same reasons. In the following, I explore one issue that has recently gained a lot of well-deserved attention: the opacity of AI decision-making.

Sometimes, when we trust, we trust not only that an agent will do something, but that we will be able to understand whether, why and how they've done it. That is, we want to be able to validate whether the AI has met expectations, to understand why they've acted as they have, and to discover things through how they've done what they've done (Watson and Floridi, 2019). Explanation is a vital tool for understanding, validating and discovering that we can take advantage of in human-human trust relationships that is often unavailable in human-AI trust relationships. For example, in trusting a doctor, a patient may want an explanation for the proposed treatment plan prior to adopting it. The explanation helps to confirm that the commitment domain is compatible with the trust domain. In other words, sometimes trust domains include the expectation that the trustee will be able to tell us why they committed to their chosen course of action.

In cases where an explanation is included in the trust domain and a human agent refuses to provide one, this is often indicative that the agent has not actually done what we expected. While they may have legitimate reasons not to explain their actions, it is often still suspicious and we are justifiably inclined to count these people

as untrustworthy. However, AI doesn't fail to offer explanations for suspicious reasons. Many forms of AI are just too computationally complex for people to make sense of the process by which they produce an output, even though those outputs may meet our other expectations. Thus, it isn't clearly justifiable to count AI as untrustworthy when it fails to provide an explanation. Indeed, despite being unexplainable, AI is often incredibly useful. We must be careful, then, not to refuse to trust AI simply because it does not offer an explanation for what it does. While the inability to offer an explanation alone does not necessarily constitute a reason to count an AI as untrustworthy, there are at least three situations in which we are justified in including an explanation in our trust domain and thus justified in counting an AI untrustworthy when it cannot provide one. We have reason to seek an explanation when a) we aren't certain what to expect, b) when the trustee meets the expectations in our trust domain in an unexpected way, or c) when we want to ensure that the trustee meets the expectations in the right way. Unfortunately, many cases where AI is especially useful are often also cases in which an explanation should be expected. Thus, there are many cases where we have good reason not to count AI as even naively trustworthy.

It is important here to distinguish between several different types of AI because not all types of AI have this explainability problem. In some cases, we can look at an algorithm and see exactly what it has done and why. When a system has hard coded rules, for example, it is easy to see why it performs a given action because it was explicitly coded to do so. Many natural language processing systems still rely on programmers explicitly coding machine responses to inputs. These systems rely on programmers developing databases of possible inputs, that is, thinking of what people might possibly say to the system or identifying relevant key words and then assigning those inputs to the appropriate outputs, or responses.

Joseph Weizenbaum's chatbot ELIZA is an early example of such a system (Weizenbaum, 1966). ELIZA, which famously mimicked a Rogerian psychotherapist, worked by identifying key words in the human participant's text and then transforming that text according to a rule associated with the keyword. While computationally very simple, ELIZA had surprisingly human-like responses. Weizenbaum relates the following typical conversation, where ELIZA's contributions are capitalized:

> Men are all alike.
>
> IN WHAT WAY?
>
> They're always bugging us about something or other.
>
> CAN YOU THINK OF A SPECIFIC EXAMPLE?
>
> Well, my boyfriend made me come here.
>
> YOUR BOYFRIEND MADE YOU COME HERE
>
> He says I'm depressed much of the time.
>
> I AM SORRY TO HEAR YOU ARE DEPRESSED
>
> It's true. I am unhappy
>
> DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
>
> I need some help that much seems certain.
>
> ...(Weizenbaum, 1966, p. 37)

Despite these human-like interactions, ELIZA was fairly limited in its capacity. Its ability to engage in conversation was limited to the amount of detail included in the script, which included both the keywords and their associated rules. ELIZA's behavior was entirely explainable. One only had to look at the script of keywords and rules to know why ELIZA produced a given output.

Indeed, Weizenbaum himself notes the following:

> "It is said that to explain is to explain away. This maxim is nowhere so well fulfilled as in the area of computer programming, especially in what is called heuristic programming and artificial intelligence. For in those realms machines are made to behave in wondrous ways, often sufficient to dazzle even the most experienced observer. But once a particular program is unmasked, once its inner workings are explained in language sufficiently plain to induce understanding, its magic crumbles away; it stands revealed as a mere collection of procedures, each quite comprehensible." (Weizenbaum, 1966, p. 36)

While ELIZA relied on preprogrammed rules, newer forms of AI employ statistical techniques to learn rules. These often have seemingly magical results. Contrast ELIZA with GPT-3, OPENAI's language model that, like ELIZA, produces human-like text. Given a short prompt, GPT-3 generates a response. These responses are often surprisingly human-like. Floridi and Chiriatti (2020) provide the following example of GPT-3 in action. They provided the machine with the first sentence of Jane Austen's unfinished novel *Sanditon*:

> "A gentleman and a lady travelling from Tunbridge towards that part of the Sussex coast which lies between Hastings and Eastbourne, being induced by business to quit the high road and attempt a very rough lane, were overturned in toiling up its long ascent, half rock, half sand."

GPT-3 continued the text as follows:

> "The gentleman was very little hurt; but the lady had two of her ribs broken, and sustained some other injuries, from which she did not recover for several months. She never saw her husband again. He was seized with

> fever on the third day after the accident, and died in less than a week. This was in the summer of 1812. In the following February the lady gave birth to a daughter, who reached her fifth month in all the vigour of health." (Floridi and Chiriatti, 2020)

Like ELIZA, however, once unmasked, the program is far less magical than it may seem. GPT-3 uses statistical learning and an enormous amount of data scraped from the internet, including encyclopedic sources like Wikipedia, to predict what text should follow from the prompt. GPT-3 is very good at reproducing the patterns it detects in the training data. This, of course, has some downsides. For example, Floridi and Chiriatti demonstrate that GPT-3 produces racist text, is poor at mathematics, and does not cope well with prompts that require semantic understanding. For example, when they prompted GPT-3 with the simple mathematical equation "solve for x:x+4 =10", it correctly answered "6". However, when they changed the equation to "solve for x: x + 40000 = 100000", it answered "50000" (Floridi and Chiriatti, 2020).

Unlike ELIZA's rule-based method, however, the statistical methods GPT-3 utilizes admit far less comprehensibility for human observers. While at a high level the procedures are comprehensible (we know it is generating new text based on its training data) the actual computations performed during text production are not. GPT-3 uses unsupervised deep learning techniques to search for patterns in the data. In order to make sense of why this poses a problem for explainability, I will briefly describe what deep learning entails.

Deep learning is just one form of a broader class of machine learning algorithms, all of which, to varying degrees, pose a challenge to human understanding and explanation. Whereas previous iterations of statistical analyses required humans to identify which features of some phenomenon were relevant to the research question, machine learning extracts such features from the data by itself. In language process-

ing, for example, relevant features might include how common a word is, where in a sentence it is likely to occur, and what words are likely to occur nearby. Deep learning is called “deep" because the raw data passes through multiple layers of nodes. At each layer, the data is transformed into a higher-level, more abstract representation of the original data according to the features that it identifies. As more data is passed through the layers during training, the system learns which features matter and is fine-tuned to use those features to identify relevant patterns in the data.

When the training data used is labeled, the learning process is called “supervised”, when the training data isn’t labeled it is called “unsupervised”. Supervised learning is often used for categorization tasks, such as categorizing a set of images as cats or dogs. In learning to distinguish dogs from cats, the system is fed a set of images labeled as cats or dogs and predicts what the image contains. The prediction is then compared to the assigned label and if the prediction is wrong, then the machine reconfigures itself accordingly. In unsupervised learning, however, there are no labels to compare the predictions to. Unsupervised machine learning is thus most frequently used to cluster data according to similarities it discovers in the data. GPT-3 is an unsupervised machine learning system. So, given a prompt like the above examples, GPT-3 will produce new text based on the features that it has learned from previous training.

However, it is difficult for humans to tell what features such algorithms extract and why those features are relevant. The transformations that occur between layers are computationally complex in such a way that it isn’t easy for humans to make sense of what is happening. Machine learning is often referred to as a “black box” because the mappings from input to output are opaque to observers. It is difficult to assess the naive trustworthiness of such “black box” technologies, especially in the following contexts.

4.1.1 Contexts requiring validation

In some contexts, when people use AI, it is challenging to determine whether the AI is meeting the expectations in the trust domain or not. This can happen when people don't know exactly what to expect and/or when there isn't an easy way to check whether or not the agent has done what they've trusted it to do. In such cases, it is possible that the AI *is* naively trustworthy in that it is actually meeting our expectations. However, there are some cases where the trust domain includes the expectation that we can validate whether it has met our other expectations. Often, the ability to validate whether expectations have been met is tied up with the trustee's ability to explain what it has done and why.

For example, projects at the Large Hadron Collider (LHC) generate hundreds of millions of collisions per second, but physicists are only interested in those collisions that are interesting. Such interesting events include those that gave proof to the existence of the Higgs Boson or those which might provide evidence of new physics beyond the Standard Model. It is simply not possible for humans to manually sort through the hundreds of millions of data points that the LHC generates and this is where AI has immense potential to help. Currently, researchers are interested in using machine learning to identify interesting data points which can then be turned into a catalogue of anomalous data points which can form the basis for new hypotheses (Ngadiuba and Pierini, 2021). However, in trusting algorithms to identify which events are interesting and which are not, there is a risk that potentially interesting data will be thrown out (Castelvecchi, 2015; Creel, 2020). Validation that the algorithm is identifying all and only interesting events is thus something that physicists have good reason to include in their trust domain.

4.1.2 Contexts requiring discovery

When an AI produces an unexpected output, we often want to know why it produced that output. Whereas the babysitter who unexpectedly painted the nursery purple was able to explain why they did it (to improve the child's well-being) an algorithm cannot do this.

Returning to the LHC example, in addition to the difficulty in verifying that the algorithm is detecting what it is supposed to detect, there is an additional risk that when the algorithm correctly identifies an interesting event, it won't be clear why that event is interesting (Roxlo and Reece, 2018). That is, it isn't clear what features in the data the machine based its decision on. However, understanding why the event is interesting is one of primary aims of physicists. What is needed in this case is an explanation that ties the algorithm's output to the physicists' expectation that the algorithm will assist in furthering their understanding of how the world works. In order to meet that expectation, these algorithms need to be transparent (Creel, 2020).

4.1.3 Contexts requiring generalizability

The lack of transparency also makes it difficult to determine whether the technology is really doing what we expect it to do, even when it produces the correct results. For example, in image recognition, a common worry is that the machine is correctly categorizing images, but for spurious reasons and thus its ability to correctly categorize novel images is diminished. Suppose the machine is designed to identify whether or not there is a dog in the picture. However, in the training data, all of the images of dogs have grass in the background. While it may correctly identify all the images of dogs, thus doing what it was designed to do, it may not actually be basing its decision on the presence of dogs, but on the grass background. This may not seem particularly problematic. After all, if the machine is producing results that meet expectations,

then it shouldn't matter whether their "thought process" mirrors our own. Unfortunately, however, if the reasons for the AI's decisions are inconsistent with what we expect the reasons to be, then it's functioning may break down in unexpected ways. If the machine is entrusted with the task of identifying dog images in a different data set, it may fail to do so accurately if there are images of dogs inside or images of grass with no dogs. While we can ask a person about their decision making process, we cannot query a machine in the same way. This makes it more difficult to know whether trust expectations are met and thus whether the agent is trustworthy.

A noteworthy problem to highlight here is that of adversarial attacks. Adversarial attacks involve altering an input so that it changes the output in ways that are undesirable. Essentially, such attacks amount to gaming the system. While adversarial attacks can be maliciously deployed to disrupt a system, developers also use them to locate and correct bugs in the program or to test the robustness of their technology. They are frequently used in image recognition, where altering the value of even one pixel in an image can sometimes change the output in surprising ways. For example, in one study, after one pixel in each image was altered the model misclassified a teapot as a joystick, a hamster as a nipple, and a baby bassinet as a paper towel (Su et al., 2019). Adversarial attacks are not limited to data based attacks, but can be deployed in physical, real-time settings. For example, objects can be added to the visual field which surprisingly alter how a visual recognition system classifies an object. Adversarial patches are stickers that can be added to a scene which will influence the model's classification. One such patch caused an AI system to classify every image that contained it as a toaster (Brown et al., 2017). Other researchers printed a 3D turtle that is consistently classified as a rifle, even when the angle of the turtle is altered (Athalye et al., 2018).

The ability of adversarial examples to disrupt an AI's functioning poses an important problem for trustworthiness. An AI's ability to withstand such adversarial

attacks, that is, its robustness, is a key requirement for naive trustworthiness. Indeed, the EU's High Level Expert Group on Trustworthy AI lists robustness as one of the central features of trustworthy AI (High Level Expert Group on Artificial Intelligence, 2019). We should not count an AI trustworthy if it is systematically susceptible to failing to meet the expectations in the trust domain. Note that robustness is evaluated relative to the trust domain. Thus, to count as trustworthy, it need not resist all adversarial attacks - only those it might encounter in the setting in which it is deployed. Thus, if an AI system is deployed in a relatively controlled setting, the bar for naive trustworthiness is considerably lower than if it is deployed in a range of real-world settings. For example, a self-driving car deployed in a lab environment or other controlled setting, such as a warehouse, meets the standard for naive trustworthiness more easily than one deployed in the real-world, where it must be capable of reliably generalizing its training in real-time in a wide range of settings.

4.1.4 Explainable AI

We are not entirely without help here. While we cannot ask a machine learning system why it makes the decisions that it does, there are a growing range of tools available to make machine learning more transparent or "explainable". There are a variety of post-hoc methods of interpretation, that is, methods that aim at explaining model behavior after the model has been trained. Some methods specific to interpreting neural networks include learned features, saliency maps, concept detection, adversarial examples, counterfactual explanations and influential instances (Molnar, 2019). Researchers have a range of goals in creating more explainable AI (XAI), including understanding causality, model transferability, model informativeness, fairness, and privacy awareness (Arrieta et al., 2020). When assessing the naive trustworthiness of a system, the best method of interpretation will depend on the potential truster's goals.

For example, if someone wants to better understand which features causally contribute to an algorithm's output, they may employ counterfactual tools to better understand how influential each feature is to the output. Suppose someone wants to understand why a hiring algorithm ranks candidates as it does. They could identify what changes to the input features (such as the wording of their resume, level of education, experience, etc.) change the output. Put as a counterfactual: how would the output change if the candidate had more experience? What if they had less education? Or if they'd changed the wording on their resume slightly? The thought is that if they can identify the feature that needs the least amount of change to alter the output, they can get a sense of how important that feature is to the model's decision. So if altering the candidate's level of education, leads to a significant change in the model output, then we know that education importantly contributes to the model output.

Counterfactual explanations don't require "looking into" the black box at all. They simply require looking at the inputs and outputs. Thus, if a potential truster's goal requires understanding what is going on inside the system, counterfactual explanations are of limited value. There are, however, increasingly more ways to try to understand what is happening within the model itself. Learned features is one such technique that is primarily used for understanding visual recognition systems (Molnar, 2019). Using this method, researchers can choose a particular neuron, channel within a network layer, or an entire network layer to focus on. After the model is trained, they hold everything else fixed but maximize the activation of the part of the network they are interested in. They then can iteratively send an image of random noise through the network, altering the image each time until the image maximally activates the desired unit. The resulting image reflects that unit's contribution to the model. It may be a pattern, texture or object. While this method can help us understand what is going on within a model, often the resulting images do not

contain a humanly recognizable concept and thus are of limited value. Additionally, in most neural network models, there are simply too many units to look at and thus this method is of limited practical value (Molnar, 2019).

Computer scientists are actively working on refining these and other techniques with the aim of improving explainability. As these methods improve, some worries related to the naive trustworthiness of such "black box" systems may diminish as we gain the ability to tell whether the system is actually doing what we expect it to do. Furthermore, as noted previously, there are many forms of AI that do not involve machine learning, but are rule based or utilize interpretable statistical methods. Such interpretable methods include models like linear regression, logistic regression and decision trees. These systems do not pose the same problems for naive trustworthiness because it is easier to understand whether they are meeting a truster's expectations.

4.1.5 Naively Trustworthy within a Context

Let us suppose, however, that explainability is, in principle, not possible in some situations. Given what I've said above, it might seem as though black-box AI systems are necessarily untrustworthy because we cannot easily understand why they do what they do. However, someone might object that as long as the machine produces the right outputs in the contexts that the truster needs it to, then it should not matter whether it does so for the right reasons. I am inclined to agree with this objection. In some cases, we might count the agent naively trustworthy just so long as it acts as expected.

It is worth noting that if we don't understand the reasoning behind the agents actions, then we don't have a good idea of what might cause it to do something unexpected. However in situations where knowing *why* an algorithm produces the result that it does is central to the trust domain, as in the case of the physicists at

the LHC, then to the extent that that algorithm remains opaque, we should not count it naively trustworthy.

However, there are lots of contexts where it may not be particularly important to know why an AI does what it does. Suppose, for example, that someone designs an algorithm that will regulate the temperature in their home while also minimizing the amount of energy expended. As long as the house remains a comfortable temperature for its inhabitants and also minimizes their energy bill, then its residents can reasonably count it naively trustworthy, even if they don't know the exact features that contribute to the algorithm's decisions.

To summarize thus far, whether an AI should be counted naively trustworthy depends on the extent to which an explanation is included in the trust domain, the extent to which such an explanation is available, and, of course, the actual functioning of the AI. However, even if we determine that an AI system is naively trustworthy, that does not guarantee that it we should trust it. This would require that the system is also robustly trustworthy. We shall now turn to the question of how to think about the robust trustworthiness of AI systems.

4.2 Robustly Trustworthy Technology

In order to determine whether AI agents are robustly trustworthy, we need to not only know whether the agents meet our expectations, but whether in meeting those expectations they meet the needs or mitigate the harms relevant to the domain of trust. In other words, in establishing whether an AI is robustly trustworthy, we must look beyond the actual capacities of artificial agents, to see whether and how engaging with those agents meets human needs or affects humans in harmful ways. Mark Coeckelbergh advocates for such a shift in thinking:

> "Instead of a philosophy of mind concerning what robots really are or really (can) think, let us turn to a philosophy of interaction and take

> seriously the ethical significance of appearance. It is a turn from the 'inside' (what is 'in the mind' of robots) to the 'outside' (what robots do to us). Let us ask: What is the ethical implication of living with personal robots, that is, of interacting with them in a personal, social, and emotional context? How do we perceive them, and what do they do to us as social and emotional beings?" (Coeckelbergh, 2009, p. 219)

Let's return to the example of a companion robot. In the previous section, I argued that it is possible to find such a bot naively trustworthy through altering our expectations to accommodate its limited abilities. A full account of what we should expect of a trustworthy companion is beyond the scope of this chapter. However, I will briefly sketch out several reasons why I believe that in rendering the bot naively trustworthy through altering our expectations of companionship, the agent becomes robustly untrustworthy. Primarily, in order to accommodate the agent's actual abilities we must adopt a view of companionship which, on most philosophical accounts of friendship, is incredibly impoverished and fails to meet the relevant social needs associated with companionship.

For example, one way to conceptualize the harm that this impoverished view of companionship represents is through the lens of virtue ethics. Charles Ess summarizes the view as follows:

> "Most briefly, virtue ethics argues that we require certain virtues (habits, practices, or excellence) in order to pursue and attain a good life, a life of flourishing – where friendship is the primary example of a human relationship that requires and fosters such virtues. ...acquiring these virtues is *difficult* – and so require practice, and so require relationships that thus push us to take up and practice what we might otherwise cheerfully avoid." (Ess, 2016, p. 66).

While Ess is primarily concerned with the morality of sexual relationships with robots, the worry he relates extends to companion robots as well. A companion that does exactly what we expect of it without reciprocating the complex emotions and behaviors that accompany the process of virtue acquisition, may make it more difficult to develop the kinds of virtues that make for a good life. If this is true, then we should not trust artificial agents in this way because it will inhibit our ability to have a virtuous, flourishing life.

Alternatively, we might approach the issue from the perspective of care ethics, which takes the caring relation as fundamental to the development of the ethical self. In order to care, a person must become *engrossed* in the other, that is, they must try to apprehend the other's reality. Additionally, this engrossment must lead them to act on behalf of the other. According to Nel Noddings, an agent is caring when it "is sufficiently engrossed in the other to listen to him and to take pleasure or pain in what he recounts. Whatever she does for the cared-for is embedded in a relationship that reveals itself as engrossment and in an attitude that warms and comforts the cared-for" (Noddings, 1984, p. 19). While we might take Noddings' account of caring as evidence that a companion bot cannot care for us, the more important point, I think, is that we cannot rightly care for them. We cannot apprehend the reality of the other because there is no other. Nor does our attempt at caring warm or comfort the cared-for because the companion bot cannot *feel* anything. In choosing companionship with a robot, then, we cease to care, and in doing so, undermine our ethical selves. Shannon Vallor explores this consequence in the context of carebots: "...how many of us be seduced by the possibility opened up by carebots of *not* caring, at least not in the intimate, direct relations that presuppose engrossment in the reality of the suffering or vulnerable other? And how can we protect ourselves against such a seduction unless we attend more carefully to the ways in which engagement in caring practices is critical for our own well-being, and not just that of those who need

us?" (Vallor, 2011, p.262). Adapting our view of companionship to accommodate the abilities of robots enables us to avoid the difficulties and risks of genuine caring. However, according to care ethics, this ultimately harms us as we cease to engage in the activity that undergirds our sense of self.

As social beings, we are inherently vulnerable to loneliness, but seeking deep relationships with other people is risky. The previous chapters explored the many different ways in which trust between people can be broken. It may be tempting, then, to turn to companion robots that are designed to adapt to one's expectations in order to avoid the risk of trust-breaking. However, virtue ethics and care ethics both offer reasons to suggest that it is harmful to attempt to meet one's social needs without actually engaging in social behavior. Robustly trustworthy companions are those whose relationships propel us to live more virtuous, caring lives. In this, AI fails us. While someone may find it naively trustworthy through altering their expectations of companionship, doing so reinforces the loneliness or other social need that motivated the trust in the first place.

Should we then assume that *all* AI is robustly untrustworthy because it requires us to adopt impoverished versions of interpersonal relationships? I do not think so. The core assumption undergirding the worry that AI is necessarily robustly untrustworthy is that all trust in AI involves harmfully renegotiating our trust domains or harmfully imagining or misperceiving the abilities of an AI. There are three reasons to reject this assumption. First, as in the case of naive trustworthiness, not all trust in AI requires us to change our trust domains or believe anything untrue. When we trust AI with things within its abilities and when it is able to meet the relevant needs or mitigate the relevant harms, then we can count it robustly trustworthy. Secondly, changing the expectations within our trust domain to accommodate the abilities of an AI is not always harmful. Indeed, changing our expectations is often necessary for making beneficial changes. Thirdly, and perhaps most controversially, believing

something untrue about the AI is not always harmful. I will consider each of these reasons in greater detail.

4.2.1 AI Acting Within Its Abilities Isn't Always Harmful

As noted in the previous section, there are many things that we can trust to AI that it is capable of doing without our changing to accommodate their limitations. The interesting question is whether in meeting our expectations, it also meets the relevant needs or mitigates the relevant harms associated with the domain of trust. Recall that for robust trustworthiness, the trustee's actions must not reinforce existing pathogenic vulnerabilities or create new ones, but, where possible, mitigate them. AI that is used for nefarious ends, then, is not robustly trustworthy, even if it meets the truster's expectations. An AI used by a greedy hacker to steal money, for example, is not robustly trustworthy because the trust is motivated by greed and AI's actions reinforce that greed.

However, AI can also be trusted to address and mitigate pathogenic vulnerabilities. For example, there is a growing body of research and collaborative efforts aimed at using AI to address pressing social needs. These efforts fall under the umbrella of "AI for Social Good" (AI4SG), defined as:

> "AI4SG $=_{def.}$ the design, development, and deployment of AI systems in ways that (i) prevent, mitigate or resolve problems adversely affecting human life and/or wellbeing of the natural world, and/or (ii) enable socially preferable and/or environmentally sustainable developments." (Floridi et al., 2020)

For example, there is a non-profit organisation "AI for Good" which seeks to develop collaborations that address problems in a number of different spheres, including education, healthcare, climate, inequality, poverty, and others (AI for Good, 2022). From predicting the spread of dengue fever over time (Carvajal et al., 2018),

to identifying students at risk of falling behind in school (Lakkaraju et al., 2015), to understanding and combating climate change (Cowls et al., 2021) people are utilizing AI to mitigate pathogenic vulnerabilities. To the extent that AI is successful in addressing these issues, it can be counted robustly trustworthy. Of course, as I will discuss later, there are a lot of obstacles that inhibit AI from successfully resolving these problems.

4.2.2 AI Accommodation Isn't Always Harmful

There are a growing number of tasks that humans have historically undertaken that are now done through AI-enabled automation. Rather than viewing this automation revolution as the result of human-like AI slowly taking over our responsibilities, it is better viewed as the result of de-coupling problems from the need for intelligence to solve them. That is, instead of creating machines that can do things like us, we are adapting the problem space to accommodate the abilities of AI. On this view, AI is a "growing resource of interactive, autonomous, and often self-learning ... *agency*, that can deal with tasks that would otherwise require human intelligence and intervention to be performed successfully." (Floridi, 2019, p. 2). Understood as such, "AI achieves its problem-solving goals by detaching the ability to perform a task successfully from any need to be intelligent in doing so." (Floridi, 2019, p. 9).

Altering our expectations of how certain tasks should be done or what certain roles require to accommodate the abilities of AI has the potential to free people up from boring, repetitive or dangerous tasks, thus allowing people to engage in more meaningful activities (Smids et al., 2020). For example, rather than sending a human in to diffuse a bomb, robots can be sent instead, thus shielding humans from harm. Robots in warehouses can do manual labor tasks that are both boring and physically taxing for humans, thus allowing humans to do more intellectually stimulating tasks that do not wear the body down. Changing our expectations to accommodate AI in

settings like these, at least on the face of it, seems beneficial, rather than harmful. After all, humans are vulnerable to boredom, fatigue, and pain and trusting AI with tasks that exacerbate these vulnerabilities can mitigate the harm that humans experience.

4.2.3 AI Illusions Aren't Always Harmful

Let us suppose that someone does misperceive or imagine AI as having properties that it does not actually have. Is this necessarily harmful? Is it always harmful to believe something false or act as though things are not what they are? Consider the following example. In the film *Cast Away*, the character played by Tom Hanks is stranded alone on a desert island. In his loneliness, his only "companion" is a volleyball that he names "Winston". He treats Winston as though it is a person, talking to it, arguing with it, and eventually grieving its loss. While there is clearly something wrong about this because the volleyball is not a person, Winston's "friendship" provides the character with the will to survive. Perhaps in some cases, then, trusting AI while knowing that it isn't actually meeting one's trust may actually benefit the truster, despite acting in discordance with one's beliefs. That said, the range of cases where this holds is undoubtedly quite narrow. Hanks' character was in incredibly unusual circumstances: entirely alone without the option for human companionship. Hopefully it is not controversial to assume that in most contexts, deriving the will to survive from companionship with a volleyball signals the presence of some underlying pathogenic vulnerability. Similarly, in most contexts, preferring the "companionship" of an AI is also a likely signal of some underlying pathogenic vulnerability.

4.2.4 Obstacles to Robustly Trustworthy AI

At this point, it is hopefully apparent that it is at least theoretically possible for AI to count as robustly trustworthy. That is, there are situations in which an AI can

genuinely meet the need or mitigate the harm relevant to the trust domain. However, for each of the above scenarios—AI working within existing expectations, adapting our expectations to accommodate AI's limitations, and accepting the illusion that AI meets existing expectations—there are situations where AI fails to be robustly trustworthy.

Firstly, when we trust AI to solve a problem that is within its capabilities, but have a hidden expectation that it will also solve some deeper problem, our trust can reinforce that deeper problem. In other words, AI can successfully function as a band-aid. However, putting a band-aid on a broken bone and saying the problem has been fixed does not solve the problem and, indeed, makes it significantly worse. AI can be naively trustworthy with respect to some surface level issue, but consequently it may cover up the deeper, more complex problem that generated the initial problem.

For example, humans are unfortunately prone to making biased decisions that negatively affect other people and exacerbate inequalities. In an effort to address this bias, people have turned to AI because the assumption is that machines don't see race, gender, disability status, etc., and therefore their decisions will be neutral. Unfortunately, as has been seen in a wide range of cases from prison recidivism risk scoring (Larson et al., 2016) to hiring decisions (Dastin, 2018), AI mirrors the biases of human decision-makers. In the case of Amazon's hiring algorithm, it consistently ranked women lower then men because the algorithm was trained on data drawn from the previous ten years. Over those ten years, more men than women were hired and the machine thus learned that men were preferable to women. However, the reasons why more men than women were hired during that time-frame are rooted in a variety of historically and socially complex issues that are unrelated to women's actual potential (Hicks, 2021). In turning to AI to address the problem of identifying good job candidates, with the underlying expectation that it would solve the problem of human bias and without addressing the background conditions that led to fewer

women being hired in the first place, it actually reinforced the gender discrimination that it was expected to solve.

Secondly, when we adapt our trust domains to accommodate the abilities of AI, we may disrupt existing social systems in harmful ways. This is notable when the disruption displaces workers or undermines people's dignity. For example, while automation has the potential to make workers' lives easier, safer, more interesting, and more fulfilling, it also has the potential to do the opposite. Smids et al (2020) recount how metro drivers in Paris felt they had been deprived of meaningful work after their jobs were partly outsourced to self-driving metros. Even though they were offered alternative employment as managers, they were not able to directly respond to emergencies and felt less connected to the lives of passengers. Thus, even though they technically received a promotion into a managerial position, they felt that their jobs had become less meaningful (Smids et al., 2020).

Finally, when people accept the illusion that AI is doing things that it cannot or has features that it actually lacks, their lives may become increasingly divorced from the shared social world, leaving them isolated and lonely. Believing true things is presumably *prima facie* a good thing for humans. Therefore, except in the kind of exceptional circumstances that Tom Hanks' character found himself in, believing things to be what they are not is not wise. Additionally, it is important for social functioning that members of a given society have some shared vision of reality. As discussed in chapter three, in order to maintain trust, the parties must have some shared understanding of the trust domain, which, in part, depends on agreement about the world and agreement about social norms. If these are dispensed with in an effort to integrate AI into one's life, this has the wider consequence of undermining trust in other humans and trust in society as a whole.

4.3 Disagreement and Trust-Breaking

Just as in the case of human-human trust relationships, there is a lot of room for trust-breaking in human-AI trust relationships. The source of the broken trust may lie at the individual, societal or cultural level because individuals, societies and cultures vary in the needs and vulnerabilities that precede trust. Thus, a given AI application may meet the needs of one individual, but reinforce the pathogenic vulnerabilities of another. Similarly, it may meet the needs of one society or culture, but harm another.

Additionally, similarly to human-human trust relationships, it is not always clear which party is at fault for the broken trust: the human or the AI. Disagreements about fault, of course, are not taken up by the AI technologies, which cannot justify or defend their actions, but by people. The disagreement may only manifest in the truster, as confusion over whether they should have acted differently, or it may be a public disagreement involving the developers of the technology, those who deployed the technology, relevant legal systems, users, and society writ large. These disagreements encompass decisions about whether the AI meets the needs relevant to the trust domain and whether introducing a change in a trust domain to accommodate the AI is harmful or not. Does it meet the human truster's needs? Does it actually mitigate the harm that it was intended to mitigate? Are the disruptions that accommodating an AI requires ultimately good for humans, perhaps despite short-term harms?

The answers to these questions are not always obvious and require not only individual decisions about AI use, but decisions on a societal and cultural scale. Indeed, it is important that the decisions about how to integrate AI into society are made with input from those whom the decisions will affect. Without such input, AI, like humans, risks blamelessly or faultlessly breaking trust with users due to its design failing to reflect the interests and values of those who are affected by its use.

In the following, I explore how trust-breaking occurs at both the individual and cultural levels and discuss difficulties in determining fault, and therefore, trustworthiness.

4.3.1 Individual-Level Trust-Breaking

On the individual level, the vulnerabilities that motivate trust in a given AI application are many and vary in the degree to which they are pathogenic. Consider the weight loss app, Noom and its target user population. Noom "uses science and personalization to help you lose weight and keep it off for good. [They]'ll help you better understand your relationship with food, how to be more mindful of your habits, and give you the knowledge and support you need for long-lasting change." (Noom, 2022). Noom combines AI and human support systems to encourage users to lose weight. Users are assigned human coaches and support groups, who offer support via chat. While humans offer coaching to users, AI is used to connect users to coaches, existing support groups, and relevant educational articles. It is also used to calculate a person's personalized "calorie budget", that is, how many calories a user can take in and still meet their weight loss goals within their desired time-frame.

Noom pitches itself as a way to change behavior through utilizing psychological research, rather than as a traditional diet. Noom does not rule out any foods. Rather, it labels foods as "red", "yellow", or "green", based on calorie density. Higher calorie foods are labeled as "red", to indicate users should limit these foods, while low calorie foods are labeled "green", to indicate users can eat more of these. Users then log everything that they eat, noting its color and adapting their eating habits to stay within their AI-generated calorie budget. Users weigh in daily to track their weight loss over time. If users forget to weigh in, the app sends a reminder to step on the scale. The app purportedly helps people develop a healthy relationship with food and some company sponsored research suggests that it may be useful in preventing

diabetes through weight loss interventions (Toro-Ramos et al., 2020). Unfortunately, there are also numerous anecdotal reports of individuals whose use of Noom triggered the onset or relapse of an eating disorder, although there are no robust studies that confirm this occurs on a widespread scale (Greenway, 2021; Sole-Smith, 2021).

An individual newly diagnosed with diabetes may really benefit from the recommendations and support that the app offers. However, the individual seeking to improve their health by losing some weight who ends up with an eating disorder may feel that their trust has been broken. Does this mean that the app is robustly untrustworthy? What exactly has happened in such cases and which party, if any, is at fault for the broken trust?

Let's begin by analyzing how the trust was broken through comparing users' trust domains and the app's commitment domain. Let's suppose there are two users, A and B, who both trust Noom with respect to the domain of weight loss help. Both expect that the app will help them lose a certain number of pounds, will provide personalized diet advice, access to a supportive community, and information about nutrition. Both also expect that the app will assist them in becoming healthier. However, their expectations surrounding health differ. "Health", for A, means managing their diabetes, an expectation that is fairly fine-grained. As such, they have a goal to lose a specific number of pounds, as recommended by their primary care physician. For B, however, their expectation of improved health remains course-grained; they just want to be healthier and believe losing weight is the way to meet this goal.

$$\begin{aligned} \mathit{Weight\ Loss\ Help}|A = \{&\mathit{lose\ x\ lbs},\ \mathit{personalized\ calorie\ plan}\ , \mathit{supportive} \\ &\mathit{community},\ \mathit{nutrition\ education}, \mathit{manage\ diabetes}\} \end{aligned}$$

$$Weight\ Loss\ Help|B = \{lose\ weight,\ personalized\ calorie\ plan,\ supportive\ community,\ nutrition\ education,\ improve\ health\}$$

$$Weight\ Loss\ Help*|Noom = \{provide\ personalized\ calorie\ plan,\ supportive\ community,\ nutrition\ education,\ provide\ accountability\ via\ push\ notifications,\ improve\ user\ health\}$$

Given that the trust and commitment domains align in both cases, there shouldn't be any issues with trust-breaking. However, there is an important problem in B's trust relationship with the app. Recall that course-grained trust domains can indicate the presence of hidden expectations; expectations that trusters may not have thought of before or may not be consciously aware of. Let's suppose that in B's case, these hidden expectations include the expectation that improving health will improve self-image. They expect that in losing weight, they will become more healthy, and subsequently they will also become more confident, more in control, more beautiful and more like-able. These hidden expectations reflect complex social and historical narratives about weight, health, beauty, and human value; narratives that are often harmful and difficult to escape, especially for women (Manne, 2022).

The app, in this case, provides a "band-aid solution" to the complex problem that B is dealing with. While the app successfully assists B in losing weight, it avoids dealing with the harmful narratives that motivated the trust in the first place. Addressing the surface level issue (weight), in B's case, reinforces a pathogenic preceding vulnerability (poor self-image) and generates a new pathogenic vulnerability (an eating disorder). B's trust in the app is thus broken due to an indirect misunderstanding

mismatch: Noom fails to meet B's hidden expectation that their self-image will improve as a result of the trust relationship. In fact, the app's recommended daily weigh-ins make B *more* self-conscious. Additionally, the emphasis on losing weight via maintaining a calorie deficit leads B to develop harmful, obsessive behaviors regarding their food.

Which party, if any, is at fault for this broken trust? If someone is at fault, are they blameworthy? For this to be a case of faultless broken trust, it must meet the following conditions:

> (a) the truster must trust the trustee with respect to some domain,
>
> (b) the trustee must meet that trust domain with a set of commitments,
>
> (c) the trust and commitment domains are compatible from the trustee's perspective, but incompatible from the truster's perspective
>
> (d) neither the truster nor trustee has made a mistake in forming their respective trust and commitment domains.

In this case, conditions (a) and (b) are obviously met. From the app's perspective, the commitment domain is compatible with the trust domain - at least the trust domain as it is prior to the hidden expectations coming to light, so condition (c) is met. However, in this case, it seems quite clear that mistakes have been made in this trust relationship and this is, consequently, not a case of faultless broken trust. B has clearly made a mistake because they've expected something from the app that it cannot offer: improved self-image. If the fault is B's alone, then they can renegotiate the trust domain to reflect the app's actual commitments: to help users lose weight. This renegotiation, of course, is more easily said than done as it requires that B challenge narratives about weight and beauty, health, confidence, etc. Theoretically, however, we can imagine that B overcomes these harmful narratives. Interestingly, however, in doing this, the need for B to trust the app dissolves entirely: it becomes

apparent that they didn't actually need to lose weight in order to improve their health or self-image. There is no need for a band-aid if the broken bone is healed.

Unfortunately, given that the app's profitability depends on users continuing their monthly subscriptions, the app is designed in such a way as to reinforce the harmful narrative that weight loss is, prima facie, beneficial for health and well-being. For example, the app builds in several assumptions that health professionals have questioned in recent years, including the assumption that BMI is a reliable indicator of health status, that weight loss is best achieved by limiting the intake of calorie dense foods (regardless of nutrient content), and that daily weigh-ins are a beneficial form of accountability. For potential users like *B*, who are pathogenically vulnerable to developing eating disorders due to existing beliefs and behavioral tendencies, and whose health is not actually compromised by their weight, the app is robustly untrustworthy.

For users like *A*, however, who do not hold harmful narratives about their weight and whose health would actually benefit from weight loss, Noom meets a legitimate need and can count as robustly trustworthy. It should be noted, however, that harmful narratives about weight and health are pervasive in American culture, influencing not only ideals of beauty in popular culture, but also health policy and access to quality healthcare (Hunger et al., 2016; Major et al., 2018). Thus, individuals who are not affected by these narratives or who will remain resistant to forming them, given the app's design, are likely to be few and far between. People are, after all, vulnerable to viewing themselves in an unjustifiably negative light, and to the extent that trusting the app generates or reinforces such views, it should be counted untrustworthy.

This discussion has hopefully highlighted how individual differences in the vulnerabilities that precede trust relationship render trustees as more or less trustworthy. Someone might object at this point that, in this case, it is not AI alone whose trustworthiness is at issue, but the app environment as a whole - including the coaches,

support groups, articles, design choices and user-interface. However, the AI's recommendations for users' personalized calorie budget and recommended educational articles play an important role in how users engage with the app and thus how those users are made more or less vulnerable. If we focus more narrowly on what exactly users trust the AI to do (provide a personalized calorie budget that enables weight loss), rather than the app as a whole, the result is the same: for some it will be trustworthy and for others not. That is, for some, it will mitigate the pathogenic vulnerability that motivated the trust, but for others, it will exacerbate it.

4.3.2 Cultural-Level Trust-Breaking

In an increasingly globalized world, it is not unusual for an AI application designed and developed in one culture to be deployed in another. Unfortunately, this can have negative consequences when an AI does not account for cultural differences in expectations or the ways that people of different cultures are differently vulnerable. Trust-breaking can occur, for example, when the commitments that fulfill a set of expectations in one culture differ from the commitments that fulfill that same set of expectations in a different culture. An AI application designed from the perspective of the first culture may faultlessly break trust with users from the second culture.

In Chapter One, I discussed the problem of privacy. What is expected of "privacy preserving tech" differs across cultures, with important differences in how privacy is understood in relation to autonomy. We are now in a position to make sense of why this is a problem for cross-cultural trust and trustworthiness. For example, suppose a company based in India designs an AI-enabled app that will be deployed in both the US and in India.

They know that both American and Indian users expect that their privacy will be preserved while using the app:

$$App|Indian\ Users = \{\{preserve\ privacy\},\ etc.\}$$

$$App|American\ Users = \{\{preserve\ privacy\},\ etc.\}$$

However, the Indian developers fail to account for how privacy is understood differently in the US. Subhajit Basu describes some of these differences as follows:

> "For example, while it is not common practice in the UK for general practitioners (GPs) to discuss patient information of the wife with her husband, such discussion is quite common in India, where GPs regularly discuss such issues with the husband or other members of the family. Does this reduced control necessarily imply reduced privacy? On the contrary, this practice conforms to a consistent set of rules as developed within the Indian society that the level of privacy should not interfere with an individual's ability to feel emotionally safe and secure through social linkage. I believe it also means that 'being private' in traditional Indian concept includes sharing with 'family' and within the 'familial space' and evidence suggests that the desire to demarcate one's life from others in the family is nearly non-existent but it is separate from the outside world as the privacy 'of the family' is greatly valued." (Basu, 2012, p.19).

Basu argues that in countries like the US and UK, privacy is generally understood as applying at the level of the individual, rather than the group, because privacy is valued due to the role that it plays in promoting autonomy and individuality. In India, however, individuals are understood as interconnected with others such that privacy

is not valued at the level of the individual, but at the level of groups - such as families. Supposing that this is correct, let's say that the developers prioritize privacy at the level of families through facilitating information sharing between family members, but encrypting data shared between groups. For the developers then, "preserving privacy" results in the following commitment domain:

$$App * | Developers = \{\{make\ data\ shareable\ between\ group\ members,\ encrypt\ data\ shared\ between\ groups\}, etc.\}$$

These commitments meet user expectations in one context (India), but fail to meet user expectations in a different context (US). In this case, all the conditions for faultless trust-breaking are met. The American users trust the app to preserve their privacy, the app commits to preserve privacy, the commitment domain and the trust domain are compatible when evaluated relative to the Indian perspective of privacy, but incompatible when evaluated relative to the American context. Yet, assuming that privacy concepts vary across cultures, neither party has made a mistake.

Once the trust has been broken, however, users must renegotiate their domains in order to accommodate the app if they wish to count the app trustworthy. In doing so, they must alter their expectations regarding privacy. In order to do this, however, they must shift their values accordingly—making autonomy subject to interconnectivity. Does this alteration render the app robustly untrustworthy for the user? Whether there is a right way to think about privacy specifically is a matter beyond the scope of this paper, so I will not address that here. Instead, I will focus more generally on how adapting to a different culture's norms in order to accommodate an AI technology might be understood as harmful: through a form of "othering".

Halycon M. Lawrence argues that voice recognition technologies perpetrate this harm on users when the technology is not trained on the voices and dialects of those for whom the technology is deployed.

Lawrence argues that:

> "If you possess a foreign accent or speak in a dialect, speech technologies practice a form of "othering" that is biased and disciplinary, demanding a form of postcolonial assimilation to standard accents that "silences" the speaker's sociohistorical reality.
>
> Because these technologies have not been fundamentally designed to process non-standard and foreign-accented speech, speakers often have to make adjustments to their speech—that is, change their accents—to reduce recognition errors. The result is the sustained marginalization and deligitimization of nonstandard and foreign-accented speakers of the English language." (Lawrence, 2021, p. 181)

According to Lawrence, then, deploying a technology that requires users to adapt aspects of the culture where the technology was developed can amount to a form of colonialism. At the very least, widespread deployment of technologies across cultures can increase cultural homogenization through encouraging users to adapt to the culture of the technology's origin. The extent to which such homogenization is harmful is unclear. Assuming that there is value in diversity, the trustworthiness of a given AI technology may extend only to those cultures whose values are expressed by the technology. For members of cultures whose unique cultural identities are vulnerable to disappearing and who wish to preserve their cultural identities, AI technologies that do not represent their cultural values should not be counted trustworthy. This is not, however, due to some inherent flaw in the technology, but due to an inherent limitation. AI cannot reflect everyone's cultural values or ways of life because these values are often in conflict. It is not inappropriate for an AI to optimize for privacy at the level of the family or to recognize the voices of native English speakers. It may be inappropriate, however, to deploy that technology in a diverse cultural context or across cultures with the expectation that users can adapt to the technology without

consequence. Whether an AI technology should be trusted, then, may depend on particular cultural contexts.

4.4 Conclusion

In this dissertation, I set out to argue that individual end users can reasonably count AI as worthy of trust. I first argued that people can (and probably do) trust AI in a variety of ways. Even on more demanding accounts of trust that require that agents be sentient or have intentional attitudes, people may come to trust AI through imagining or misperceiving it as possessing properties that it lacks. On less demanding accounts, such as Nguyen's account of trust as an unquestioning attitude, people can trust AI with things that fall within the AI's abilities or through adapting their domains of trust to accommodate its limitations. Whether trust is appropriate, however, depends on the trustworthiness of the AI.

Next, I distinguished two different kinds of trustworthiness, naive and robust trustworthiness, and demonstrated that AI can meet the requirements for both. In particular, AI can be robustly trustworthy in scenarios where it is deployed to meet legitimate needs and the ability to meet those needs falls within the range of its abilities. It can also be robustly trustworthy when we alter our expectations to accommodate its abilities and this accommodation is not harmful, as in some cases of automating dangerous or boring tasks. Finally, it can be robustly trustworthy even when people imagine or misperceive it as doing something it is not actually doing, although instances of this are likely few and far between. Assessing the trustworthiness of an AI application, then, requires careful attention both to abilities of the AI and to the domain of trust and how that domain changes over time.

I then argued that the domain specificity of trust creates room for disagreements regarding the trustworthiness of an agent. In the case of robustly trustworthy AI, people may disagree about whether AI is being deployed as a "band-aid" to cover

deeper problems, about whether adapting to AI's abilities and limitations causes trusters harm, and whether believing false things about the AI isolates or otherwise harms users. These conflicts reflect underlying disagreements about which needs are legitimate and what counts as harmful. I have characterized these as disagreements about which vulnerabilities are pathogenic (and should be eliminated) and which must be lived with, and also as disagreements about the best way to eliminate or manage vulnerabilities. Given these disagreements, it is likely that people will disagree about whether a particular AI application is trustworthy or who is at fault when trust is broken.

Finally, I argued that the trustworthiness of an AI should be determined relative to individual users and/or users' cultures. Determining whether an AI is trustworthy requires dialogue between users, developers and the broader society whose lives are also impacted by the deployment of new technologies. This dialogue must be sensitive to changes in the individual and the society to which they belong. The inclusion of user perspectives in development is important, especially in cross-cultural deployment, because users are differently vulnerable, especially so across cultures. For example, a society where weight does not carry the same social meaning as it does in many so-called Western cultures is not similarly vulnerable to the harms that a weight-loss app poses. Trustworthiness is thus sensitive to context, and as context changes, so too must our practices of extending trust.

Bibliography

AI for Good (2022). What if ai were developed to serve humanity rather than commerce? https://ai4good.org/about-us/.

Alfano, M. and Hujits, N. (2020). Trust in institutions and governance. In Simon, J., editor, *The Routledge Handbook of Trust and Philosophy*, chapter 20, pages 256–270. Routledge.

Alonso, F. M. (2014). What is reliance? *Canadian Journal of Philosophy*, 44(2):163–183.

Anderson, J. (2014). Autonomy and vulnerability entwined. In Mackenzie, C., Rogers, W., and Dodds, S., editors, *Vulnerability: New Essays in Ethics and Feminist Philosophy*, chapter 5, pages 134–161. Oxford University Press.

Arrieta, A. B., Diaz-Rodriguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Asada, M. (2019). Artificial pain may induce empathy, morality, and ethics in the conscious mind of robots. *Philosophies*, 4(3).

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2018). Synthesizing robust adversarial examples. In *Internation conference on machine learning*, pages 284–293. PMLR.

Baier, A. (1986). Trust and antitrust. *Ethics*, 96(2):231–260.

Bartneck, C., Rosalia, C., Menges, R., and Deckers, I. (2005). Robot abuse – a limitation of the media equation. In *Proceedings of the Interact 2005 Workshop on Agent Abuse*, Rome.

Basu, S. (2012). Privacy protection: A tale of two cultures. *Masaryk University Journal of Law and Technology*.

Bronfman, Z., Ginsburg, S., and Jablonka, E. (2021). When will robots be sentient. *Journal of Artificial Intelligence and Consciousness*, 8(2):183–203.

Brown, T. B., Mané, D., Roy, A., Abadi, M., and Gilmer, J. (2017). Adversarial patch. *arXiv:1712.09665*.

Capurro, R. (2005). Privacy. an intercultural perspective. *Ethics and Information Technology*, 7:37–47.

Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D. M., and Watanabe, K. (2018). Machine learning methods reveal the temporal patter of dengue incidence using meterological factors in metropolitan manila, philippines. *BMC Infections Diseases*, 18(183).

Castelvecchi, D. (2015). Artificial intelligence called in to tackle lhc data deluge. *Nature News*, 528(18-19).

Coeckelbergh, M. (2009). Personal robots, appearance, and human good: A methodological reflection on roboethics. *International Journal of Social Robotics*, 1(3):217–221.

Cowls, J., Tsamados, A., Taddeo, M., and Floridi, L. (2021). The ai gambit: leveraging artificial intelligence to combat climate change–opportunities, challenges, and recommendations. *AI and Society*.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87:568–589.

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*.

D'Cruz, J. (2018). Trust within limits. *International Journal of Philosophical Studies*, 26(2):240–250.

D'Cruz, J. (2020). Trust and distrust. In Simon, J., editor, *The Routledge Handbook of Trust and Philosophy*, pages 41–51. Taylor & Francis, New York, NY.

Dimock, S. (2020). Trust and cooperation. In Simon, J., editor, *The Routledge Handbook of Trust and Philosophy*, chapter 13, pages 160–174. Routledge.

Domenicucci, J. and Holton, R. (2017). Trust as a two-place relation. In Faulkner, P. and Simpson, T., editors, *The Philosophy of Trust*, pages 149–160. Oxford University Press.

Dworkin, G. (1982). Autonomy and informed consent. In *Making Health Care Decisions*, volume Three: Appendices of *Studies on the Foundations of Informed Consent*, pages 63–81. President's Commission for the Study of Ethical Problems in Medicine and Biomeditcal and Behavioral Research., Washington, D.C.

Elgin, C. Z. (2008). Trustworthiness. *Philosophical Papers*, 37(3):371–387.

Ess, C. (2006). Ethical pluralism and global information ethis. *Ethics and Information Technology*, 8:215–226.

Ess, C. (2016). What's love got to do with it? In Nørskov, M., editor, *Social Robots: Boundaries, Potential, Challenges*, chapter 4, pages 57–79. Routledge, New York, NY.

Faulkner, P. (2015). The attitude of trust is basic. *Analysis*, 75(3):424–429.

Fineman, M. A. (2010). The vulnerable subject and the responsive state. *Emory Law Journal*, 60(2).

Floridi, L. (2019). What the near future of artificial intelligence could be. *Philosophy & Technology*, 32(1):1–15.

Floridi, L. and Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Floridi, L., Cowls, J., King, T. C., and Taddeo, M. (2020). How to design ai for social good: Seven essential factors. *Science and Engineering Ethics*, 26:1771–1796.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, New York.

Frost-Arnold, K. (2014). The cognitive attitude of rational trust. *Synthese*, 191:1957–1974.

Greenway, K. H. (2021). Noom lured me in with its wellness hype — and wrecked my relationship with food. *Cosmopolitan*.

Hardin, R. (2001). Conceptions and explanations of trust. In Cook, K. S., editor, *Trust in Society*, pages 3–39. Russell Sage Foundation.

Hardin, R. (2002). *Trust and Trustworthiness*. Russell Sage Foundation.

Hardin, R. (2006). *Trust*. Cambridge: Polity Press.

Haslanger, S. (2000). Gender and race: (what) are they? (what) do we want them to be? *Nous*, 34:31–55.

Hawley, K. (2014). Trust, distrust, and commitment. *Nous*, 48(1):1–20.

Henkin, L. (1974). Privacy and autonomy. *Columbia Law Review*, 74(8):1410–1433.

Hicks, M. (2021). Sexism is a feature, not a bug. In Mullaney, T. S., Peters, B., Hicks, M., and Philip, K., editors, *Your Computer Is On Fire*, chapter 6, pages 135–158. The MIT Press.

High Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. European Commission.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11:19–29.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy*, 72(1):63–76.

Hongladarom, S. (2016). *A Buddhist Theory of Privacy*. Springer Briefs in Philosophy. Springer.

Hunger, J. M., Tomiyama, A., Nguyen-Cuu, J., and Wells, C. (2016). Moving to a personalized medicine approach to promote health across the weight spectrum. *International Journal of Obesity*.

Intuition Robotics (2017). ELLIQ - the active aging companion. `https://www.youtube.com/watch?time_continue=1&v=URcuVfzwB4g&feature=emb_logo` Date of Access: 1/23/2020.

Intuition Robotics (2019). ELLIQ - beta users testimonials. `https://www.youtube.com/watch?v=emrqHpC8Bs8` Date of Access: 1/23/2020.

Johnson, C. R. (2020). Epistemic vulnerability. *International Journal of Philosophical Studies*, 28(5):677–691.

Jones, K. (1996). Trust as an affective attitude. *Ethics*, 107(1).

Jones, K. (2004). Trust and terror. In Walker, M. U. and DesAutels, P., editors, *Moral Psychology: Feminist Ethics and Social Theory*, chapter 1, pages 3–18. Rowman and Littlefield.

Jones, K. (2012). Trustworthiness. *Ethics*, 123(1):61–85.

Jones, K. (2015). Trust: Philosophical aspects. *International Encyclopedia of the Social & Behavioral Sciences, 2nd Ed.*, 24.

Jones, K. (2019). Trust, distrust, and affective looping. *Philosophical Studies*, 176:955–968.

Jozuka, E. (2018). Beyond dimensions: The man who married a hologram. *CNN*, December. https://www.cnn.com/2018/12/28/health/rise-of-digisexuals-intl/index.html.

Keren, A. (2014). Trust and belief: a preemptive reasons account. *Synthese*, 191:2593–2615.

Keren, A. (2020). Trust and belief. In Simon, J., editor, *The Routledge Handbook of Trust and Philosophy*, chapter 9, pages 109–120. Routledge.

Kittay, E. F. (1999). *Love's Labor: Essays on Women, Equality, and Dependency*. Routledge, New York.

Kölbel, M. (2004). Iii—faultless disagreement. In *Proceedings of the Aristotelian Society*, volume 104, Oxford, UK. Oxford University Press.

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., and Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica*.

Lawrence, H. M. (2021). Siri disciplines. In Mullaney, T. S., Peters, B., Hicks, M., and Philip, K., editors, *Your Computer Is On Fire*, pages 179–197. MIT Press.

Leskin, P. (2018). Over a million people asked amazon's alexa to marry them in 2017 and it turned them all down. *Business Insider*, October. https://www.businessinsider.com/amazons-alexa-got-over-1-million-marriage-proposals-in-2017-2018-10.

Locke, R. M. (2001). Building trust. In *Annual Meetings of the American Political Science Association*.

Ma, Y. (2019). Relational privacy: Where the east and the west could meet. In *Proceeding of the Association for Information Science and Technology*.

Mackenzie, C. (2014). The importance of relational autonomy and capabilities for an ethics of vulnerability. In Mackenzie, C., Rogers, W., and Dodds, S., editors, *Vulnerability: New Essays in Ethics and Feminist Philosophy*, chapter 1, pages 33–59. Oxford University Press.

Mackenzie, C., Rogers, W., and Dodds, S. (2014). Introduction: What is vulnerability and why does it matter for moral theory. In Mackenzie, C., Rogers, W., and Dodds, S., editors, *Vulnerability: New Essays in Ethics and Feminist Philosophy*, pages 1–29. Oxford University Press.

Major, B., Tomiyama, A. J., and Hunger, J. M. (2018). The negative and bidirectional effects of weight stigma on health. In Major, B., Dovidio, J., and Link, B., editors, *Oxford Handbook of Stigma, Discrimination, and Health*, chapter 27, pages 499–519. Oxford University Press.

Manne, K. (2022). Diet culture is unhealthy. it's also immoral. *The New York Times*.

Markovà, I., Linell, P., and Gillespie, A. (2008). Trust and distrust in society. In Markovà, I. and Gillespie, A., editors, *Trust and Distrust: Sociocultural Perspectives*, chapter 1, pages 3–27. Information Age Publishing, Charlotte, NC.

Mattu, S. and Hill, K. (2018). The house that spied on me. *Gizmodo*, https://gizmodo.com/the-house-that-spied-on-me-1822429852.

McArthur, N. and Twist, M. L. C. (2017). The rise of digisexuality: therapeutic challenges and possibilities. *Sexual and Relationship Therapy*, 32(3-4):334–344.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (1955). A proposal for the dartmouth summer research project on artificial intelligence.

Medina, J. (2020). Trust and epistemic injustice. In Simon, J., editor, *The Routledge Handbook of Trust and Philosophy*, pages 52–63. Taylor & Francis, New York, NY.

Misselhorn, C. (2009). Empathy with inanimate objects and the uncanny valley. *Minds and Machines*, 19:345–359.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, pages 1–19.

Molnar, C. (2019). *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/.

Morrone, A., Tontoranelli, N., and Ranuzzi, G. (2009). How good is trust? measuring trust and its role for the progress of societies. OECD Statistics Working Papers.

Ngadiuba, J. and Pierini, M. (2021). Hunting anomalies with an ai trigger. *CERN Courier*, https://cerncourier.com/a/hunting-anomalies-with-an-ai-trigger/.

Nguyen, C. T. (2019). Trust as an unquestioning attitude. *Oxford Studies in Epistemology*.

Nickel, P. J., Franssen, M., and Kroes, P. (2010). Can we make sense of the notion of trustworthy technology? *Knowledge, Technology and Policy*, 23:429–444.

Noddings, N. (1984). *Caring: A Feminine Approach to Ethics and Moral Education*. University of California Press, Berkeley and Los Angelos, CA.

Noom (2022). Noom weight. https://web.noom.com/weight-loss/.

O'Neill, O. (2018). Linking trust to trustworthiness. *International Journal of Philosophical Studies*, 26(2):293–300.

O'Neill, O. (2020). Questioning trust. In Simon, J., editor, *The Routledge Handbook of Trust and Philosophy*, chapter 1, pages 17–27. Routledge.

Ortiz, D. R. (1989). Privacy, autonomy, and consent. *Harvard Journal of Law and Public Policy*, 12.

Parthemore, J. and Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness*, 5(2):105–129.

Pitt, J. C. (2010). It's not about technology. *Knowledge, Technology and Policy*, 23:44–454.

Potter, N. N. (2002). *How Can I Be Trusted?* Rowman & Littlefield, Maryland.

Prinzing, M. (2018). The revisionist's rubric: conceptual engineering and the discontinuity objection. *Inquiry*, 61(8):854–880.

Redstone, J. (2016). Makin sense of empathy with sociable robots. In Nørskov, M., editor, *Social Robots: Boundaries, Potential, Challenges*, chapter 2, pages 19–38. Routledge, New York, NY.

Reviglio, U. and Alunge, R. (2020). “i am datafied because we are datified": an ubuntu perspective on (relational) privacy. *Philosophy & Technology*, 33:595–612.

Rogers, W. (2014). Vulnerability and bioethics. In Mackenzie, C., Rogers, W., and Dodds, S., editors, *Vulnerability: New Essays in Ethics and Feminist Philosophy*, chapter 2, pages 60–87. Oxford University Press.

Roxlo, T. and Reece, M. (2018). Opening the black box of neural nets: Case studies in stop/top discrimination. https://arxiv.org/abs /1804.09278.

Scheman, N. (2020). Trust and trustworthiness. In Simon, J., editor, *The Routledge Handbook of Trust and Philosophy*, chapter 2, pages 28–40. Routledge.

Shevlin, H. (2018). To build conscious machines, focus on general intelligence: A framework for the assessment of consciousness in biological and artificial systems. In *CUER Workship Proceedings*.

Simpson, T. (2017). Trust and evidence. In Faulkner, P. and Simpson, T., editors, *The Philosophy of Trust*, pages 177–194. Oxford University Press.

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., Pistrang, N., and Sanchez-Vives, M. V. (2006). A virtual reprise of the stanley milgram obedience experiments. *PLoS ONE*, 1(1):e39.

Smids, J., Nyholm, S., and Berkers, H. (2020). Robots in the workplace: a threat to - or opportunity for - meaningful work? *Philosophy & Technology*, 33:503–522.

Sole-Smith, V. (2021). The dieter’s diet. *Bustle*.

Spellings, M. and Glotzer, S. C. (2018). Machine learning for crystal identification and discovery. *AIChE Journal*, 64(6):2198–2206.

Strawson, P. (1963). Canrap's view on constructed systems versus natural languages in analytic philosophy. In Schilpp, P., editor, *The Philosophy of Rudolf Carnap*, chapter 16, pages 503–518. Open Court, La Salle.

Su, J., Vargas, D. V., and Sakurai, K. (2019). One pixel attack for fooling deep neural networks. In *IEEE Transaction on Evolutionary Computation*.

Sullins, J. P. (2006). When is a robot a moral agent? *International Review of Information Ethics*, 6.

Toro-Ramos, T., Michaelides, A., Anton, M., Karim, Z., Kang-Oh, L., Argyrou, C., Loukaidou, E., Charitou, M. M., Sze, W., and Miller, J. D. (2020). Mobile delivery of the diabetes prevention program in people with prediabetes: Randomized controlled trial. *JMIR Mhealth Uhealth*, 8(7).

Townley, C. and Garfield, J. L. (2013). Public trust. In Townley, C. and Makela, P., editors, *Trust*. Brill Rodopi.

Tutić, A. and Voss, T. (2020). Trust and game theory. In Simon, J., editor, *The Routledge Handbook of Trust and Philosophy*, chapter 14, pages 175–188. Routledge, New York, NY.

Ullmann-Margalit, E. (2004). Trust, distrust and in between. In Hardin, R., editor, *Distrust*, chapter 3, pages 60–82. Russell Sage Foundation, New York.

Vallor, S. (2011). Knowing what to wish for. *Techne*, 15(2):137–155.

Watson, D. and Floridi, L. (2019). The explanation game: A formal framework for interpretable machine learning. Available at SSRN 3509737.

Weizenbaum, J. (1966). ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Westin, A. F. (1967). *Privacy and Freedom.* Atheneum, New York.

Whyte, D. (2020). *Consolations: The Solace, Nourishment and Underlying Meaning of Everyday Words (Revised Edition).* Many Rivers Press, Langley, WA.